IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 March 2024, accepted 8 May 2024, date of publication 20 May 2024, date of current version 11 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3401669

RESEARCH ARTICLE

HMSAM-UNet: A Hierarchical Multi-Scale Attention Module-Based Convolutional Neural Network for Improved CT Image Segmentation

NA LIU[1], ZHONGHUA LU[1], WENYONG LIAN[2], MIN TIAN[1], CHIYUE MA[1], AND LIJUAN PENG[3]

[1]College of Mechanical and Electrical Engineering, Shihezi University, Shihezi 832000, China
[2]General Hospital of the Third Division of Xinjiang Production and Construction Corps, Tumxuk, Xinjiang 659003, China
[3]College of Sciences, Shihezi University, Shihezi 832000, China

Corresponding author: Zhonghua Lu (luzhonghua@stu.shzu.edu.cn)

This work was supported in part by the Division City Key Areas of Science and Technology Research Program Projects under Grant S202202GG048, in part by the Division City Science and Technology Innovation Talent Program under Grant KJ2023CX05, in part by the Research Start-Up Funds for High-Level Talents at Shihezi University under Grant RCZK2021B16, and in part by the National Natural Science Foundation of China under Grant 71971123.

ABSTRACT While modern deep learning methods have made significant progress in medical image segmentation, some challenges remain, including accurately capturing features at multiple scales, limited ability to detect critical regions, and susceptibility to noise and background interference. To address these challenges, a new neural network called HMSAM-UNet is introduced in this work. A novel module, the Hierarchical Multi-Scale Attention Module (HMSAM), was designed in HMSAM-UNet to improve the precision and accuracy of CT image segmentation significantly. Specifically, HMSAM integrates the Hierarchical Attention Mechanism and Inception Module via residual connections. The Hierarchical Attention Mechanism can highlight important regions by learning attention weights, dramatically enhancing the model's ability to perceive critical areas for more accurate localization and segmentation of target structures in CT images. Meanwhile, incorporating the Inception module effectively strengthens the network's capacity to capture multi-scale features, substantially improving the model's ability to comprehend the structural characteristics of CT images. The results show that the average loss achieved by the proposed model has a 50.04% reduction compared to the original U-Net architecture. Furthermore, compared to other deep learning models such as FCN, DeepLabV3, PSPNet, Unet, UNet++, and SegNet, the model proposed in this work attains an average Dice coefficient of 98.72, and an average IoU score of 97.46 on the three datasets, both of which are the highest among all compared models.

INDEX TERMS CT image segmentation, hierarchical attention mechanism, inception module, multi-scale, critical region perception.

I. INTRODUCTION

Accurate segmentation of lesion regions in CT images is crucial for diagnosis, treatment, and assessment of efficacy in medical image analysis. While exact results can be achieved through manual segmentation by experts, this approach is typically costly and requires many trained professionals for large-scale clinical applications [1]. Therefore, the development and application of automated medical image segmentation technology have become a hot spot in current research [2], [3], [4]. Compared with manual segmentation, automated segmentation technology can provide a reliable diagnostic process for large-scale clinical applications and effectively reduce the risk of human error. Therefore, there is

The associate editor coordinating the review of this manuscript and approving it for publication was Nuno M. Garcia.

an urgent need for an automated medical image segmentation technique in the current CT image segmentation field.

In recent years, deep learning methods, particularly convolutional neural networks (CNNs), have been widely applied in medical image segmentation, outperforming conventional algorithms for computer vision tasks and CT image segmentation. Fully Convolutional Networks (FCNs) are one of the most advanced techniques to generate pixel-level image labels at full resolution. In addition, many works have used FCN as a starting point to develop more profound and complex segmentation architectures such as the SegNet, FPN, PSPNet, U-net, and DeepLab families. These methods perform well in natural images but have some challenges in medical images [5], [6], [7].

Compared with natural images, CT images have some complexity, similar to medical images, but they still face some challenges in the segmentation task. First, CT images are acquired in a way that is quite different from natural pictures, usually by volume sampling of the subject's body. This sampling method may lead to many laminated structures, overlapping organs, and irregular shapes in the image, making it difficult to accurately recognize and segment the region of interest [8], [9]. Second, parts of interest in CT images usually consist of organs, lesions, or other medical entities, and their boundaries are often ambiguous. This ambiguity complicates the localization and boundary extraction of the region of interest, requiring the model to have strong perceptual ability and accurate location localization.

Meanwhile, the homogeneity of areas of interest in CT images, i.e., their diversity in shape, colour, size, etc., makes it more challenging for algorithms to identify and segment these regions. For example, specific lesions may present irregular shapes in the images, and organs may vary among individuals, which requires the model to be flexible enough to adapt to different feature representations. Furthermore, CT images often contain substantial noise and artifacts, which can impact the accuracy and reliability of segmenting the region of interest. To address this issue, dedicated image enhancement and denoising techniques are required to optimize the quality of medical images, thereby improving the model's recognition and segmentation of regions of interest.

U-Net, a commonly employed technique for image segmentation tasks, predicts pixel labels in images based on training data [10], [11], [12]. In the field of medical imaging specifically, U-Net has become one of the leading segmentation tools for various image modalities, including CT scans, MRI, X-rays, and microscopy. The method makes effective use of training data for image segmentation and generates high-quality segmentation maps even with limited labeled data.

Although U-Net performs well in medical image segmentation, it has some drawbacks and limitations. For example, when segmenting small objects in an image, U-Net is prone to detail loss and excessive smoothing, affecting accuracy, especially when dealing with small things in segmented images. To overcome these problems, researchers have proposed improved schemes such as U-Net++ and SegNet; however, they still need to overcome challenges such as high computational effort, memory footprint, and uneven category distribution. In addition, traditional deep learning methods such as FCN and DeepLabv3 also suffer from relatively weak processing of complexity in medical images [13], [14], [15]. Therefore, there is a critical need for new and improved solutions in the current medical image segmentation field to enhance the accuracy and efficiency of medical image segmentation tasks. Based on the preceding, this paper proposes an original convolutional neural network, HMSAM-UNet, with the goal of increasing the performance and precision of medical image segmentation. The overall layout of this paper is depicted in Figure 1.

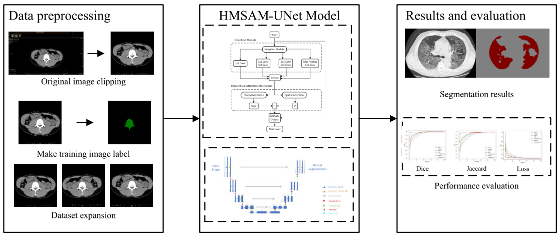


FIGURE 1. Overall Roadmap.

To validate the efficacy of the newly proposed HMSAM-UNet, three datasets - kidney stone, tuberculosis, and ultrasound - are generated in this paper. These datasets encompass various medical image segmentation challenges and possess authentic medical image data with high reference values. Leveraging these datasets, this paper compares and validates the new model against existing benchmark models, demonstrating the superior performance and efficiency of the proposed model. The main contributions of this paper are as follows:

(1) In this paper, a novel network model named HMSAM-UNet was designed, which demonstrates superior performance in CT image segmentation tasks with higher precision and accuracy than other modern deep learning networks. In addition, HMSAM-UNet also shows good robustness and can maintain excellent performance in different CT image segmentation tasks.

(2) A new module, HMSAM, is designed to innovatively integrate the Hierarchical Attention Mechanism and Inception model with the U-Net network to realize the effective fusion of multi-scale features. Through this design, the model can focus and emphasize more on the critical regions in CT images, thus effectively solving the problem of insufficient ability of the traditional model to perceive the key areas.

(3) The performance of the proposed HMSAM-UNet model is analyzed in comparison with other semantic segmentation models, including U-Net, U-Net++,

UNetr, COTR, SegNet, DeepLabV3, PSPNet and FCN. The results of the evaluation experiments demonstrate that the algorithm proposed in this work achieves higher efficiency and accuracy in medical image segmentation compared to the other methods.

The rest of this paper is organized as follows: Section II presents related work in the field of medical image segmentation. Section III presents three new datasets and the methodology for constructing the HMSAM-UNet model. Section IV outlines experimental evaluations designed to validate the effectiveness of HMSAM-UNet for medical image segmentation and provides discussion. Finally, Section V concludes the paper.

II. RELATED WORKS

Medical image segmentation has seen many noteworthy research results emerge in recent years. To present these research advances comprehensively, this paper will take a three-pronged approach. Firstly, this paper will introduce U-Net and its improved methods and explore their applications and performance in medical image segmentation tasks. Second, this paper will focus on enhanced networks for other classical deep learning networks and innovative models that improve segmentation accuracy and efficiency by introducing new mechanisms and strategies. Finally, this paper will explore the promise and potential of a creative research direction - novel deep learning-based network models for medical image segmentation. By categorizing these three aspects, readers will gain a more comprehensive and profound grasp of the current research status and future trajectory in medical imaging.segmentation [16], [17], [18], [19].

U-Net, as a classical deep-learning method in the field of medical image segmentation, is widely utilized for segmentation tasks involving CT images, MRI, and other medical modalities. In a recent study, Yin et al. [20] proposed a method named SD-UNet for the problem of segmenting lung infection regions in COVID-19 CT images. This method combines the self-attention mechanism (SA) and the dense space pyramid pooling module (Dense ASPP), which can fuse global and multiscale information and effectively solve challenges such as fuzzy boundaries and low contrast. In addition, Li et al. [21] investigated automatic liver and tumor segmentation in CT images and proposed a method called Hybrid Dense Connected U-Net (H-DenseUNet) in their paper. The technique combines 2D and 3D modules to extract features efficiently and aggregate contextual information. On the MICCAI 2017 dataset and the 3DIRCADb dataset, H-DenseUNet exhibits segmentation results superior to other methods. Additionally, Kushner et al. [22] probed the significance of automated liver and tumor segmentation in the clinical interpretation and treatment planning of hepatic diseases and put forward a multiscale approach. The method effectively improved the segmentation performance by enhancing the sensory field of the convolutional neural network (CNN). In a study by Geeta Rani et al. [23], the authors proposed a KUB-UNet model for solving the problem of semantic segmentation of kidneys, ureters, and bladder in KUB X-ray images. The model utilizes the ability of adaptive local receptive fields and feature reuse, which enables effective capture of specific details and structures in medical images, thus significantly improving the accuracy of segmentation of urinary system organs. However, as stated above, while these methods have made some headway on medical image segmentation tasks, they still encounter challenges when applied to complex and variable CT images. Specifically, they are prone to issues like loss of detail, fuzzy boundaries, and limited ability to identify critical regions, which impacts the accuracy of segmentation outcomes.

In addition to U-Net and its improved methods, other classical deep learning models, such as SegNet, PSPNet, FCN, and DeepLabV3, have also received extensive attention and improvements from researchers. These models introduce new strategies, such as null convolution, attention mechanism, feature fusion, etc., to improve segmentation accuracy and performance. In the paper by Xing et al. [24], in order to achieve accurate segmentation of medical images with different modalities, they proposed a deep learning-based automatic segmentation model called CM-SegNet. The model adopts multi-scale input and encoding-decoding ideas and consists of a multilayer perceptron and convolutional module, which can effectively extract global and local image information to realize the accurate segmentation task. The experimental results show that the CM-SegNet model outperforms other methods with better segmentation performance and shorter training time, which has high clinical application value. In another piece of work, Yamanakkanavar and Lee [25] introduced MF2-Net, a multipath feature fusion CNN encompassing multiple encoder paths to seize layer-specific multiscale information. Each encoder path employs a stacked asymmetric kernel module termed SGC to efficiently encode contextual particulars in high-level features and accurately conflate neighboring feature cues. At the bottleneck layer, the encoded elements are connected to capture the rich semantic features of the input image. Furthermore, the segmentation boundaries are honed at the decoder phase through the bootstrap block mechanism. Their experimental results demonstrate that the SGC module significantly improves segmentation accuracy, and MF2-Net outperforms existing methods on medical image segmentation tasks. By effectively conflating multiscale feature cues and honing segmentation margins, their proposed approach attains state-of-the-art performance. In a paper [26], Srivastava et al. proposed a Multiscale Residual Fusion Network called MSRF-Net specifically for medical image segmentation. This network can efficiently handle the segmentation task for objects of different sizes and is especially suitable for small biased datasets. MSRF-Net employs a Dual-Scale Dense Fusion (DSDF) block for exchanging multiscale features of other receptive fields for multiscale fusion while preserving the resolution, improving the information transfer and being able to propagate high and low-level features, thus achieving accurate segmentation results. Extensive experiments have shown that MSRF-Net outperforms other methods on several medical datasets and obtains improved results. Cheng et al. [27] proposed an efficient and precise expansive fully convolutional network (FCN) for biomedical image segmentation termed Fully Convolutional Attention Network (FCANet). FCANet combines two attentional modules to aggregate long-range and short-range contextual information. The spatial attention module can aggregate features at each location, thus spatially promoting similar characteristics. The channel attention module, on the other hand, emphasizes the channel dependency between any two channels. FCANet integrates two attention modules to weight and combine their output features. This improves feature representation and segmentation accuracy for biomedical images. Experiments demonstrate FCANet's ability to substantially enhance biomedical image segmentation performance. However, despite these enhanced models boosting the accuracy and efficiency of medical image segmentation to some degree, they still have some performance and robustness constraints when confronting intricate and variable medical images. In particular, these models may need to perform better when dealing with images with solid interference or noise or when facing unbalanced datasets with few categories and samples, resulting in imprecise segmentation boundaries or mis-segmentation. Furthermore, existing models need improvement in perceiving critical regions to better capture intricate details and structures in medical images. This can enhance segmentation accuracy and robustness. Advancing medical image segmentation requires continued optimization of deep learning models to address the growing complexity and diversity of medical images.

In recent times, Transformer-centered approaches have garnered substantial attention for medical image segmentation, with mounting researcher interest. Capitalizing on the triumph of Transformers in natural language processing, these methodologies exhibit formidable global feature modeling capabilities for medical images. For example, Ma et al. [28] proposed a Hierarchical Contextualized Attention Transformer Network (HT-Net) to achieve excellent performance in medical CT image processing, which effectively learns the relationship between remote pixels and captures rich semantic information through the fusion of multiscale and Transformer. Meanwhile, the algorithm of Pan et al. [29] successfully combines a U-shaped Convolutional Neural Network (CNN) with a Visual Transformer (VIT) in multi-organ segmentation of male pelvic CT images, which facilitates long-range dependency through a self-attentive mechanism to achieve the delivery and prediction of high-resolution feature maps. However, while Transformer-based methods have made significant progress in global feature modelling, they still need to improve in perceiving details and local information. Given the large number of complex structures and subtle features in CT images, local information is crucial for accurate segmentation results. Moving forward, an important research direction involves developing new network architectures that can better fuse local and global knowledge. This will enable more precise detail perception and improved segmentation performance for medical images.

Based on the above, the primary objective of this paper is to enhance the precision and robustness of CT image segmentation, with a particular emphasis on approaches to refine several classical and innovative models. However, these methods have limitations in discerning details and local information, especially when facing complex and variable CT images. To conquer these challenges, this study put forward a network model called HMSAM-UNet, which dramatically strengthens the capacity to glean insights from critical regions while maintaining excellent performance and robustness, thus demonstrating superb competency in CT image segmentation tasks.

III. MATERIALS AND METHODS

A. HMSAM-UNET

The current U-Net model has a particular application basis within medical image segmentation. Limited by its sensory area and the problem of segmentation accuracy, its performance in some specific application scenarios could be more satisfactory. To address the limitations of the U-Net model, this work proposes a new network, HMSAM-UNet, a novel CNN based on the U-Net architecture that integrates the hierarchical attention mechanism and Inception module.

B. INTRODUCTION TO U-NET

U-Net networks are characterized by symmetric encoder and decoder structures for pixel-level semantic segmentation. The encoder can progressively reduce the image size and extract feature representations at different scales, while the decoder can restore the feature map to its original dimensions by upsampling, thereby enabling pixel-level semantic segmentation.

The U-Net network has been extensively applied in medical image segmentation, for instance, in segmenting organs such as liver, lung and heart. Compared with the traditional manual feature extraction and segmentation methods, the U-Net network can automatically learn the features and patterns in medical images with higher segmentation accuracy and better robustness. In addition, the U-Net network can handle medical images with multiple channels (e.g., multi-sequence MRI) and improve the segmentation effect by combining various information.

C. MODULE INTRODUCTION

1) HIERARCHICAL ATTENTION MECHANISM

The Hierarchical Attention Mechanism (HAM) is a critical technique introduced in deep learning models to enhance the model's attention to the input data. The mechanism aims to implement a cascading attention architecture that selectively focuses on crucial information in the input data through a hierarchical structure from global to local. At each hierarchy, the adaptive weighting of different features is achieved by learning parameterized attention weights, calculating the importance distribution of the input data, and applying the consequences to the input feature map. Subsequently, the weighted features from different layers are fused to obtain a more discriminative feature representation.

Recent studies shed light on the efficacy of hierarchical attention mechanisms in various domains. For instance, in medical image segmentation, the work by Ding et al. [30] introduces the Hierarchical Attention Network (HANet) for this purpose. This approach presents an innovative method that reconstructs the self-attention mechanism from the perspective of high-order graphs. By embedding a Hierarchical Attention (HA) module within HANet, context information from multiple levels of neighbours is captured, effectively mitigating noise and enhancing segmentation accuracy, as demonstrated in experiments across tasks such as optic disc/cup segmentation, vessel segmentation, and lung segmentation on medical datasets.

Similarly, Wang et al. [31] propose a Hierarchical Attention Network (HANet) in image captioning. This network synchronously computes attention over features at different semantic levels, enabling the prediction of other words based on distinct features. The Multimodal Residual Module (MRM) is introduced to learn joint representations from various modalities, facilitating adaptability to diverse image contexts. Experiments on the MSCOCO dataset validate the effectiveness of HANet, surpassing state-of-the-art methods in terms of BLEU and CIDEr scores.

These studies collectively underscore the adaptability and robustness of hierarchical attention mechanisms across different domains. The hierarchical attention mechanism's ability to dynamically allocate attention based on varying features enables models to adapt to diverse contexts effectively. This adaptability is crucial for enhancing model generalization and robustness, as evidenced by experimental validations across different scenarios and tasks. Thus, the hierarchical attention mechanism improves performance and enhances the model's ability to handle changes in various environments, ultimately contributing to its applicability and efficacy.

2) INCEPTION MODEL

The Inception module is a means of efficiently extracting features and reducing the number of parameters. The module is specifically designed and implemented as follows.

Firstly, the Inception module employs a multi-branch network structure, where each branch uses a different-sized convolutional kernel to capture multi-scale information in the image. This is consistent with how the human eye perceives scene information in other sensory domains. Larger convolutional seeds extract global features, while smaller seeds are better at learning local features. The multi-branch structure allows for both international and regional features. Second, a 1 × 1 convolutional kernel is introduced in the module for channel compression. This reduces the number of parameters on one hand and accelerates the computation process on the other. After the channel compression, 3 × 3 and 5 × 5 convolutions are added, which act as a channel decomposition and improve the expressive power of the model.

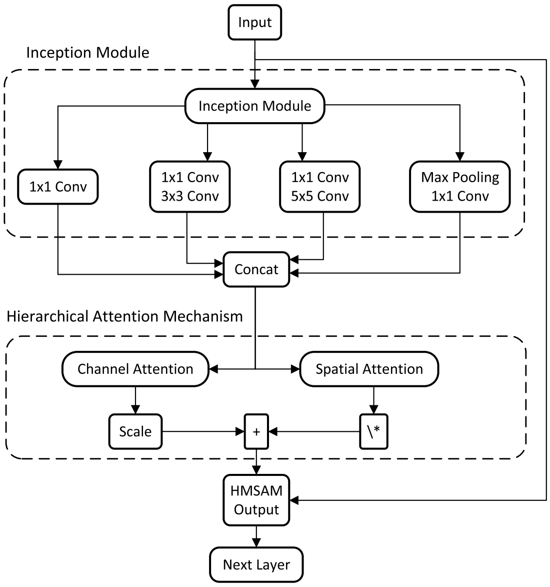


FIGURE 2. HMSAM structure diagram.

Again, the maximum pooling branch enhances the stability and robustness of the model to geometric transformations of the image. Pooling allows features to remain invariant to position and rotation transformations.

Finally, the width of each branch of the Inception module can be customised. The structure is robust and easily scalable. The branch width can be adjusted according to the task requirements and computational resources to achieve efficient feature learning.

Overall, the Inception module extends the width and depth of the network with the same number of parameters and improves the model's feature expressiveness and computational efficiency through multi-scale convolution, channel decomposition, compression, pooling and other technical means.

D. HMSAM STRUCTURE

The HMSAM module consists of two main components: the multiscale Inception module and the Hierarchical Attention Mechanism. Its structure is shown in Figure 2. The multiscale Inception module utilizes four parallel path structures to extract different scale features: the first path is a 1 × 1 convolution to capture spatial correlation; the second path is a 1 × 1 convolution followed by a 3 × 3 convolution to obtain local parts; the third path uses a 1 × 1 convolution followed

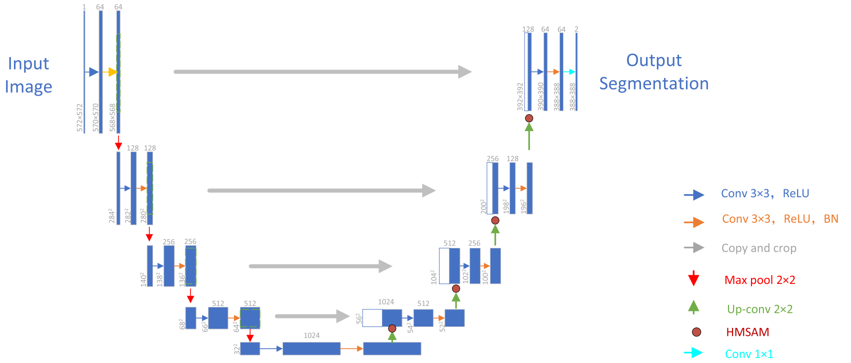


FIGURE 3. HMSAM-UNet structure diagram.

by a 5 × 5 convolution to aggregate elements from a broader region; and the last path undergoes max pooling followed by a 1 × 1 convolution to obtain high-level semantic features. All four paths apply the ReLU activation function.

After obtaining the features from the four paths, they are concatenated as the output of the Inception module. This produces a feature representation that incorporates different scale features.

Next, the feature map is fed into the hierarchical attention mechanism. Here, 1 × 1 convolution is utilized to generate spatial attention weights and dot products with the feature map to obtain spatial attention features. Concurrently, channel attention weights are also produced using global average pooling for channel scaling. Finally, the spatial attention features and channel attention features are summed to acquire the feature representation with fusion attention.

The final output of the HMSAM module is the residual concatenated features, i.e., the above-described attention features are added to the module's input features. This further enhances the feature representation.

E. HMSAM-UNET STRUCTURE

HMSAM-UNet is a fully convolutional network (FCN) designed with an encoder-decoder architecture for the semantic segmentation of medical images. The entire network can be divided into the encoder part and the decoder part. Its structure is shown in Figure 3.

The encoder part of the network uses convolutional layers, batch normalization, activation functions, and max pooling for feature extraction. In order to enhance the feature representation capability of the encoder, the network introduces a newly proposed HMSAM module after each downsampling layer, which first utilizes a multi-scale Inception structure to capture feature information at different scales. The HMSAM module first captures feature information at different scales using a multi-scale Inception structure. Then, it incorporates a hierarchical attention mechanism into the Inception output to enhance the feature representation of critical regions by learning the attention weights. Finally, the production of the HMSAM module is channel-adjusted by 1 × 1 convolution and combined with the output of the encoder through residual concatenation to form an enhanced feature representation.

In the decoder part, the network uses upsampling to recover the spatial size of the encoder output. At the same time, the feature maps of corresponding scales in the encoder are concatenated to the upsampled features through skip connections so that the decoder can utilize richer feature information. After the decoder gradually restores the spatial resolution, the last layer outputs the category prediction of each pixel through 1 × 1 convolution to complete the image segmentation.

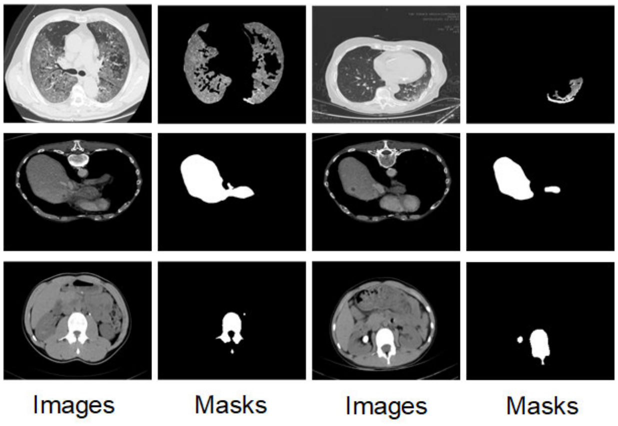


FIGURE 4. Example of a dataset.

IV. EXPERIMENTS AND RESULTS

A. EXPERIMENT SETTINGS

1) IMAGE DATASET

This paper used three image datasets for model training, which are lung, liver, and kidney stone CT image datasets. Specific examples are shown in Figure 4.

2) TRAINING, VALIDATION, AND TEST DATA SETUP

For the lung CT dataset, 2100 images were used for training and 500 for testing. The liver CT dataset consisted of 400 training and 100 test images. Similarly, the stone CT dataset had 400 training images and 100 test images. Data augmentation through rotation, noise injection, cropping etc. was applied to the original and labeled images to increase diversity and volume, improving generalization. During training, a validation set enabled model adjustment and optimization to ensure effectiveness. The multi-dataset approach with augmentation and validation helps improve and verify the HMSAM-UNet model's segmentation capability.

To summarize, data augmentation and validation sets were utilized during HMSAM-UNet's training process. The training set updated model parameters, the validation set tuned hyperparameters and monitored performance, and the test set evaluated final generalization capability. This approach of partitioning the data into training, validation, and testing sets facilitated efficient model training, tuning, and performance appraisal. The proof set, in particular, assisted optimization and verification of the model by furnishing an unbiased means to analyze hyperparameters and track progress discrete from the test set. This rigorous methodology ensures HMSAM-UNet is properly trained and validated for optimal medical image segmentation [32], [33], [34].

3) LAB ENVIRONMENT

In this study, NVIDIA GeForce RTX 3080 Ti GPU was used as the deep learning computational gas pedal and CUDA 11.2 was installed as the parallel computing platform and programming model. In addition, Python 3.8 was chosen as the primary programming language, and TensorFlow 2.9.0 was installed as the deep learning framework for implementing various neural network models for training and inference. Together, these components build an experimental environment that provides mighty computing power and a flexible programming environment for deep learning research, thus making it possible to conduct complex and meaningful experiments. Such a configuration enables the performance and effectiveness of different deep learning algorithms on diverse tasks to be explored in experiments [35], [36], [37].

4) HYPERPARAMETERS

This paper proposes a new loss function called Boundary Dice Loss for CT image segmentation. It comprises three components: Categorical Cross-Entropy Loss, Dice Loss, and Boundary Loss. Collectively, these facilitate pondering the similarity between predicted and ground truth segmentations in conjunction with boundary constraints to refine accuracy. By holistically accounting for segmentation similarity and boundaries, Boundary Dice Loss allows improving model segmentation performance on CT images. The multi-component design provides comprehensive optimization to enhance segmentation precision.

To elucidate, Categorical Cross-Entropy Loss gauges the disparity between the model's predicted outputs and the ground truth segmentation labels for the multi-class segmentation undertaking. It computes the cross-entropy loss for semantic segmentation. Its formula is as Equation 1.

TABLE 1. Confusion matrix example.

	Positive	Negative
Predicted	*TP*	*FP*
Negative	*FN*	*TN*

$$Categorical\_Cross - EntropyLoss = -\sum_{c=1}^{C} y_{\text{true}}^{(c)} \log(y_{\text{pred}}^{(c)}) \tag{1}$$

where C is the total number of categories, $y_{true}^{(c)}$ is the actual binary split pixel value (0 or 1) of sort c in the label, and $y_{pred}^{(c)}$ is the pixel value of category c in the predicted output of the model (taking matters between 0 and 1).

Dice Loss utilizes the Dice Coefficient to measure similarity between the predicted and ground truth segmentations. It provides a loss function based on this common segmentation evaluation metric. Its formula is as Equation 2.

$$DiceLoss = 1.0 - \frac{2 \cdot \text{intersection} + \text{smooth}}{\text{union} + \text{soft}} \tag{2}$$

where corner is the intersection of the predicted segmentation and the accurate segmentation; union is the concatenation of the predicted segmentation and the accurate segmentation; smooth is a smoothing term, usually taken as a minimal number, such as $1e^{-7}$, to prevent the denominator from being zero.

Boundary Loss constrains the similarity between the boundaries of the predicted and actual segmentations. It achieves this by calculating the difference in boundary gradients. Its formula is as Equation 3.

$$BoundaryLoss = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \left|\left| \nabla y_{\text{true}}^{(ij)} - \nabla y_{\text{pred}}^{(ij)} \right|\right|_1 \tag{3}$$

where H and W are the height and width of the image, respectively; $y_{true}^{(ij)}$ and $y_{pred}^{(ij)}$ are the pixel values in the actual boundary and the predicted boundary of the model, respectively; ∇ denotes the gradient operator; $||\cdot||_1$ indicates the L1 paradigm, which is used to compute the absolute sum of the gradient differences.

In addition, this uses the Adam optimizer that can dynamically adjust the learning rate to 3×10^{-4} during the training process and sets the batch size to 32. Considering the computational performance of the workstation, this paper finally sets 100 rounds for training.

B. EVALUATION INDEX

To evaluate the proposed model's performance, this work utilizes several common segmentation metrics: Dice coefficient, Jaccard index, Precision, Recall, F1 score, and Pixel Accuracy. Together these provide a comprehensive quantitative assessment. The confusion matrix involved in this paper is shown in Table 1.

The Dice coefficient measures the similarity between two sets and is widely used to evaluate segmentation. For binary tasks, it assesses the overlap between predictions and ground truth labels. It provides a quantitative measure of how well the segmentations match. The Dice coefficient is calculated using the Equation 4.

$$\text{Dice} = \frac{2 \cdot |A \cap B|}{|A| + |B|} \tag{4}$$

where A is the set of predicted results, B is the set of proper labels, $|A|$ denotes the number of elements in set A, $|B|$ represents the number of factors in set B, and $|A \cap B|$ indicates the number of elements in the intersection of A and B.

Jaccard Index (also known as Intersection over Union, IoU): Similar to Dice coefficient, the Jaccard index measures similarity between two sets and is commonly applied in segmentation and detection tasks. For binary problems, it assesses the similarity between predicted and true labels. It provides another metric to quantitatively evaluate the correspondence between segmentations. The formula for the Jaccard index is Equation 5.

$$\text{Jaccard} = \frac{|A \cap B|}{|A \cup B|} \tag{5}$$

where $|A \cup B|$ denotes the number of elements of the concatenated set of A and B.

Precision measures the proportion of predicted positive samples that are actually positive. In other words, it calculates the percentage of samples the model predicts as positive that are genuinely positive. Precision evaluates how many of the model's optimistic predictions are correct. The formula for Precision is Equation 6.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

Recall measures the proportion of valid positive samples that are correctly predicted as positive. In other words, it calculates the percentage of positive examples that the model correctly identifies out of all actual positive cases. Recall evaluates how many relevant instances are captured in the predictions. Memory is calculated using Equation 7.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

Accuracy measures the overall proportion of correctly classified samples out of all samples. It evaluates the total predictive accuracy of a classification model. Accuracy is computed as Equation 8.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{8}$$

C. EXPERIMENTAL RESULTS

Figure 5 shows the training results for the three datasets. Observing the figure, when using the lung lesion region segmentation dataset, the prediction results of U-Net++, UNetr, SegNet, DeepLabV3, and PSPNet can roughly describe the contour range of the lesion region. However, the prediction results of these models have some bias and inaccuracy compared with the actual areas.

Especially in the edge part of some delicate or complex lesions, the segmentation effect of these models could be better; FCN performs more poorly, and its prediction results lose a lot of detailed information, which leads to many delicate areas of lesion regions being ignored. On the other hand, the segmentation results of U-Net and COTR are relatively good and can have a high degree of overlap with the natural region. Especially, COTR has better detail information grasping ability than U-Net, making it slightly better in segmentation results. However, there needs to be more clarity in the boundary segmentation of U-Net and COTR, and effective boundary recognition is impossible for some large segmented regions. In contrast, HMSAM-UNet has the most apparent segmentation of the boundary area while ensuring the ability to acquire detailed detail information, and this difference is more evident in those other two datasets. Overall, the overall segmentation area of HMSAM-UNet is closer to the actual situation and has a higher practical application value.

D. MODEL EVALUATION

This paper evaluates the proposed model and six comparative models, and the results are shown in Figure 6 and Table 2.

Firstly, the training process of the seven models was comprehensively analyzed, including the variation of Dice coefficients, IoU coefficients, and loss functions. Figure 6 shows the interpretation of these training metrics.

From the figure 6, it can be learned that COTR, FCN, PSPNet and DeepLabV3 showed a gradient explosion phenomenon when training with Dataset 1. This is because these two models encountered too many parameters during the training process, which caused the gradient values to become abnormally large, thus causing the model's parameters to deviate from the optimal solution and leading to unstable training. In contrast, the other models maintained a normal training state during the training process without the problem of gradient explosion. In addition, U-Net and HMSAM-UNet show better robustness throughout the training process. They have nearly the same convergence speed while keeping the training curve smooth. This indicates that U-Net and HMSAM-UNet can perform well for different training data and parameter settings, with strong adaptability and generalization ability.

Table 2 shows the training results of each model on the validation set, and the results show that HMSAM-UNet scores better than other models on the vast majority of evaluation functions.

The experimental results show that the HMSAM-UNet model achieves the lowest Loss coefficient among all the comparative models, with an average Loss of 7.34 in 3 training sessions, which is 50.04% lower than the average Loss

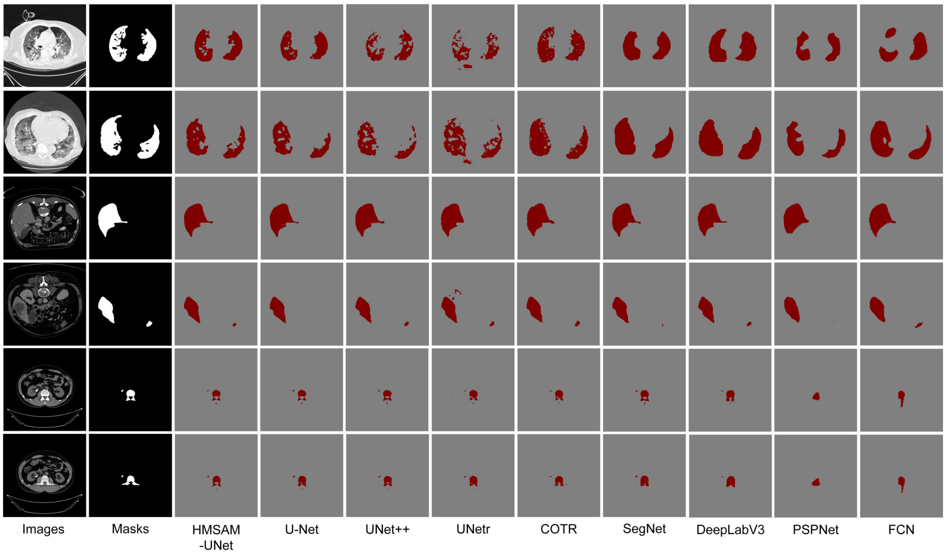


FIGURE 5. Segmented image.

TABLE 2. Test set results.

		Indicator(%)					
	Model	Loss	Dice	IoU	Precision	Recall	Accuracy
Dataset 1	HMSAM-UNet	1.99	98.88	97.78	99.44	99.72	99.44
	U-Net	6.25	95.03	90.53	99.36	99.68	99.35
	U-Net++	5.65	95.46	91.32	99.36	99.68	99.36
	UNetr	3.82	97.45	95.02	99.07	99.53	99.06
	COTR	2.02	98.69	98.01	99.22	99.61	99.22
	SegNet	6.72	94.58	89.73	99.19	99.59	99.15
	DeepLabV3	3.08	98.31	96.69	98.91	99.46	98.91
	PSPNet	1.62	99.17	98.35	99.39	99.69	99.38
	FCN	2.16	99.13	98.29	99.26	99.63	99.26
Dataset 2	HMSAM-UNet	6.40	98.49	97.02	99.44	99.72	99.44
	U-Net	24.79	95.81	91.95	99.36	99.68	99.35
	U-Net++	10.34	97.82	95.74	99.36	99.68	99.36
	UNetr	9.31	96.78	93.77	99.07	99.53	99.06
	COTR	6.50	97.81	96.65	99.22	99.61	99.22
	SegNet	10.78	97.26	94.67	99.19	99.59	99.15
	DeepLabV3	7.54	98.30	96.44	98.91	99.46	98.91
	PSPNet	7.46	98.33	96.71	99.39	99.69	99.38
	FCN	9.03	98.47	97.00	99.26	99.63	99.26
Dataset 3	HMSAM-UNet	13.62	98.78	97.59	97.59	98.86	99.43
	U-Net	14.77	98.63	97.29	97.29	98.85	99.43
	U-Net++	15.08	98.66	97.35	97.35	98.85	99.42
	UNetr	15.25	98.71	97.45	97.65	98.82	99.41
	COTR	17.90	98.71	97.54	97.64	98.82	99.41
	SegNet	16.95	98.59	97.22	97.22	98.76	99.38
	DeepLabV3	17.89	98.70	97.43	97.43	98.71	99.35
	PSPNet	16.07	98.66	97.35	97.35	98.71	99.36
	FCN	17.79	98.52	97.09	97.09	98.54	99.27

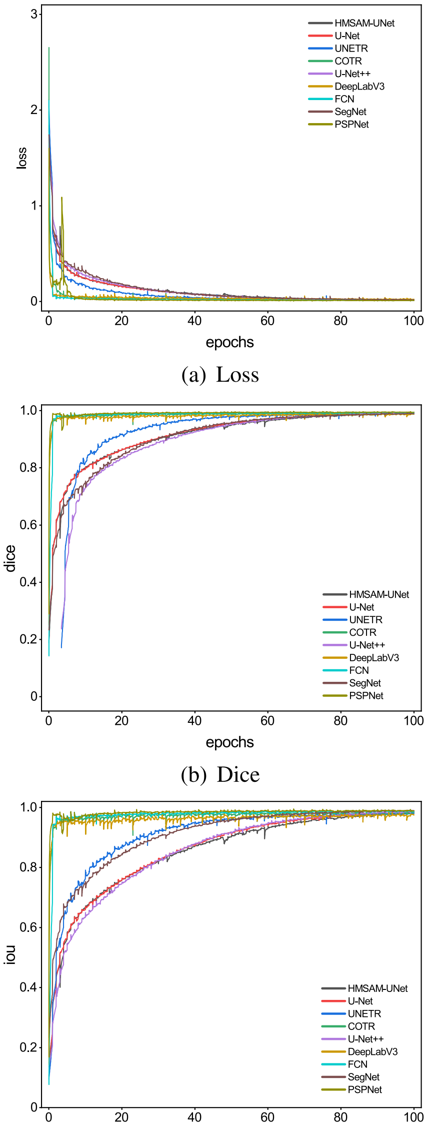
of 15.27 of U-Net. The Dice coefficient of HMSAM-UNet is also higher than the other models, with 98.72 in 3 training sessions, while the average Dice coefficient of U-Net is 96.49, which is 2.31% higher than the other models. Regarding the IoU coefficient, HMSAM-UNet has an average score of 97.46, maintaining the highest IoU coefficient in all three training sets. Regarding moderate accuracy, HMSAM-UNet scored close to COTR in dataset 1 while keeping the highest score in all other datasets, with an average score of 98.82, which is an improvement of 0.47 compared to U-Net. In addition, HMSAM-UNet also performs well in terms of average recall and accuracy, which are 99.43 and 99.44, an improvement of 0.03 and 0.06, respectively. Overall, HMSAM-UNet shows better results in segmentation tasks, indicating that it can effectively handle different datasets and charges.

E. HAM EVALUATION EXPERIMENT

This experiment compares two deep learning models, the traditional U-Net and the U-Net, with the introduction of the Hierarchical Multiscale Attention Mechanism (HAM) (UNet-HAM). Both models were tested on three different

TABLE 3. HAM Evaluation Experiment.

	Dataset 1			Dataset 2			Dataset 3		
Model	Loss	Dice	IoU	Loss	Dice	IoU	Loss	Dice	IoU
U-Net	6.25	95.03	90.53	24.79	95.81	91.95	14.77	98.63	97.29
UNet-HAM	4.29	98.56	93.16	19.75	98.95	91.61	15.56	98.63	97.30



(a) Loss

(b) Dice

(c) IoU

FIGURE 6. Changes in training process indicators in Dataset 1.

datasets to evaluate their performance fully. As shown in Table 3, the experimental results show that UNet-HAM outperforms the traditional U-Net on either dataset. Specifically, UNet-HAM significantly improves two important performance metrics, Dice and IoU. At the same time, UNet-HAM has a lower loss value than U-Net. This result verifies that HAM can effectively improve the model's performance and adapt it to different data distributions better.

TABLE 4. Ablation experiment.

Model	HAM	Inception	RC	Loss	Dice	IoU
U-Net				24.79	95.81	91.95
UNet-Inception		✓		10.97	97.46	95.03
UNet-Inception-HAM	✓	✓		9.25	98.01	96.19
HMSAM-UNet	✓	✓	✓	6.40	98.49	97.02

F. ABLATION EXPERIMENT

To further assess the impact of introducing different modules (e.g., Inception module, Hierarchical Attention Mechanism, and Residual Connection) into the U-Net model on the model performance, an ablation experiment is designed in this paper. The ablation experiment uses a modified U-Net model, HMSAM-UNet, as the baseline model. This model integrates the HMSAM module in each downsampling layer of the encoder, which contains the Inception module, the Hierarchical Attention Mechanism, and the Residual Connection.

This research set up two comparison models to evaluate the effect of these new modules individually and in combination. The first comparison model, UNet-Inception, is a regular U-Net model with the Inception module added after a regular convolutional layer. The second comparison model, UNet-Inception-HAM, adds an Attention Mechanism module to UNet-Inception. All four models were validated on Dataset 2.

The experimental results are shown in Table 4, where the HMSAM-UNet model outperforms the other three models in all evaluation metrics. This indicates that the simultaneous introduction of the Inception module, the hierarchical attention mechanism and residual concatenation can significantly improve the model performance, and the fact that the UNet-Inception-HAM model outperforms the UNet-Inception model confirms that the introduction of the hierarchical attention mechanism alone can also bring about a certain degree of performance improvement. On the other hand, the plain U-Net model performs the worst. The results of this ablation experiment validate that introducing these novel modules (e.g., Inception, Attention Mechanism, Residual Connection, etc.) into the U-Net model helps improve the model performance and that a combination of these modules works best.

G. MODEL TRAINING TIME EVALUATION

Table 5 shows the training time of each model, and the results show that HMSAM-UNet has a faster training speed than other models. This short training advantage is mainly because HMSAM-UNet adopts feature splicing in the decoder part, which is used to realize the fusion of multi-

TABLE 5. Model inference times.

Model	Time(s)	Model	Time(s)	Model	Time(s)
HMSAM-UNet	2183	U-Net	2343	U-Net++	2263
UNetr	2713	COTR	2643	SegNet	2216
DeepLabV3	3340	PSPNet	2238	FCN	2956

scale information. Compared to the fully connected or fully convolutional layer approach, the feature splicing approach has less computational effort, which is conducive to speeding up the training process. Overall, HMSAM-UNet has high practicality.

V. DISCUSSION

In this study, an innovative model, named HMSAM-UNet, is designed to enhance the model's ability to capture multiscale features and focus on critical regions by introducing the Hierarchical Multiscale Attention Module (HMSAM), which incorporates the Inception module, which acquires feature representations at different scales in parallel, and the hierarchical attention mechanism, which focuses on critical regions through adaptive weighting. The latter focuses the model on important areas of the input data through adaptive weighting. This design enables HMSAM-UNet to efficiently extract multi-resolution semantic and detailed information from CT images and focus on key structures such as lesions and organs, thus achieving more accurate segmentation.

Compared with the original U-Net, the experimental results show that HMSAM-UNet reduces the average loss value on the three datasets by 50.04%, mainly attributed to the multi-faceted optimisation design. Firstly, the multiscale feature fusion of HMSAM makes the feature representation richer, and the model can portray the target structure more accurately; secondly, the attention mechanism enables the model to adaptively focus on the key regions, which reduces the risk of misclassification; furthermore, the structure of 1 × 1 convolution and residual linkage improves the effectiveness of the model and avoids information loss and gradient problems. Together, these innovative designs enhance the performance of HMSAM-UNet, enabling it to excel in processing complex, detail-rich CT images and in segmenting clear boundaries and small critical regions, which are essential for clinical diagnosis and treatment planning.

In addition, HMSAM-UNet shows good adaptability and generalisation ability, achieving the best segmentation accuracy on different datasets (e.g., lung, liver, kidney stones, etc.), which proves that the model is promising to be applied to a wide range of medical image segmentation scenarios, e.g., lesion and thorax segmentation in CT images of the lungs, organ segmentation in CT images, and renal stone segmentation in CT images of the kidneys.

However, despite the excellent achievements of HMSAM-UNet, there is still room for further improvement in future iterations. For example, introducing more diverse data can improve the model's generalisation. In addition, increasing the interpretability of the model and reducing the number of parameters are also directions that need to be worked on to improve the interpretability and deployment friendliness of the model.

When extending HMSAM-UNet to other medical imaging domains (e.g., MRI, ultrasound, etc.), it is necessary to focus on the special challenges of different imaging modalities. For example, MRI often suffers from strong noise and offset fields, while ultrasound images are characterised by significant speckle noise and low contrast. To address these challenges, the preprocessing, loss function and other aspects of HMSAM-UNet may need to be specially designed to ensure good segmentation performance. Different imaging modalities typically require stronger attentional focus and more detailed feature capture capabilities than CT images.

VI. CONCLUSION

To enhance the precision of CT image segmentation, this paper puts forward a new convolutional neural network model named HMSAM-UNet. The unique characteristic of HMSAM-UNet is the design of a new module, the Hierarchical Multi-Scale Attention Module, which ingeniously combines the Hierarchical Attention Mechanism and the Inception module to actualize multi-scale feature fusion, which dramatically refines the model's segmentation effect on the critical regions. By bringing in the layer attention mechanism, HMSAM-UNet can adaptively concentrate on features at different levels. This enables the model to seize better details and structures at various scales in medical images. In the meantime, by employing the Inception module, HMSAM-UNet can extract features efficiently, effectively reducing the number of parameters of the model and bettering the computational efficiency. To verify the performance of HMSAM-UNet, a series of evaluation experiments are conducted in this paper, compared with the other six models. The experimental results demonstrate that HMSAM-UNet has made significant advancements in CT image segmentation tasks and is markedly superior to other methods. Its ability to accurately segment complex CT images is impressive and brings new prospects to medical image segmentation.

Although HMSAM-UNet shows superior performance in simulation experiments, it still needs further improvement and validation due to the limitations of research capabilities and environmental conditions. In this paper, HMSAM-UNet mainly focuses on scenarios applicable to CT image segmentation and does not fully consider other types of medical images. Future research plans include further optimizing HMSAM-UNet for a broader range of medical images and further enhancing the utility and adaptability of the model. Through continuous efforts and improvements, HMSAM-UNet is expected to become an essential innovation in medical image segmentation, providing more reliable and efficient solutions for medical image diagnosis and treatment.

REFERENCES

[1] S. Mishra, H. K. Tripathy, and B. Acharya, "A precise analysis of deep learning for medical image processing," *Bio-Inspired Neurocomput.*, vol. 903, pp. 25–41, Jul. 2020. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-15-5495-7_2#citeas

[2] M. Kaur, S. Sofat, and D. K. Chouhan, "Review of automated segmentation approaches for knee images," *IET Image Process.*, vol. 15, no. 2, pp. 302–324, Feb. 2021.

[3] C. Kaur and U. Garg, "Artificial intelligence techniques for cancer detection in medical image processing: A review," *Mater. Today, Proc.*, vol. 81, pp. 806–809, Jan. 2023.

[4] M.-H. Sheu, S. M. S. Morsalin, S.-H. Wang, L.-K. Wei, S.-C. Hsia, and C.-Y. Chang, "FHI-UNet: Faster heterogeneous images semantic segmentation design and edge AI implementation for visible and thermal images processing," *IEEE Access*, vol. 10, pp. 18596–18607, 2022.

[5] B. N. Kumar, T. R. Mahesh, G. Geetha, and S. Guluwadi, "Redefining retinal lesion segmentation: A quantum leap with DL-UNet enhanced auto encoder–decoder for fundus image analysis," *IEEE Access*, vol. 11, pp. 70853–70864, 2023.

[6] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "DS-TransUNet: Dual swin transformer U-Net for medical image segmentation," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–15, 2022.

[7] B. Chen, Y. Liu, Z. Zhang, G. Lu, and A. W. K. Kong, "TransAttUNet: Multi-level attention-guided U-Net with transformer for medical image segmentation," 2021, *arXiv:2107.05274*.

[8] S.-T. Tran, M.-H. Nguyen, H.-P. Dang, and T.-T. Nguyen, "Automatic polyp segmentation using modified recurrent residual UNet network," *IEEE Access*, vol. 10, pp. 65951–65961, 2022.

[9] S. Jain, T. V. Vyvere, V. Terzopoulos, D. M. Sima, E. Roura, A. Maas, G. Wilms, and J. Verheyden, "Automatic quantification of computed tomography features in acute traumatic brain injury," *J. Neurotrauma*, vol. 36, no. 11, pp. 1794–1803, Jun. 2019.

[10] X. Liu, L. Song, S. Liu, and Y. Zhang, "A review of deep-learning-based medical image segmentation methods," *Sustainability*, vol. 13, no. 3, p. 1224, Jan. 2021.

[11] A. Iqbal, M. Sharif, M. Yasmin, M. Raza, and S. Aftab, "Generative adversarial networks and its applications in the biomedical image segmentation: A comprehensive survey," *Int. J. Multimedia Inf. Retr.*, vol. 11, no. 3, pp. 333–368, Sep. 2022.

[12] G. Mahalaxmi, T. Tirupal, S. Shanawaz, S. Swarnakar, and S. V. Krishna, "A comparison and survey on brain tumour detection techniques using MRI images," *Current Signal Transduction Therapy*, vol. 18, no. 1, pp. 14–23, Mar. 2023.

[13] H. Liu, H. Wang, Y. Wu, and L. Xing, "Superpixel region merging based on deep network for medical image segmentation," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 4, pp. 1–22, Aug. 2020.

[14] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "DoubleU-Net: A deep convolutional neural network for medical image segmentation," in *Proc. IEEE 33rd Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jul. 2020, pp. 558–564.

[15] M. Diwakar, P. Sharma, S. Swarnakar, and P. Kumar, "Image security using cellular automata rules," in *Proc. 3rd Int. Conf. Soft Comput. Problem Solving*, vol. 1. New Delhi, India: Springer, 2014, pp. 403–412. [Online]. Available: https://link.springer.com/chapter/10.1007/978-81-322-1771-8_35#citeas

[16] M. A. Abdou, "Literature review: Efficient deep neural networks techniques for medical image analysis," *Neural Comput. Appl.*, vol. 34, no. 8, pp. 5791–5812, Apr. 2022.

[17] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical image analysis using convolutional neural networks: A review," *J. Med. Syst.*, vol. 42, no. 11, pp. 1–13, Nov. 2018.

[18] H. Yu, L. T. Yang, Q. Zhang, D. Armstrong, and M. J. Deen, "Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives," *Neurocomputing*, vol. 444, pp. 92–110, Jul. 2021.

[19] S. Suganyadevi, V. Seethalakshmi, and K. Balasamy, "A review on deep learning in medical image analysis," *Int. J. Multimedia Inf. Retr.*, vol. 11, no. 1, pp. 19–38, Mar. 2022.

[20] S. Yin, H. Deng, Z. Xu, Q. Zhu, and J. Cheng, "SD-UNet: A novel segmentation framework for CT images of lung infections," *Electronics*, vol. 11, no. 1, p. 130, Jan. 2022.

[21] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[22] D. T. Kushnure and S. N. Talbar, "MS-UNet: A multi-scale UNet with feature recalibration approach for automatic liver and tumor segmentation in CT images," *Computerized Med. Imag. Graph.*, vol. 89, Apr. 2021, Art. no. 101885.

[23] G. Rani, P. Thakkar, A. Verma, V. Mehta, R. Chavan, V. S. Dhaka, R. K. Sharma, E. Vocaturo, and E. Zumpano, "KUB-UNet: Segmentation of organs of urinary system from a KUB X-ray image," *Comput. Methods Programs Biomed.*, vol. 224, Sep. 2022, Art. no. 107031.

[24] W. Xing, Z. Zhu, D. Hou, Y. Yue, F. Dai, Y. Li, L. Tong, Y. Song, and D. Ta, "CM-SegNet: A deep learning-based automatic segmentation approach for medical images by combining convolution and multilayer perceptron," *Comput. Biol. Med.*, vol. 147, Aug. 2022, Art. no. 105797.

[25] N. Yamanakkanavar and B. Lee, "MF2-net: A multipath feature fusion network for medical image segmentation," *Eng. Appl. Artif. Intell.*, vol. 114, Sep. 2022, Art. no. 105004.

[26] A. Srivastava, D. Jha, S. Chanda, U. Pal, H. D. Johansen, D. Johansen, M. A. Riegler, S. Ali, and P. Halvorsen, "MSRF-net: A multi-scale residual fusion network for biomedical image segmentation," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 5, pp. 2252–2263, May 2022.

[27] J. Cheng, S. Tian, L. Yu, H. Lu, and X. Lv, "Fully convolutional attention network for biomedical image segmentation," *Artif. Intell. Med.*, vol. 107, Jul. 2020, Art. no. 101899.

[28] M. Ma, H. Xia, Y. Tan, H. Li, and S. Song, "HT-net: Hierarchical context-attention transformer network for medical ct image segmentation," *Int. J. Speech Technol.*, vol. 52, no. 9, pp. 10692–10705, Jul. 2022.

[29] S. Pan, Y. Lei, T. Wang, J. Wynne, C.-W. Chang, J. Roper, A. B. Jani, P. Patel, J. D. Bradley, T. Liu, and X. Yang, "Male pelvic multi-organ segmentation using token-based transformer vnet," *Phys. Med. Biol.*, vol. 67, no. 20, Oct. 2022, Art. no. 205012.

[30] F. Ding, G. Yang, J. Liu, J. Wu, D. Ding, J. Xv, G. Cheng, and X. Li, "Hierarchical attention networks for medical image segmentation," 2019, *arXiv:1911.08777*.

[31] W. Wang, Z. Chen, and H. Hu, "Hierarchical attention network for image captioning," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 8957–8964.

[32] Z.-J. Gao, Y. He, and Y. Li, "A novel lightweight swin-UNet network for semantic segmentation of COVID-19 lesion in CT images," *IEEE Access*, vol. 11, pp. 950–962, 2023.

[33] Q. Xu, Z. Ma, N. He, and W. Duan, "DCSAU-net: A deeper and more compact split-attention U-Net for medical image segmentation," *Comput. Biol. Med.*, vol. 154, Mar. 2023, Art. no. 106626.

[34] S. Pan, X. Liu, N. Xie, and Y. Chong, "EG-TransUNet: A transformer-based U-Net with enhanced and guided models for biomedical image segmentation," *BMC Bioinf.*, vol. 24, no. 1, p. 85, Mar. 2023.

[35] G. Chen, J. Yin, Y. Dai, J. Zhang, X. Yin, and L. Cui, "A novel convolutional neural network for kidney ultrasound images segmentation," *Comput. Methods Programs Biomed.*, vol. 218, May 2022, Art. no. 106712.

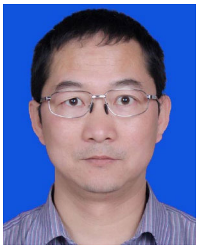
[36] J. Zhang, C. Li, S. Kosov, M. Grzegorzek, K. Shirahama, T. Jiang, C. Sun, Z. Li, and H. Li, "LCU-net: A novel low-cost U-net for environmental microorganism image segmentation," *Pattern Recognit.*, vol. 115, Jul. 2021, Art. no. 107885.

[37] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82031–82057, 2021.

NA LIU received the Ph.D. degree in management science and engineering from the School of Management and Economics, Tianjin University, Tianjin, China, in 2020. She is currently an Associate Professor with Shihezi University. Her research interests include data mining and intelligent decision-making, project management, industrial engineering, and mechanical engineering.

ZHONGHUA LU is currently pursuing the master's degree with the College of Mechanical and Electrical Engineering, Shihezi University, Shihezi, China. His research interests include image processing, machine learning, computer vision, pattern recognition, and deep learning.

WENYONG LIAN received the B.S. degree in clinical medicine from Xinjiang Medical University, in 1992. He is currently the Chief Physician of the Department of Urology, General Hospital of the Third Division of Xinjiang Production and Construction Corps, where he is also the Vice President.



MIN TIAN received the Ph.D. degree in agricultural informatics from Shihezi University, Shihezi, China, in 2011. He is currently a Professor with Shihezi University. His research interests include the Internet of Things, wireless communications, detection control, and information processing.

CHIYUE MA is currently pursuing the master's degree with the College of Mechanical and Electrical Engineering, Shihezi University, Shihezi, China. His research interests include image processing, computer vision, industrial engineering, and mechanical engineering.

LIJUAN PENG received the M.S. degree in applied mathematics from Xi'an University of Science and Technology, Xi'an, China, in 2012. She is an Associate Professor with Shihezi University, Shihezi, China. Her research interests include network security and cryptography theory, big data, mathematics classroom teaching research, and textbook research.