Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 17, No. 1, 2026

Kidney Tumor Image Segmentation Algorithm based on Deep Learning

Jian Lin *, Yan Chen

Wenzhou Polytechnic, Wenzhou, Zhejiang, 325035, China

*** Corresponding author:** Jian Lin (Email: 604330846@qq.com)

Abstract: This paper focuses on the segmentation algorithm of kidney tumor image based on deep learning, expounds the research significance of the subject, combs the research status at home and abroad in this field, and defines the research methods and technical routes. In this paper, the overall design of the algorithm is completed, and the preprocessing work such as data collection and annotation of kidney tumor image is completed. On the basis of v-net neural network, the feature fusion module and dual attention mechanism are introduced to improve the segmentation model. By setting experimental parameters to carry out experiments, the segmentation effect of the model is analyzed from the qualitative and quantitative perspectives. This study can provide technical support for the clinical diagnosis of renal tumor, and provide reference for algorithm optimization in related fields.

Keywords: Renal Tumor; Image Segmentation; Deep Learning; Segmentation Model; Feature Fusion.

1. Introduction

1.1. Research Significance

According to the global cancer statistics, in 2020, there were more than 430000 new cases of kidney tumors and 170000 deaths worldwide, accounting for 2% to 3% of the total incidence of systemic tumors, and showing an upward trend year by year [1]. Early accurate segmentation of renal tumors is the core basis for clinical diagnosis, surgical planning and prognosis evaluation. The traditional manual segmentation method relies on the clinical experience of radiologists. A single abdominal CT image contains multi-layer cross sections. The average time for manual segmentation of the complete tumor area is 45-60 minutes, and the average dice similarity coefficient of segmentation results between different physicians is only 0.78, which is poor consistency, and cannot meet the diagnosis and treatment needs of large-scale clinical cases [2].

The traditional segmentation algorithm relies on manual extraction of tumor gray and texture features, and the segmentation accuracy of small renal tumors with uneven density and fuzzy boundary is only about 0.62, which is difficult to adapt to the features of large morphological differences and unstable location of renal tumors [3]. The segmentation algorithm based on deep learning can automatically extract multi-scale image features, skip the manual feature extraction link, and greatly improve the segmentation efficiency. In this study, the existing algorithms are optimized for the problems of fuzzy boundary segmentation and feature redundancy of small tumors. The optimized algorithm can compress the segmentation time of a single CT image to less than 8 seconds, providing technical support for rapid clinical diagnosis.

From the perspective of clinical application, accurate automatic segmentation of renal tumors can help doctors quickly determine the adjacent relationship between tumor volume, location and surrounding tissues, reduce the risk of vascular injury during partial nephrectomy, and improve the safety of surgery. From the perspective of technology development, the improved segmentation framework proposed in this study can provide reusable technical ideas for image segmentation of other abdominal organ tumors, and promote the application of deep learning technology in the field of medical imaging.

1.2. Research Status at Home and Abroad

Image segmentation divides the image into non overlapping regions according to features such as gray level, edge and shape [4]. Medical image segmentation is to extract the lesion area required by doctors from medical images, so that doctors can more intuitively and clearly observe the anatomical structure and analyze the lesions, and provide scientific basis for clinical diagnosis and operation plan [5]. It helps to identify lesions and quantify the size, shape and location of physiological structures by accurately distinguishing the boundaries of different tissues or organs. Medical image segmentation tasks can be divided into traditional methods and deep learning methods.

Medical image segmentation based on traditional methods: traditional methods use artificial design features, and use the differences between the target and background in gray, contrast, texture and other aspects to complete the segmentation. The common methods are threshold method, multi-image set registration method and region growth method.

Threshold segmentation is divided into global threshold method and local threshold method. The global threshold method selects a fixed threshold according to the gray distribution of the whole image; The local threshold method divides the image into multiple regions, and sets thresholds respectively to adapt to local gray changes. Kim et al. used the threshold method to segment the kidney after removing the spine with similar gray level to the kidney [6]. This method has the advantages of simple design, low complexity and wide application range.

The multi atlas registration method needs to collect and preprocess the images from different sources, and then extract the features such as edges, corners and gray histogram to calculate the similarity between images. At the same time, the

rigid body, affine or nonlinear registration models are selected, and the least square method, gradient descent method and other optimization parameters are used. Finally, the registration results are evaluated. Yang et al. proposed a multi atlas registration strategy from coarse to fine [7]. First, the image was registered with the atlas as a whole to obtain the coarse segmentation and positioning of the kidney, and then the region of interest of the kidney was collected and registered to the cropped atlas to construct a high-resolution atlas to obtain the final segmentation.

The growth method starts from the seed point, and gradually includes the adjacent pixels that meet the gray similarity conditions. When the gray difference between the pixel and the seed point is less than the given threshold, the growth stops, but it is sensitive to parameters and noise. Lin et al. processed abdominal CT in two stages in combination with renal anatomical characteristics [8]: in the first stage, candidate regions were extracted according to the geometric position of the kidney; In the second stage, kidney segmentation is completed by spine localization, oval candidate selection extraction, direction model and adaptive region growth.

Traditional methods are susceptible to motion artifacts, metal artifacts and other noise, and the segmentation results are prone to bias. Threshold, radius and other parameters depend on experience or manual adjustment, and different tasks need to be adjusted again. In addition, when processing large-scale data, the computational complexity is high and the speed is slow, which is difficult to meet the requirements of real-time segmentation.

With the development of artificial intelligence and neural network technology, deep learning method shows obvious advantages in medical image segmentation, which can effectively improve the accuracy of complex image segmentation.

Convolutional neural network (CNN) automatically extracts local features and hierarchical representation through convolution layer, pooling layer and full connection layer. Haghighi et al. used CNN to roughly segment the kidney, and then automatically complete the kidney segmentation [9]. The full convolution network (FCN) replaces the full connection layer of the traditional CNN with the convolution layer, realizes the classification of each pixel, and restores the feature image to the size of the original image by up sampling, preserving the detail information. This end-to-end learning method significantly improves the segmentation performance. Sharma et al. used FCN to automatically segment polycystic kidney [10]. On this basis, models with better performance continue to emerge.

U-net is an important improvement of FCN, with a "U" shaped structure. The left shrinkage path extracts advanced features through convolution and pooling, and the right expansion path is mapped to pixel level segmentation results through up sampling and convolution. The features on both sides are fused by jumping connection to restore details and improve accuracy [11]. Zhang Fang et al. [12] proposed deep multi-scale aggregation 3D u-net for kidney and kidney tumor segmentation, adding three subsampling operations on the basis of u-net++, and introducing densely nested 3D u-net structure and jumping connection between decoder layers and sub networks, which enhanced the segmentation accuracy of small-scale tumors and edges. Liu et al. proposed a u-net architecture completely based on transformer, which uses the improved self-attention mechanism cswin to parallel calculate the attention on horizontal and vertical slices [13]. Half of the attention head processes the horizontal direction, the other half processes the vertical direction, and then splices along the channel, achieving good results in kidney segmentation.

For 3D images, milletari et al. proposed v-net based on volume full convolution [14]. Pabitha et al. further proposed the adaptive recalibration context extrusion and excitation self-attention v-net (arcsav net) [15], which integrates the adaptive recalibration context extrusion and excitation module, spatial attention, multi-scale feature fusion, context-based confidence estimation and Bayesian superparameter optimization, and can more effectively segment the renal tumor in the renal cell carcinoma image.

1.3. Research Methods and Technical Route

This study uses the public data set kits19 as the research data. This data set contains 300 cases of abdominal CT scanning images of kidney tumors. The thickness range of each image is 0.5mm-5.0mm, and the pixel resolution is 512 × 512. All samples are manually labeled by professional radiologists, and the labeling accuracy reaches pixel level. Aiming at the problem of large gray range and low proportion of tumor area in the original CT image, the gray truncation processing based on the window width and window level was adopted. The window width of kidney area was set to 400hu and the window level was set to 50hu, and the truncated gray value was normalized to the [0,1] interval. At the same time, the input size was unified to 256 × 256 × 64 by random clipping. The data amplification was completed by random flipping, random rotation and Gaussian noise addition. After amplification, the training sample size was increased by three times, which effectively reduced the risk of model over fitting.

Based on the framework of v-net network, the multi-scale feature fusion module is introduced to enhance the transmission of high-level and low-level semantic information, and the channel attention mechanism and spatial attention mechanism are added to strengthen the feature weight of tumor region and suppress the invalid features of background region. The experiment uses dice similarity coefficient, IOU intersection and union ratio, pixel accuracy and hd95 Hausdorff distance to complete the segmentation performance evaluation, and verifies the effectiveness of the improved algorithm by comparing the original v-net, u-net and attention u-net.

The technical route of the study is as follows: firstly, the pretreatment of kits19 data set and the division of training verification set are completed, and the proportion of training set, verification set and test set is set to 7:2:1; Then, an improved v-net segmentation model with embedded attention mechanism was constructed. The initial learning rate was set to 1e-4, the batch size was set to 4, and the number of training iterations was set to 100. The adamw optimizer was used to update the model parameters, and the combination of weighted dice loss and cross entropy loss was used as the loss function; Then the model training and parameter optimization are completed, and the qualitative and quantitative analysis of segmentation performance is completed on the test set; Finally, the research conclusion is summarized, the limitations of the algorithm are analyzed, and the future improvement direction is proposed.

2. Design of Kidney Tumor Image Segmentation Algorithm Based on Deep Learning

2.1. Data Preprocessing

2.1.1. Data Collection

In this study, kits19, a public medical image data set, was selected as the core data source. This data set was released by the International Conference on medical image computing and computer-aided intervention, and was specially used for the segmentation algorithm research of kidney and kidney tumors. It contained the abdominal three-phase enhanced CT scanning data of 210 patients with kidney tumors. All CT images were stored in DICOM format, the layer thickness was uniformly resampled to 1mm, the pixel size of a single CT slice was 512 × 512, and the pixel gray range ranged from -1000hu to 1000hu, covering the complete gray distribution information of renal parenchyma and tumor area. In order to ensure the diversity of data, this study additionally obtained the CT enhanced image data of 30 cases of renal clear cell carcinoma diagnosed by clinical pathology from the medical imaging department of a tertiary hospital. All cases have removed the patient's privacy information and obtained the approval of the hospital's ethics committee. The total sample size reached 240 cases, meeting the requirements of data scale for deep learning model training. In CT images, the boundary of renal tumor usually has little gray difference with surrounding fat and adjacent organs, and the direct input model will produce a large number of unrelated background interference. In this study, the original data were preliminarily trimmed to retain the range of interest including the kidney and tumor region, and a single slice was compressed to 256 × 256 pixels, reducing the subsequent calculation while retaining the complete segmentation target features. At the same time, 12 samples of minimal lesions with severe scanning artifacts and tumor area diameter less than 5mm were eliminated, and 228 effective samples were finally obtained. According to the ratio of 8:1:1, the samples were divided into training set, validation set and test set, including 182 cases of training set, 23 cases of validation set and 23 cases of test set, providing a stable data basis for subsequent model training and performance verification.

2.1.2. Data Annotation

In this study, the tagging rules of the open medical data set kits19 (kidney tumor segmentation challenge 2019) were adopted, and the tagging was completed by two radiologists with more than five years' experience in abdominal image reading. The tagging content included two categories: renal parenchyma area and renal tumor area. All original CT images were in DICOM format of axial scanning, with a layer thickness range of 0.5mm5.0mm, a single sample slice number range of 32512, and a unified image size of 512 × 512 pixels.

Labelme medical image annotation tool is used for annotation, and the contour is drawn layer by layer in three-dimensional space to avoid the interruption error of organ continuity caused by single-layer annotation. For small renal tumors (maximum diameter<3cm), when labeling, enlarge to 1.5 times of the original image for edge delineation to reduce the edge blur error caused by volume effect. After all, 210 samples were labeled, the third chief physician with more than 10 years' clinical experience conducted a cross check. For samples with different labeling results, the final labeling results were determined through joint film reading and discussion. Finally, three samples with labeling consistency less than 0.75 were eliminated, and 207 effective labeling samples were obtained.

In the labeling process, Dice coefficient was used to measure the consistency of different taggers. The consistency of kidney region labeling was 0.89, and the consistency of kidney tumor region labeling was 0.82, which met the labeling requirements of medical image segmentation research. Finally, all annotation results are saved as PNG format masks consistent with the original CT image size, in which the pixel value of the background area is set to 0, the pixel value of the renal parenchyma area is set to 1, and the pixel value of the renal tumor area is set to 2. According to the ratio of 8:1:1, they are randomly divided into training set, validation set and test set, with 166 cases in the training set, 21 cases in the validation set, and 20 cases in the test set, providing standardized annotation data for subsequent model training and performance evaluation.

2.2. Model Selection and Improvement

2.2.1. Model Selection

The core goal of medical image segmentation task is to accurately separate the kidney region and tumor region from CT images, and provide a clear anatomical basis for subsequent clinical diagnosis and surgical planning. The CT image of renal tumor has the characteristics of fuzzy boundary, large difference in tumor size and shape, and high gray similarity of surrounding tissues, which puts forward higher requirements for the ability of feature extraction and multi-scale information fusion of segmentation model. The early traditional segmentation algorithms such as threshold segmentation and region growth method rely on manual setting of feature threshold, and the segmentation accuracy of kidney tumor image with uneven gray level is very low, which can not meet the needs of clinical diagnosis.

In the existing deep learning segmentation model, the full convolution network (FCN) achieves end-to-end segmentation by replacing the full connection layer with the full convolution layer, but a large number of shallow spatial details will be lost in the down sampling process, resulting in blurred edges of small tumor segmentation. The u-net model has achieved good results in the field of medical image segmentation by fusing deep and shallow features through encoder decoder structure and jump connection. However, for 3D CT volume data, u-net uses two-dimensional convolution to extract features, which cannot effectively use the interlayer spatial information of CT sequence, and is prone to the problem of discontinuity of tumor area.

V-net model is a full convolution segmentation network specially designed for three-dimensional medical images. It uses three-dimensional convolution kernel to extract spatial features, and can completely capture the context information of kidney tumor in CT sequence. Its built-in skip connection structure can integrate the features of different depths, and retain the details of tumor edge. At the same time, v-net uses Dice coefficient as the loss function, which can effectively alleviate the common category imbalance problem in medical image segmentation. In this study, the renal tumor area accounts for only 2.1%~7.3% of the total image volume. Dice loss can better focus on the segmentation accuracy of the foreground area. Compared with 3D u-net, the parameter scale of v-net is reduced by 19%, and the training speed is increased by 23%, which is more suitable for the memory

limit of the single NVIDIA RTX 3090 video card used in this study. Therefore, this study chooses v-net as the basic segmentation model for subsequent research.

2.2.2. Model Improvement

When the original v-net model is used to process CT images of renal tumors, there are some problems, such as insufficient segmentation accuracy of small lesions, fuzzy boundary, feature redundancy interference and so on. According to the characteristics of large differences in the size of renal tumors, with diameters ranging from 8mm to 62mm, and small tumors accounting for 21.7%, a 1 × 1 × 1 short connection branch was added in the encoder downsampling phase to preserve the shallow spatial details and avoid the loss of small tumor features in the downsampling process. The number of short connection output channels was adjusted to 1/2 of the original number, and the increase of model parameters was controlled to be no more than 12%. To solve the problem of mixed tumor features in different modality CT images, a channel attention module is added at the output of each coding module to allocate dynamic weights to different feature channels containing tumor texture, gray level and edge information, and suppress redundant feature channels with weights less than 0.05, so as to reduce the interference of unrelated tissues on the segmentation results. To solve the problem of fuzzy boundary segmentation of the original model, a boundary monitoring branch is added to the output of the last layer of the decoder, and the contour region of the segmentation result is given a weight of 2.5 times of cross entropy loss, so as to strengthen the learning ability of the model to the tumor boundary. The weighted combination loss function of dice loss and cross entropy loss is introduced. The dice loss weight is set to 0.6, and the cross entropy loss weight is set to 0.4, which solves the sample imbalance problem caused by the original single loss function that the proportion of renal tumor area is only 1.2%~3.8%, and improves the learning sensitivity of the model to small target areas. The parameters of the improved model are controlled at 31.2m, only 4.7m more than the original v-net, and the increase in computational complexity is controlled within an acceptable range. At the same time, it adapts to the segmentation reasoning requirements of clinical CT images.

3. Neural Network Algorithm Design

3.1. V-net Neural Network

V-net is a fully convolutional neural network designed for 3D medical image segmentation. Its structure retains the core features of the classical encoding decoding architecture, and makes targeted adaptation for the characteristics of 3D volumetric data. The renal tumor in medical CT image belongs to the continuous lesions in three-dimensional space. The traditional two-dimensional segmentation method can only deal with the single slice information, and can not use the spatial correlation information of the lesions in the continuous layer. V-net can better capture the morphological continuity and spatial position characteristics of renal tumor by extracting the spatial features through three-dimensional convolution, which is more suitable for the segmentation task of renal tumor CT image.

The coding path of v-net includes four down sampling modules, each module is composed of two stacked 3D convolution layers, the convolution core size is set to 3 × 3 × 3, and the step size is 1. The low resolution semantic features of renal tumors are gradually extracted through convolution operation. After each module feature extraction is completed, 2 × 2 × 2 3D convolution with the step size of 2 is used to realize down sampling, which compresses the size of the feature map to 1/2 of the original, and doubles the number of feature channels, gradually improving the ability of the network to capture abstract features. In the coding process, the network will retain the feature map of each stage for subsequent jump connection to realize the fusion of shallow spatial detail information and deep semantic information.

The decoding path also corresponds to four upsampling modules. Each module first doubles the size of the feature map by deconvolution, while reducing the number of feature channels. Then, the upsampling results and the feature map reserved in the corresponding encoding phase are stitched in channel dimensions. The stitched features are then fused through two three-dimensional convolution layers. The final output layer uses 1 × 1 × 1 three-dimensional convolution to map the number of characteristic channels to the number of segmentation categories. For the kidney tumor segmentation task, the output channel is set to 2, corresponding to the background region and the kidney tumor region respectively.

V-net uses dice loss function optimization model to directly optimize the overlapping degree evaluation index of segmentation tasks, which can effectively alleviate the category imbalance caused by the small proportion of lesions in kidney tumor segmentation. Compared with cross entropy loss, v-net has faster convergence speed and higher segmentation accuracy in small target segmentation tasks. The range of receptive field of the original v-net is limited, and the feature extraction ability of the infiltrating region with fuzzy boundary in renal tumor is insufficient, so the feature extraction ability needs to be optimized through additional modules.

3.2. Feature Fusion Module

The integration ability of the original v-net network for different scale features is insufficient. The size difference of renal tumor lesions can reach more than 12 times. Small lesion features are easy to be submerged by large-scale background features, resulting in the decline of segmentation accuracy. The core function of the feature fusion module is to effectively fuse the multi-scale spatial features output by the encoder in different down sampling stages with the semantic features of the corresponding stage of the decoder, preserve the edge details of small lesions, and strengthen the global semantic expression ability of large lesions. This module adopts the adaptive weighted fusion strategy to reduce the dimension of the four layers of features with the output resolution of 128 × 128 × 64, 64 × 64 × 32, 32 × 32 × 16 and 16 × 16 × 8 by 1 × 1 × 1 convolution, and compress the number of channels to 1/2 of the original dimension, reducing the amount of redundant calculation. The dimension reduced features are uniformly interpolated and sampled to the size of the corresponding stage of the decoder, and the mutual information coefficient of each layer of features and the current features of the decoder is calculated. The mutual information value is used as the weight to weight the features of each scale, replacing the traditional direct splicing fusion method, and improving the contribution of effective features. Ablation experiments on CT image datasets of renal tumors showed that after adding this module, the Dice coefficient of the model for small lesions with diameter less than 10mm increased by 4.7%, and the intersection ratio increased by 3.9%, which effectively alleviated the problem of missing

segmentation of small lesions, reduced the over segmentation probability of the edge of large lesions, and steadily improved the overall segmentation accuracy. The parameters of the module are only increased by 2.1m, and the reasoning time of a single CT image is increased by less than 8ms, which does not cause obvious burden on the reasoning efficiency of the model, and takes into account the balance between segmentation accuracy and computational efficiency.

3.3. Channel Attention Mechanism

In the CT image of renal tumor, the gray feature similarity of tumor area and normal renal tissue is high. The contribution of feature channels extracted at different levels to the segmentation task is significantly different. Some redundant background feature channels will also interfere with the recognition of tumor boundary. The channel attention mechanism can enhance the response of tumor related feature channels and suppress the interference of unrelated channels by assigning differentiated weights to different feature channels, so as to improve the feature extraction ability of the model for renal tumor region.

In this paper, we use the squeeze and exception (SE) structure to realize the channel attention modeling. This structure can improve the performance by adding only a few parameters, and the proportion of parameter increase is less than 2% of the total parameters of the original model. The specific calculation process is divided into two steps: compression and activation: the compression step compresses the three-dimensional feature map with size into the channel description vector through the global average pooling, so as to realize the channel dimension aggregation of global spatial information; The activation step builds the channel dependency relationship through two full connection layers. The first full connection layer compresses the number of channels to 1/16 of the original number of channels to reduce the computational complexity, uses the relu function to activate, and the second full connection layer restores the number of channels to the original number, uses the sigmoid function to map the output to the [0,1] interval, and finally obtains the weight coefficient of each channel.

By multiplying the generated channel weights with the original feature map channel by channel, the feature recalibration can be completed, which makes the model focus on the feature channels related to renal tumors. In the comparative experiment of kits data set, after adding the channel attention mechanism, the dice similarity coefficient of the model increased from 0.812 to 0.847, and the intersection to union ratio increased from 0.724 to 0.758, especially for the segmentation accuracy of small renal tumors, which verified the optimization effect of the module on feature selection.

3.4. Spatial Attention Mechanism

In the CT image features of renal tumor extracted by convolutional neural network, there are significant differences in the information density of tumor regions contained in different spatial positions. The influence of the characteristics of the boundary region between the tumor edge and normal renal tissue on the segmentation accuracy is much greater than that of the uniform background region. The spatial attention mechanism can enhance the feature response of the key regions of the tumor and suppress the interference of redundant features in the unrelated background region by modeling the attention weight of the spatial dimension. The spatial attention module designed in this study is connected to the channel attention module, and the weight of the spatial dimension of the feature map after the channel attention adjustment is re calibrated. The module inputs the three-dimensional feature map with the size c × h × w × D, where C is the number of channels, and H, W, and D correspond to the length and width and slice depth of the CT image respectively.

The module uses mean pooling and maximum pooling to compress the input feature map along the channel dimension, and obtains two 1 × h × w × D spatial feature maps respectively. After the two feature maps are spliced in the channel dimension, the number of channels is reduced to 1 through 1 × 1 × 1 three-dimensional convolution, and then the sigmoid activation function is used to generate a spatial attention weight map with a size of 1 × h × w × D, and the weight range is 0 to 1. The high weight corresponds to the tumor area that needs enhancement, and the low weight corresponds to the background area that needs inhibition. The generated spatial attention weight map and the module input feature map are multiplied pixel by pixel to obtain the final spatial enhanced feature output.

In the CT image of renal tumor, the proportion of tumor area in the whole image is usually less than 15%. After the introduction of spatial attention mechanism, the model can automatically focus on the feature learning of tumor area and reduce the computational cost of useless features in the background area. This experiment compared the segmentation performance before and after adding spatial attention mechanism. The dice similarity coefficient of the model increased from 0.821 to 0.847, and the average intersection union ratio increased from 0.735 to 0.768, especially for small tumors with a volume less than 10cm ^{3}. It verified the enhancement effect of the module on the characteristics of small target tumors.

4. Experiment and Result Analysis

4.1. Experimental Design and Parameter Setting

The open kits19 renal tumor CT data set was used to verify the experiment. The data set contains 210 sets of abdominal 3D CT scanning data. The layer thickness of each set of data ranges from 0.5mm to 5.0mm, and the pixel resolution is 512 × 512. The manual labeling of kidney and tumor area has been completed by professional radiologists. In this experiment, the data set was divided into 168 training sets, 21 validation sets and 21 test sets according to the ratio of 8:1:1, to ensure that the data distribution is consistent with the original data set and avoid the interference of sample deviation on the experimental results.

In the data pre-processing stage, the input CT image is cut to 256 × 256 × 64 three-dimensional size, and the CT value is normalized by Z-score standardization. The pixel value range is mapped to the distribution with the mean value of 0 and the standard deviation of 1 to reduce the influence of gray shift caused by the difference of CT equipment scanning parameters. At the same time, the data is enhanced by random horizontal flipping, random rotation (the rotation angle range is -15 ° to 15 °) and random elastic deformation, so as to expand the training sample size and reduce the risk of model over fitting.

The experiment built a model based on pytorch 1.12 deep learning framework, and the hardware environment used

NVIDIA RTX 3090 graphics card with 24GB memory capacity, which supported parallel computing of batch 3D images. Adamw optimizer is used in the training process. The initial learning rate is set to 1e-4, the weight attenuation coefficient is set to 1e-5, the total number of training rounds is 100, and the batch size is set to 2. The learning rate is dynamically adjusted by cosine annealing strategy. The model performance is verified every 10 rounds of training, and the model weight with the highest Dice coefficient in the validation set is reserved for subsequent testing. The loss function uses a weighted combination of cross entropy loss with weight and dice loss. Aiming at the sample imbalance problem of low proportion of kidney tumor area, the loss weights of background, kidney and tumor area are set to 0.2, 0.8 and 1.5 respectively, to improve the learning attention of the model to the tumor boundary.

4.2. Analysis of Experimental Results

4.2.1. Qualitative Analysis

This study uses the open source kits19 kidney tumor CT data set, and randomly selects 200 groups of 3D CT image samples including complete kidney and tumor annotation for visual comparative analysis. The comparison objects include the original v-net model, the v-net model with feature fusion module, the improved model with channel attention and spatial attention mechanism, and the segmentation results of the three types of models and the expert gold standard annotation.

The CT slice thickness of all samples was uniformly resampled to 1mm × 1mm × 3mm, the window width was set to 300hu, and the window center was set to 40hu to match the gray display range of renal soft tissue. The results showed that the original v-net model had obvious under segmentation problem for small renal tumors with a diameter less than 10mm. The fuzzy area at the edge of the tumor was wrongly divided into normal renal parenchyma, and some tumors nested in the renal sinus were completely missed segmentation; For the large invasive tumor, the original model is prone to over segmentation, and the normal kidney tissue with edema around the tumor is wrongly classified as the tumor area. The segmentation boundary has a strong sense of burr, and the matching degree with the gold standard is low.

After adding the feature fusion module, the recognition ability of the improved model for small tumors is significantly improved, and the proportion of missing segmentation samples is reduced from 21% of the original model to 8%, but there is still the problem of boundary dislocation in the transition region where the gray level of tumor and normal tissue is close. At the same time, the final model with dual attention mechanism can effectively focus on the characteristic information of the tumor area, suppress the interference of the characteristics of unrelated areas such as normal renal tissue and blood vessels, and the segmentation boundary is closer to the expert gold standard. It can also obtain continuous and smooth segmentation contour for tumors in complex positions such as renal hilus and renal sinus.

4.2.2. Quantitative Analysis

In this study, the segmentation performance of the improved model was evaluated by using four quantitative indicators, namely, dice similarity coefficient (DSC), Jaccard coefficient, average surface distance (ASD) and Hausdorff distance (hd), which are commonly used in the field of medical image segmentation. All index calculations were based on 210 labeled CT samples in kits2021 public data set, including 168 training sets, 21 validation sets and 21 test sets. The input images were uniformly trimmed to 256 × 256 × 64 pixels.

The experimental results showed that the improved v-net model, which fused the channel spatial attention mechanism and the multi-scale feature fusion module, achieved 0.924 DSC, 0.856 Jaccard, 0.712mm ASD and 8.245mm HD in the kidney segmentation task, and 0.831 DSC, 0.718 Jaccard, 1.862mm ASD and 15.372mm HD in the tumor segmentation task. Compared with the original v-net model, the kidney segmentation DSC increased by 4.2%, and the tumor segmentation DSC increased by 6.7%; Compared with the classical u-net model in the same field, the kidney segmentation DSC increased by 2.8%, and the tumor segmentation DSC increased by 5.1%.

Further analysis of the segmentation performance of tumors with different sizes showed that for small tumors with a diameter less than 2cm, the DSC of the model reached 0.772, which was 8.3% higher than the original v-net, indicating that the attention mechanism effectively enhanced the ability of the model to capture the characteristics of small target tumors, and alleviated the segmentation error caused by the blurred boundary between the tumor and the surrounding normal kidney tissue in CT images. Paired t-test results show that the performance difference between the model and the original v-net and u-net is statistically significant ($p<0.05$), which proves that the improved model can significantly improve the segmentation accuracy. The model parameter is 14.8m, and the reasoning time of single CT image segmentation is 1.23s, which meets the efficiency requirements of clinical auxiliary diagnosis.

5. Conclusion

In this study, the kidney tumor CT image segmentation task is taken as the core, aiming at the problems of the existing deep learning model for small volume tumor segmentation accuracy is insufficient, and the edge details are lost seriously, the multi-scale feature fusion module and parallel channel spatial attention mechanism are introduced on the basis of v-net, and an improved kidney tumor segmentation model is constructed. The open kidney tumor CT data set kits19 was used in the experiment, including 210 cases of enhanced CT scanning samples, including 189 cases as the training set and 21 cases as the validation set. The initial learning rate was set to 1e-4, the batch size was set to 4, and the mixed loss function weighted by cross entropy loss and dice loss was used. The number of training iterations was 200 rounds. The experimental results show that the dice similarity coefficient of the improved model for the kidney region is 0.971, and the dice similarity coefficient for the kidney tumor region is 0.836, which is 2.7% and 8.2% higher than the original v-net model, and the intersection ratio index is also improved from 0.712 to 0.793 of the original model, which proves that the module improvement has a significant effect on the improvement of segmentation performance.

This study confirmed that multi-scale feature fusion can retain the detailed texture information of the shallow network, alleviate the information loss caused by the deep network down sampling, and the parallel attention mechanism can strengthen the feature weight of the tumor region, inhibit the interference of the background and normal tissue features, and effectively improve the recognition ability of small tumors. The improved model can automatically and

accurately segment the renal tumor region from CT images, and can provide stable technical support for the preoperative planning and curative effect evaluation of clinical renal tumors.

Limited by the number of samples in the dataset, the generalization ability of the model for rare pathological subtypes of renal tumors still has room for improvement. In the future, the multi center sample size can be expanded, and the edge segmentation accuracy can be further optimized combined with the boundary refinement module.

Acknowledgments

This work was supported by the General scientific research projects of Zhejiang Provincial Department of Education under Grant Y202559794.

References

[1] Ye, R. G., Chen, Y. S., & Fang, J. A. (2003). Summary of the Symposium on the diagnosis, treatment and efficacy criteria of kidney disease. *Chinese Journal of Integrated Traditional and Western Medicine Nephrology*, *4*(5), 249–251.

[2] Lu, C., Yang, S. F., & Yue, H. (2008). Epidemiological investigation status of chronic kidney disease. *Medical Review*, *14*(3), 370–372.

[3] Agnes, A. S., Anitha, J., & Peter, D. J. (2018). Automatic lung segmentation in low-dose chest CT scans using convolutional deep and wide network (CDWN). *Neural Computing and Applications*, *32*(20), 1–11. https://doi.org/10.1007/s00521-018-3745-6.

[4] Sun, Q., Dai, M., Lan, Z., et al. (2022). UCR-Net: U-shaped context residual network for medical image segmentation. *Computers in Biology and Medicine*, *151*, 106203. https:// doi. org/ 10.1016/j.compbiomed.2022.106203.

[5] Zhang, N., Yu, L., Zhang, D., et al. (2022). APT-Net: Adaptive encoding and parallel decoding transformer for medical image segmentation. *Computers in Biology and Medicine*, *151*, 106292. https://doi.org/10.1016/j.compbiomed.2022.106292.

[6] Belgherbi, A., Hadjidj, I., & Bessaid, A. (2014). Morphological segmentation of the kidneys from abdominal CT images. *Journal of Mechanics in Medicine and Biology*, *14*(5), 11–12. https://doi.org/10.1142/S0219519414500573.

[7] Yang, G., Gu, J., Chen, Y., et al. (2014). Automatic kidney segmentation in CT images based on multi-atlas image registration. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 5538–5541). IEEE. https://doi.org/10. 1109/EMBC. 2014. 694 4894.

[8] Lin, D. T., Lei, C. C., & Hung, S. W. (2006). Computer-aided kidney segmentation on abdominal CT images. *IEEE Transactions on Information Technology in Biomedicine*, *10*(1), 59–65. https://doi.org/10.1109/TITB.2005.855567.

[9] Sun, P., Mo, Z., Hu, F., et al. (2022). Segmentation of kidney mass using AgDenseU-Net 2.5 D model. *Computers in Biology and Medicine*, *150*, 106223. https://doi.org/10.1016/j. compbiomed. 2022.106223.

[10] Huang, Y., Jiao, J., Yu, J., et al. (2023). RsALUNet: A reinforcement supervision U-Net-based framework for multi-ROI segmentation of medical images. *Biomedical Signal Processing and Control*, *84*, 104743. https://doi.org/10. 1016/j. bspc.2023.104743.

[11] Zhou, Y., Jiang, H., Diao, Z., et al. (2023). MRLA-Net: A tumor segmentation network embedded with a multiple receptive-field lesion attention module in PET-CT images. *Computers in Biology and Medicine*, *153*, 106538. https://doi. org/ 10.1016/j.compbiomed.2023.106538.

[12] Zhang, F., Hao, S. M., & Geng, L. (2023). Segmentation method of kidney and renal tumor based on deep multi-scale aggregation 3D U-Net. *Journal of Tianjin University of Technology*, *42*(6), 84–90.

[13] Alexandru, B. M., & Popescu, D. (2025). Two-Stage Neural Network Pipeline for Kidney and Tumor Segmentation. *IEEE Access*, 1–12. https://doi.org/10.1109/ACCESS.2025.3492761.

[14] Milletari, F., Navab, N., & Ahmadi, S. A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)* (pp. 565–571). IEEE. https://doi.org/10. 1109/ 3DV.2016.68.

[15] Pabitha, C., Benila, S., & Vanathi, B. (2025). An adaptive recalibrative contextual squeeze-and-excitation self-attention V-Net for kidney tumor segmentation in RCC imaging. *The European Physical Journal Plus*, *140*(8), 1–22. https://doi.org/ 10.1140/epjp/s13360-025-05879-9.