

Article

RDCTrans U-Net: A Hybrid Variable Architecture for Liver CT Image Segmentation

Lingyun Li ¹ and Hongbing Ma ^{1,2,*} 
¹ College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China; lingyun_li1823@163.com

² Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

* Correspondence: hbma@tsinghua.edu.cn

Abstract: Segmenting medical images is a necessary prerequisite for disease diagnosis and treatment planning. Among various medical image segmentation tasks, U-Net-based variants have been widely used in liver tumor segmentation tasks. In view of the highly variable shape and size of tumors, in order to improve the accuracy of segmentation, this paper proposes a U-Net-based hybrid variable structure—RDCTrans U-Net for liver tumor segmentation in computed tomography (CT) examinations. We design a backbone network dominated by ResNeXt50 and supplemented by dilated convolution to increase the network depth, expand the perceptual field, and improve the efficiency of feature extraction without increasing the parameters. At the same time, Transformer is introduced in down-sampling to increase the network's overall perception and global understanding of the image and to improve the accuracy of liver tumor segmentation. The method proposed in this paper tests the segmentation performance of liver tumors on the LiTS (Liver Tumor Segmentation) dataset. It obtained 89.22% mIoU and 98.91% Acc, for liver and tumor segmentation. The proposed model also achieved 93.38% Dice and 89.87% Dice, respectively. Compared with the original U-Net and the U-Net model that introduces dense connection, attention mechanism, and Transformer, respectively, the method proposed in this paper achieves SOTA (state of art) results.

Keywords: liver tumor segmentation; U-Net; ResNeXt50; dilated convolution; transformer



Citation: Li, L.; Ma, H. RDCTrans U-Net: A Hybrid Variable Architecture for Liver CT Image Segmentation. *Sensors* **2022**, *22*, 2452. <https://doi.org/10.3390/s22072452>

Academic Editors: Sejung Yang, Jang-Hwan Choi and Cristiana Corsi

Received: 23 February 2022

Accepted: 18 March 2022

Published: 23 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Liver cancer is currently one of the most common cancer diseases in the world, causing a large number of deaths every year [1,2]. Liver cancer is a malignant tumor of the liver, which can be classified into two categories: primary and secondary. Primary liver cancer originates from the epithelial or mesenchymal tissue of the liver and is a high-incidence and extremely harmful malignant tumor in China; secondary liver cancer is called sarcoma, which is relatively rare compared with primary liver cancer. As the largest solid organ of the human body, the liver undertakes various important metabolic functions of the human body. Once malignant tumors appear in the liver, they can lead to serious and life-threatening consequences. Therefore, early detection and treatment are the keys to improving the survival rate of liver cancer patients. CT-based imaging methods are commonly used to evaluate liver tumors, and CT examinations can clearly show the size and shape, number, and boundaries of lesions. Segmentation of liver lesions is a preparatory step for diagnosis and plays an indispensable role in the treatment of the disease. Liver segmentation is divided into manual segmentation and semi-automatic segmentation. However, manual segmentation largely relies on the judgment of radiologists, which is time-consuming and error-prone; adding manual intervention in the semi-automatic segmentation process will lead to bias and errors. The task of automatically segmenting liver tumor lesions becomes very challenging given the unique diversity and spread of liver tumor shapes.

In recent years, with the wide application of deep learning technology in medical image segmentation tasks, more and more researchers use deep learning to achieve automatic segmentation of liver tumors [3,4]. However, what really improves the performance of liver tumor segmentation is the U-Net proposed by Ronneberger et al. [5]. U-Net is developed based on the FCN [6] proposed by Long et al. in 2015, which makes up for more low-level semantic information with its unique U-shaped and skip connection structure, and it only needs to be trained with a small amount of data to obtain more accurate results. With the development of medical technology, the U-net architecture is widely used for various medical imaging analysis. However, as a 2D network, the U-Net network has difficulties using the 3D spatial information of liver slices and cannot automatically segment 3D liver images such as CT and MRI [7]. To solve this problem, Ahmed et al. proposed 3D U-Net [8] in 2016, which replaced all 2D operations with corresponding 3D operations on the basis of U-net architecture to produce 3D segmented images, effectively using the spatial information between adjacent liver slices and achieving better segmentation results. In 2016, Liu et al. [9] designed a network suitable for 3D medical image segmentation by introducing the residual connection idea of Res-Net [10] on the basis of U-Net, where the encoder is used to extract liver features and the decoder is used to generate full-resolution output. This method improves segmentation accuracy but requires longer training time due to the large number of parameters for 3D convolution. To this end, Li et al. [11] proposed H-Dense U-Net in 2018 to alleviate this problem. The idea is to mix the features of 2D Dense U-Net and 3D Dense U-Net to accelerate the convergence of 3D Dense U-Net, while embedding dense connection blocks in U-Net, which can further improve the segmentation accuracy of liver and tumor. In 2017, Han et al. [12] proposed to segment liver and tumor by stacking multiple 2D information into 3D information. The method is based on U-Net's skip connections and Res-Net's residual connections, and then, it provides 3D contextual information by inputting multiple consecutive single slices. However, the acquisition of 3D contextual information by adding densely connected and residually connected models is limited. Therefore, in 2020 Cai et al. [13] combined local features with their corresponding global dependencies by adding attention gates (AG) [14]. In the liver tumor segmentation task, it is guaranteed that it automatically ignores other irrelevant regions while focusing only on the liver tumor location and explicitly modeling the dependencies between channels helps to capture rich contextual dependencies. It is worth mentioning that in 2017, Vaswani et al. [15] proposed a network based entirely on the attention mechanism and successfully applied it to the NLP field, and brought profound inspiration to scholars in the CV field. Because compared to Attention, the parallel ability of Transformer's own self-attention enables it to have better adaptability in the face of big data. In addition, the Transformer model is flexible and can be applied to any type of data if it is abstracted as a series of embeddings. Therefore, in 2021, Dosovitskiy et al. [16] used Transformer directly in the image block sequence to complete the task of image classification, successfully achieved the most advanced performance on multiple image recognition benchmarks, and successfully applied Transformer to the field of computer vision. In the same year, Chen et al. [17] proposed that Transformer cannot be used purely in image segmentation tasks. After being inspired by the ViT architecture, the Transformer encoder was introduced into U-Net for down-sampling. Using Transformer to solve the long-distance dependence of a convolutional network increases the network's overall perception and global understanding of the image. At the same time, combined with U-net, it can enhance finer details by recovering local spatial information. Compared with other Transformer models such as ViT, Trans U-Net not only shows better segmentation performance in liver, stomach, and other multi-organ segmentation but is also the first visual Transformer model for medical image segmentation.

Inspired by the above, a model named RDCTrans U-Net (ResNeXt50-Dilated Convolution-Transformer U-net) was proposed. The main contributions of this paper are as follows:

1. ResNeXt can solve the gradient problem caused by reducing the increase of residual connections, so this paper uses ResNeXt50 as the down-sampling backbone to increase the depth of the network without increasing parameters.
2. Dilated convolution is an effective kernel for adjusting the receptive field of feature points without reducing the resolution of feature maps. Therefore, this paper replaces all 3×3 convolutions in the last layer (Bottleneck $\times 3$) of ResNeXt50 with dilated convolutions and improves the segmentation performance by increasing the receptive field.
3. Since the Transformer can capture global information, this paper introduces the Transformer part in the encoder part to enhance the global context encoding ability of the overall structure and the ability to distinguish semantics.

Experiments conducted on a partial LiTS dataset show that the network in this paper has superior performance on the liver segmentation task compared to U-Net and some of its variants. We demonstrate the coordination and efficiency of ResNeXt50, dilated convolution, and Transformer in down-sampling tasks through ablation experiments.

2. Related Work

2.1. U-Net

In the context of medical images, the image is expensive and complex to acquire, and this also adds to the complexity of accurately annotating images [18]. However, CNNs have shown great potential in medical image segmentation in recent years [8,19], most of which is attributed to U-Net [5]. The structure of U-Net is very similar to Seg-Net [20], consisting of an encoder and a decoder. The difference lies in the skip connections between the encoder and decoder in each layer. The architecture and data expansion of U-Net allows the learning model to have a very good generalization performance from only a few annotated samples [21]. It has become a practical standard for medical image segmentation even when the amount of labeled training data are limited [10].

U-Net is derived from the idea of improving FCN [6], but it has many improvements compared to FCN. First of all, U-Net is completely symmetric, up-sampling uses adjacent interpolation, and the decoder is processed by convolution and deepening. Second, skip connections combine global and local features to form thicker features. Finally, U-Net uses valid convolution throughout to ensure segmentation results without missing contextual features. Because U-Net performs multi-scale fusion, combines low-resolution information and high-resolution information, and provides the basis for object category identification and accurate segmentation and positioning, it is very suitable for medical image segmentation.

2.2. Res-Net

The traditional convolutional network has the problem of information loss during information transmission, and it also causes the gradient to disappear or the gradient to explode, making the deep network unable to train. Res-Net [10] solves this problem to a certain extent. Its main idea is to add a direct channel to the network and protect the integrity of the information by directly detouring the input information to the output. Compared with VGG-Net [22], the biggest difference between Res-Net is that there are many bypasses to directly connect the input to the following layers. This structure is also called shortcut.

The residual structure is shown in Figure 1. By adding the identity mapping, the original function $H(x)$ that needs to be learned is converted into $F(x) + x$, that is, $H(x) = F(x) + x$, and the entire network only needs to learn the part of the difference between input and output. This idea stems from residual vector encoding in image processing. Through an information reorganization, the input and output of this module are superimposed at the element level. Not only does it not add extra parameters and computation to the network, but it can greatly increase the training speed of the model and improve the training effect.

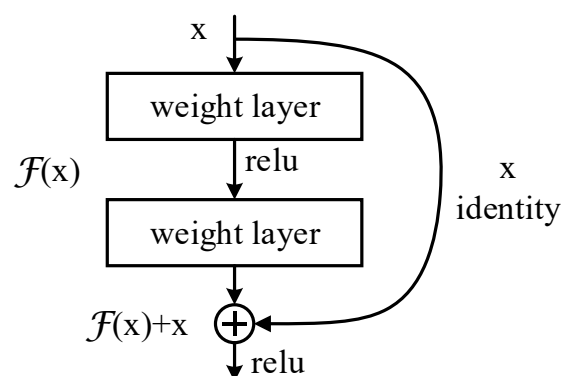


Figure 1. Residual structure.

2.3. ResNeXt

Although the proposal of the residual structure solves the problem of gradient disappearance caused by the deepening of the network layer, the modules of the same topology are stacked in Res-Net. This makes each component of the entire network more cumbersome, and the properties of the branches contained in each component are more variable. Inspired by the structure of the Inception series of networks [23,24], the ResNeXt [25] structure is designed, combining it with the residual structure in Res-Net, and simplifying the branch design method in Inception, making it modular. It can improve accuracy without increasing parameter complexity while reducing the number of hyperparameters. Especially when depth and width start to bring diminishing returns to existing models, increasing cardinality in ResNeXt is a more effective way to improve accuracy than increasing depth and width. Here cardinality is the size of the transformation set, a specific, measurable dimension of central importance.

2.4. Dilated Convolution

Common image segmentation algorithms usually use pooling and convolutional layers to obtain multi-scale contextual information. The pooling operation generally reduces the feature map size (resolution) first and then uses up-sampling to restore the image size. Although the receptive field of the neural network can be effectively increased, this operation of scaling down and then scaling up will lead to problems such as reduced resolution of feature maps and loss of spatial information. Therefore, there is a need for an operation that can increase the receptive field while keeping the size of the feature map unchanged, so as to replace the down-sampling and up-sampling operations. Under this requirement, dilated convolution was proposed by Yu et al. [26]. Dilated convolution can detect and segment large objects by expanding the receptive field without losing resolution, and increasing the resolution can precisely locate objects without introducing additional parameters or computational cost [27]. Different from the normal convolution, the dilated convolution introduces a hyper-parameter called “Hole Size”, which can get different perceptual field sizes by setting the number of holes, thereby capturing multi-scale context information and generating large-scale feature maps with rich spatial information, which can be effectively applied in the field of semantic segmentation.

2.5. Transformer

With recent advances in NLP research, some segmentation methods have explored alternatives based on channel or spatial [28,29] attention and pointwise [30] attention to better capture contextual information. However, these methods still rely mainly on convolutions and thus are more biased towards local interactions. The usual use of specialized layers to compensate for this bias shows the limitations of convolutional architectures for segmentation [28,31]. To overcome these limitations, Vaswani et al. [15] formulate the semantic segmentation problem as a sequence-to-sequence problem and exploit contextual information at each stage of the model using a Transformer architecture, which is entirely

based on Attention [14]. Transformer uses Positional Encoding to understand the order of language, self-attention mechanism (Self Attention Mechanism), and fully connected layer for calculation, which is a typical encoder-decoder structure. But the biggest difference compared with traditional CNN is its parallel training, which can greatly improve computational efficiency.

The great success of Transformers in NLP has also influenced the CV field [32,33]. In various medical image segmentation tasks, the success brought by the use of the U-Net architecture has become a reality. However, due to the inherent local nature of convolutional operations, U-Net typically exhibits limitations in explicitly modeling remote dependencies. This problem is addressed by the proposal of Trans U-Net [17]. Transformer, as a powerful encoder for medical image segmentation tasks, is combined with U-Net to enhance finer details by recovering local spatial information. In different medical applications for multi-organ segmentation, Trans U-Net achieves performance superior to various competing methods.

3. Methods

In this paper, a new segmentation architecture-RDCTrans U-Net (ResNeXt50-Dilated Convolution-Transformer U-Net) is proposed. The network structure is shown in Figure 2. In the encoder part, we use ResNeXt50 as shown in Figure 3b to extract feature maps and adopt skip connections in the U-Net structure to combine the multi-path feature maps of the mid-layer and deep layers, while using dilated convolution to refine the deep feature map of the fourth block of ResNeXt and fuse the global context information. The decoder part consists of four modules; each module contains one up-sampling block and two convolutional blocks; each convolutional block consists of convolutional layer, Batch Normalization, and ReLU. The structure is shown in Figure 4a.

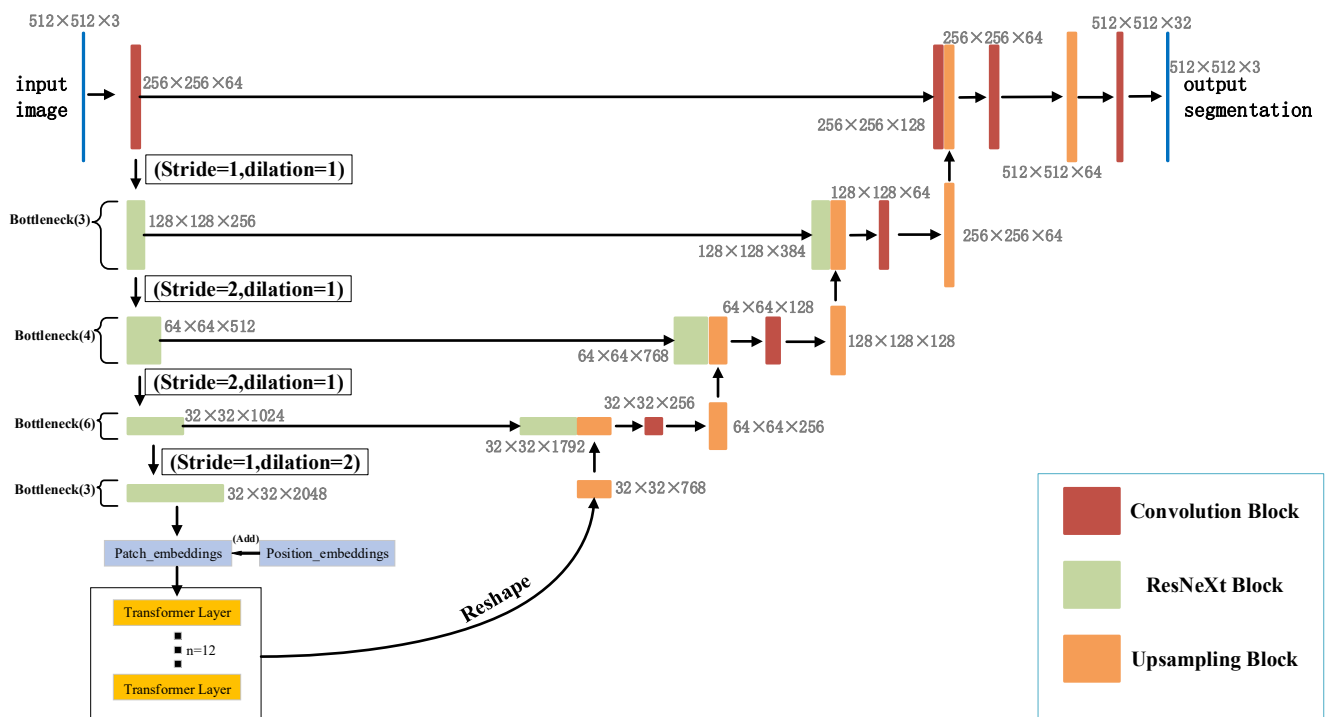


Figure 2. Network structure diagram of RDCTrans U-Net.

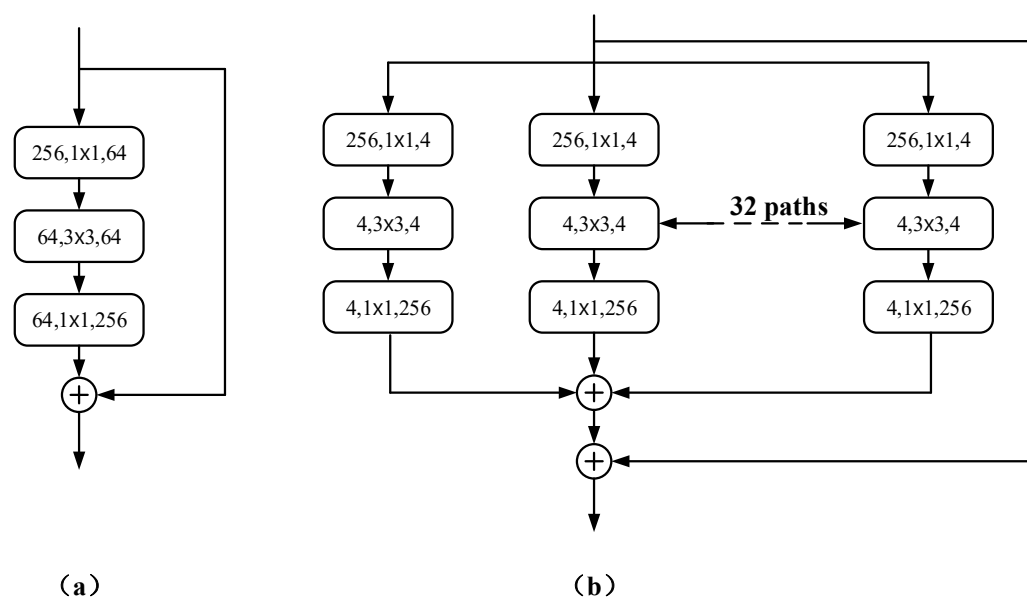


Figure 3. (a) An early version of Res-Net; (b) an aggregation block of ResNeXt50.

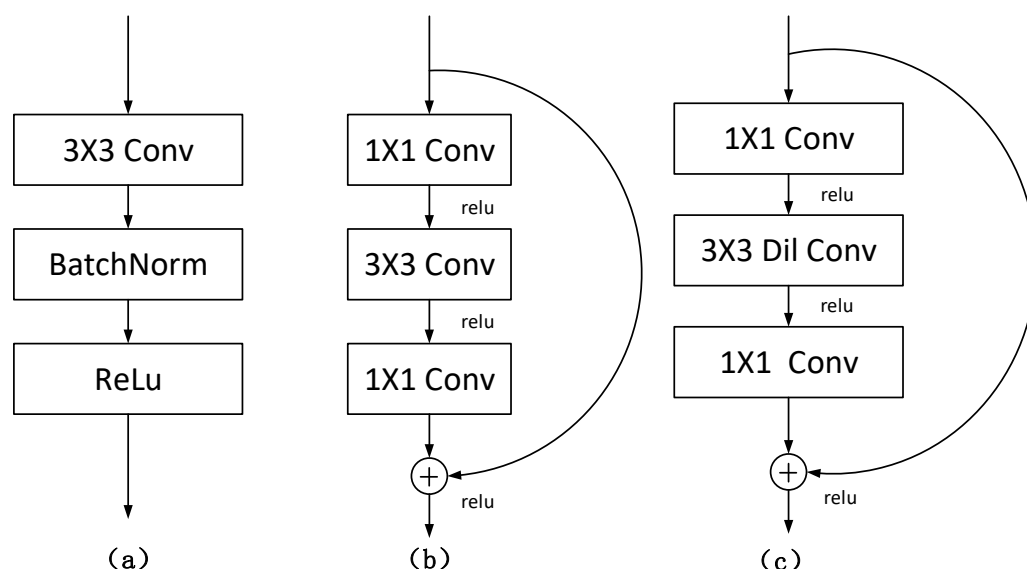


Figure 4. (a) The structure of convolution block; (b) the original structure of Bottleneck; (c) the structure of Bottleneck replaced with dilated convolution.

3.1. U-Net

U-Net [5] is divided into a down-sampling stage and an up-sampling stage, and the two stages use the same number of layers of convolution operations. The skip connection structure can connect the down-sampling layer with the up-sampling layer. After the channel features are extracted from the down-sampling layer, they can be directly transferred to the up-sampling layer, which greatly improves the segmentation accuracy. There is no fully connected layer in the network structure, and the shallow and deep layers are used to solve the problems of pixel localization and pixel classification, respectively, so as to achieve image semantic level segmentation.

In the experiment, in order to enhance the U-Net structure, the U-Net encoder part is improved into a hybrid encoder composed of three modules: ResNeXt50, dilated convolution, and Transformers, which are described in detail below.

3.2. ResNeXt50

ResNeXt [25] was created based on the idea of stacked networks such as Res-Net. It first aggregates a set of transformations with the same topology, then uses residual connections to augment blocks of multiple convolutional layers and generates gradient shortcuts that greatly reduce the risk of vanishing gradients, thus allowing training of deeper network structures.

Inspired by ViT [16], the ResNeXt50 encoder is used as the backbone to enhance the encoder structure of the baseline U-Net, where 50 represents the number of layers. The basic structure of the aggregation block in ResNeXt50 is shown in Figure 3b.

Following the highly modular design rule of Res-Net [10], it is only necessary to design template modules to determine all modules in the network. RDCTrans U-Net consists of one convolutional block and four residual blocks with the same topology. To adapt to the segmentation task, the global average pooling layers and fully connected layers in ResNeXt50 are removed. The Rectified Linear Unit (ReLU) [34] of Batch Normal (Batch Norm) [35] is added to the first convolution block. Batch Norm can make the network converge faster so that both the training set and the test set can remain independent and identically distributed and alleviate the overload problem of data initialization. The introduction of ReLU can solve the gradient vanishing problem of activation function back-propagation [36] in deep networks and alleviate the problem of network overfitting. The latter four residual blocks we replace are composed of bottleneck layers (Bottleneck) [37] with 3, 4, 6, and 3 stacked blocks, respectively, as shown in Figure 2. The use of bottleneck layers enables the reduction of parameters, making it possible to train and extract features from the data more efficiently and intuitively after dimensionality reduction. The cardinality of the aggregated blocks in ResNeXt50 is set to 32, as shown in Figure 3b.

3.3. Dilated Convolution

Dilated convolution [26] is a new type of convolution that allows aggregating multi-scale context. Dilation rates using kernels k and l of size M means sampling the input image with a stride of l , as shown in Equation (1):

$$y[i, j] = \sum_{n=1}^M \sum_{m=1}^M x[i + l * m, j + l * n] k[m, n] \quad (1)$$

All 3×3 convolutions in the last bottleneck layer of the ResNeXt50 structure are replaced by displacement-0 dilated convolutions with constant kernel size and dilation rate $l = 2$, as shown in Figure 4c. This operation can refine the fourth deep feature map, and the receptive field can be increased without changing the size of the feature map, so that ResNeXt50 can capture a wider range of contexts, and then better integrate global information.

3.4. Transformers

The feature maps learned from ResNeXt50 are divided into a series of patches. To better utilize the Transformer to learn location information, a learnable location embedding is performed on each patch to obtain the location matrix of N patches.

Transformers here is a collection of 12 concatenated Transformer encoders. The structure of a single Transformer encoder is shown in Figure 5, which is mainly composed of three parts: multi-head self-attention (MSA), multi-layer perceptron (MLP), and layer normalization (Layer Norm).

Multiheaded Self-Attention (MSA) can obtain more levels of semantic information and can reduce the total amount of calculation by reducing the dimension. It is essentially multiple independent Attention [14] calculations, and the role of integration is to prevent over-fitting. The definition of Multiheaded Self-Attention is shown in Equations (2) and (3).

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_h) W^o \quad (2)$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (3)$$

where Q , K , and V are single inputs from the same data; W^Q , W^K , and W^V are the weight matrices obtained by $linear(*)$ transformation with three different parameters, and $Concat$ denotes integration. i represents the number of multi, $i = 8$.

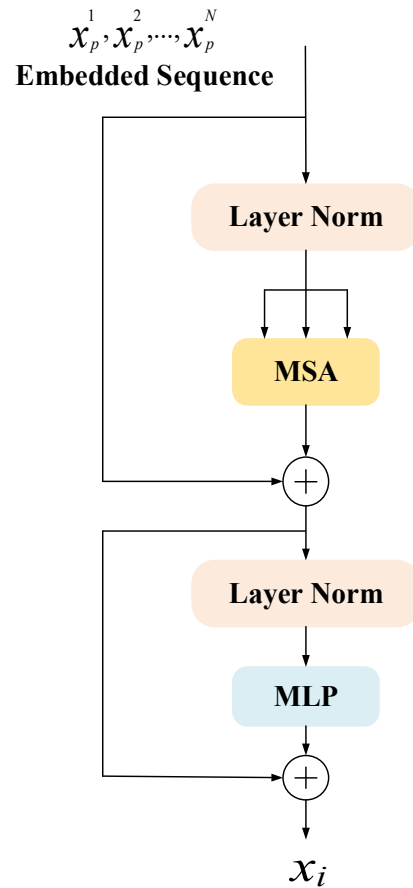


Figure 5. Encoder schematic of Transformer.

MLP (Multi-Layer Perceptron) contains three parts: input layer, hidden layer, and output layer, and the purpose is to realize the mapping from input to output. Each layer of it is fully connected to the next layer, which is called the fully connected layer, and the definition of the fully connected layer is shown in Equation (4).

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (4)$$

The fully connected layer here is a two-layer neural network that first maps the input Z to a higher-dimensional space, linearly transforms it, filters it through the nonlinear function ReLU, and then linearly transforms it to the original dimension.

Layer normalization (Layer Norm) [38] is to normalize all neurons in an intermediate layer, which can alleviate the problem of gradient disappearance and explosion in the early stage of training and improve stability. As shown in Equation (5).

$$\tilde{z}^{(l)} = \frac{z^{(l)} - \mu^{(l)}}{\sqrt{\sigma^{(l)} + \epsilon}} \odot \gamma + \beta \Leftarrow LN_{\gamma, \beta}(z^{(l)}) \quad (5)$$

where $z^{(l)}$ is the net input of the l -th layer of neurons, $\mu^{(l)}$ and $\sigma^{(l)}$ are its mean and variance, respectively, and γ, β represent the scaling and translation parameter vectors; the dimension is the same as $z^{(l)}$.

After the Transformer structure is introduced into the down-sampling part, it is convenient to use the parallelized training of Transformer to better capture full-text information. The architecture is entirely based on attention, which can suppress irrelevant background and highlight useful features, which helps to localize tumors quickly and accurately in the task of liver tumor segmentation.

4. Experiments

We used the LiTS public dataset for training to obtain a segmentation model for liver tumors. Compared with four advanced segmentation methods (U-Net [5], Attention U-Net [14], Dense U-Net [11], and Trans U-Net [17]), the method in this paper has the best segmentation results. We verify the effectiveness of our proposed model with comparative experiments on the LiTS dataset, including two ablation analysis experiments, as detailed in Section 4.4.

4.1. Datasets and Metrics

To validate the performance of our model in liver tumor segmentation, we use the MIC-CAI 2017 Liver Tumor Segmentation Challenge (LiTS) dataset [39] as the experimental dataset.

The segmentation of liver and its lesions in medical images is helpful for accurate diagnosis and therapeutic evaluation of liver cancer. The LiTS dataset includes 200 CT scans provided by clinical sites around the world, and images from each CT sequence provide liver and lesion areas through masks. CT imaging consists of three processes: First, an X-ray scan is obtained and converted into digital information. Then, the voxels are generated and separated by a computer, and the X-ray coefficients of each voxel are obtained and arranged into a digital matrix. Finally, the digital matrix corresponds to different grayscale values according to the values of its entries, and a CT grayscale image is obtained. In the LiTS dataset, each CT scan image contains a large number of axial slices, typically ranging from a few hundred to thousands, with an axial slice resolution of 512×512 pixels; the labels are divided into 3 categories: background (label 0), liver (label 1), liver tumor (label 2).

To train our proposed model, we randomly sampled 1371 2D slice images containing liver tumors from this dataset, where 1096 images are used as training set, 131 images are used as validation set, and 144 samples are used as test set. All images are normalized using Equation (6) to improve the overall training process speed.

$$value_{normalized} = \frac{value_{original} - mean}{std} \quad (6)$$

where $value_{original}$ and $value_{normalized}$ show the original image pixel value and the normalized image pixel value, respectively. Mean represents the mean value of the image pixels, and std represents the standard deviation of the image pixels. In addition, in order to satisfy the training of neural network and prevent over-fitting during model training, we also performed data flipping and data scaling operations on data images for data enhancement.

In the work covered in this paper, we use Mean Intersection over Union ($mIoU$) and Accuracy (Acc) to evaluate the comprehensive segmentation performance of our proposed model. Additionally, we use the Dice Similarity Coefficient (DSC), Precision (Pr), and Recall (Re) to evaluate the segmentation performance of the proposed model for liver and tumor, respectively.

$mIoU$ is the arithmetic mean of pixel-level intersection/union (IoU) [40] and n test image samples, defined as Equation (7). Acc is the accuracy rate [41], which refers to the proportion of correctly predicted pixels of the category (background class and target class) to the total image pixels, as shown in Equation (8). The Dice Similarity Coefficient (DSC) [42] is a measure of the ensemble similarity, as shown in Equation (9). Precision (Pr) is the ratio of the number of correctly predicted positive samples to the total number of predicted positive samples, as shown in Equation (10). Recall (Re) is the ratio of the number of correctly predicted positive samples to the total number of actual positive samples,

defined as Equation (11). The larger the value of these five indicators, the better the model segmentation performance.

$$mIoU_i = \frac{1}{n} IoU_i = \frac{1}{n} \sum_{i=1}^n \frac{TP}{TP + FP + FN} \quad (7)$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$Dice = \frac{2TP}{FP + 2TP + FN} \quad (9)$$

$$Pr = \frac{TP}{TP + FP} \quad (10)$$

$$Re = \frac{TP}{TP + FN} \quad (11)$$

where TP , TN , FP , and FN are pixel-level metrics representing the true, true negative, false positive, and false negative values in the confusion matrix, respectively.

The loss function used in the medical segmentation task in this paper is cross-entropy (CE) [43] defined as Equation (12).

$$CE(A, B) = -(A \log(B)) + (1 - A) \log(1 - B) \quad (12)$$

where B and A represent the predicted and underlying ground-truth splits, respectively. The loss function curve of the proposed model RDCTrans U-Net is shown in Figure 6. It can be seen that after 100 training iterations, the loss value is reduced to 0.1, indicating that the trained deep learning network model has converged.

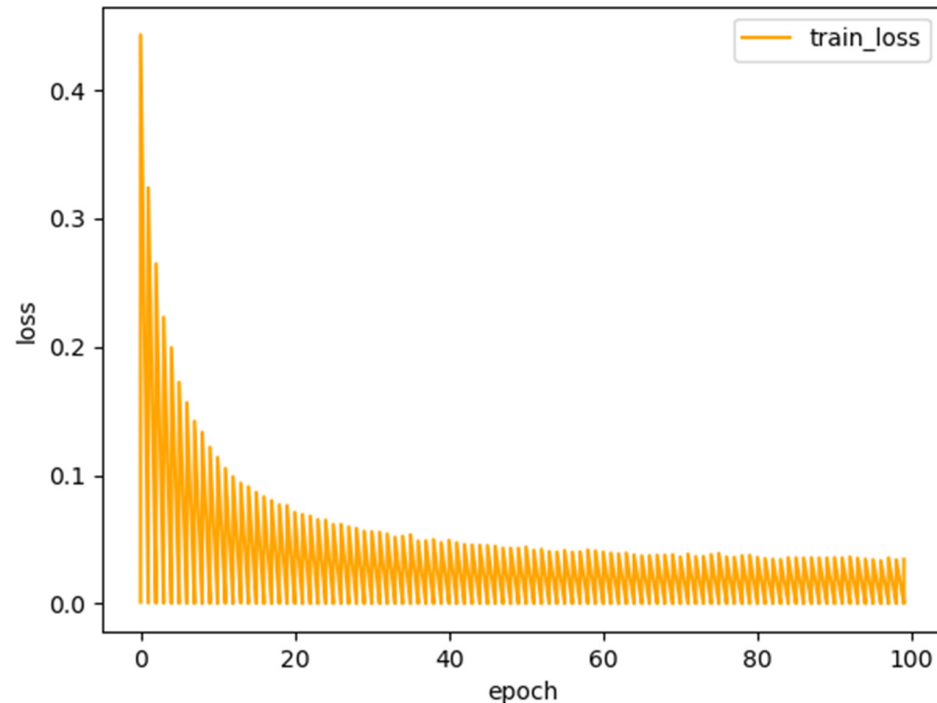


Figure 6. Loss function of the proposed model RDCTrans U-Net.

4.2. Experimental Details

In this experiment, the proposed RDCTrans U-Net is implemented in Python using the Pytorch deep learning framework. In addition, we use stochastic gradient descent (SGD) [44] with a batch size of 2, momentum of 0.9, and weight decay of 5×10^{-4} instead of Adam optimization [45], which according to a recent study [46] shows that SGD usually

leads to better performance, although Adam optimization converges faster. We trained the model using a fixed size training image (512×512) and trained 100 batches on an NVIDIA GeForce RTX 1080 Ti GPU. For a fair comparison, the parameters of all experiments were set to the same case. During the training process, the model that performs best on the validation set is selected as the final model. We use cross-entropy loss (CE) as the loss function to optimize the model.

4.3. Comparative Experiments

To verify the validity of the proposed model, we selected the original U-Net method and three recent popular U-Net variant models (Attention U-Net, Dense U-Net, and Trans U-Net) for comparison. In the stage of evaluating the model performance, firstly, in the LiTS dataset, we adopted the evaluation metrics of common medical image segmentation tasks—*Acc* (Accuracy) and *mIoU* (mean Intersection over Union), to verify the overall segmentation performance of our proposed model RDCTrans U-Net and the comparison model. The experimental results are shown in Table 1. Second, we use Dice similarity coefficient (DSC), precision rate (*Pr*), and recall rate (*Re*) to verify the segmentation performance of these models in the LiTS dataset for liver and tumor, respectively. The experimental results are shown in Table 2.

Table 1. Comparison of overall segmentation performance of each model on LiTS dataset.

Network Structure	Acc (%)	<i>mIoU</i> Score (%)
Original U-Net	96.65	74.29
Attention U-Net	98.06	83.09
Dense U-Net	96.93	79.18
Trans U-Net	98.17	83.32
RDCTrans U-Net	98.91	89.22

Table 2. The segmentation results of liver and tumor of each model on LiTS data set.

Network Structure	Dice (%)		Pr (%)		Re (%)	
	Liver	Tumor	Liver	Tumor	Liver	Tumor
Original U-Net	83.99	78.01	75.44	68.02	94.73	91.41
Attention U-Net	91.62	89.47	87.13	83.79	98.3	95.35
Dense U-Net	89.24	78.89	84.64	67.98	94.36	93.95
Trans U-Net	89.71	82.62	83.19	73.58	97.34	93.82
RDCTrans U-Net	93.38	89.87	88.65	86.52	98.89	94.31

As can be seen from Table 1, RDCTrans U-Net achieves the best performance on the LiTS dataset, with an *mIoU* value of 89.22% and an *Acc* of 98.91%. This is 14.93% higher than the *mIoU* value of the baseline method Original U-Net, and also much higher than the *mIoU* value of some classic U-Net variant methods such as Attention U-Net, Dense U-Net, and Trans U-Net. They are 6.13%, 10.04%, and 5.9% higher, respectively. Furthermore, compared with these classical methods, the *Acc* metric of our proposed model RDCTrans U-Net is 2.26% higher than the traditional U-Net, reaching 98.91%. This verifies the effectiveness of our proposed U-Net improved model.

As can be seen from Table 2, for liver segmentation, the Dice score, Precision, and Recall of the proposed model RDCTrans U-Net reach 93.38%, 88.65%, and 98.89%, respectively, which achieves the best liver segmentation performance compared to these contrasting algorithms. Meanwhile, for tumor segmentation, the Dice score and Precision of our model are higher than those of all contrasting models. It is worth mentioning that the Recall of the proposed model is higher than that of other models except that the Recall is lower than that of Attention U-Net by about 1%. This proves that the proposed model has an ideal segmentation effect in both liver and tumor.

To compare the complexity of each model, all models were trained on NVIDIA GeForce RTX 1080 Ti GPUs. Among them, the training time of Trans U-Net, Attention U-Net, and the

proposed model RDCTrans U-Net is about 20 h, the training time of Dense U-Net is about 23 h, and the training time of Original U-Net is about 16 h. Obviously, our model complexity is higher than the Original U-Net, but compared to these U-Net variant models, the model complexity does not increase. Through the analysis, it is concluded that ResNeXt is a structure that improves the accuracy without increasing the complexity of the parameters. Using it as the down-sampling backbone network will not increase the complexity of the model, while the complexity of Transformer is quadratically related to the input sequence, which leads to higher model complexity of the proposed model and Trans U-Net.

4.4. Ablation Experiments

To further determine the effectiveness of our proposed method, we performed two experimental analyses of ablation. During the ablation analysis, we only use the two metrics *Acc* and *mIoU* to evaluate the impact of each module on the model segmentation performance. Table 3 shows the comprehensive segmentation performance of the proposed model RDCTrans U-Net and two ablation models on the LiTS dataset.

Table 3. Results of the ablation study of the proposed model RDCTrans U-Net.

Network Structure	Acc (%)	<i>mIoU</i> Score (%)
ResNeXt U-Net	96.79	80.92
Dilated ResNeXt U-Net	97.32	83.15
RDCTrans U-Net	98.91	89.22

First, we remove the dilated convolution and Transformer modules and only use ResNeXt50 as the encoder for down-sampling, and we name the network model ResNeXt U-Net. Second, we only remove the Transformers module, combine ResNeXt50 and dilated convolution as the encoder part of the model, and name the model Dilated ResNeXt U-Net.

From Tables 1 and 3, we can see that ResNeXt U-Net achieves *mIoU* of 80.92% and *Acc* of 96.79% on LiTS dataset. Although it is not as good as the comprehensive segmentation effect of Attention U-Net, it is better than the comprehensive segmentation effect of the original U-Net, Dense U-Net and ResNeXt U-Net network. This proves the effectiveness of replacing the traditional CNN encoder with the Resnext50 encoder, which improves the segmentation accuracy of the model to a certain extent. To further expand the receptive field of the network, we add dilated convolution to the last bottleneck layer. The *mIoU* of Dilated ResNeXt U-Net on the LiTS dataset is 83.15%, and the *Acc* is 97.32%, which proves that adding dilated convolution can effectively improve the performance of the network.

In addition, Table 3 shows that RDCTrans U-Net achieves an *mIoU* of 89.08% and an *Acc* of 98.91% on the LiTS dataset. Comparing RDCTrans U-Net with ResNeXt U-Net and Dilated ResNeXt U-Net, our proposed model has the best overall segmentation performance. This further proves the effectiveness of our proposed hybrid encoder, and at the same time verifies that the multi-head self-attention in Transformers can effectively obtain global context information, making up for the shortcomings of convolution operations in this regard. In the encoding part, the combination of ResNeXt, dilated convolution, and Transformers can make the network segmentation of medical images more accurate.

4.5. Visual Analysis

We propose a deep learning model RDCTrans U-Net for liver and tumor segmentation, and to verify the effectiveness of our model, we conduct comparative experiments with other state-of-the-art methods. At the same time, the segmentation results of liver tumors were visualized on the LiTS dataset, as shown in Figure 7. We list 5 CT images containing liver and tumor to visually see that the segmentation results of our proposed model are significantly better than other models.

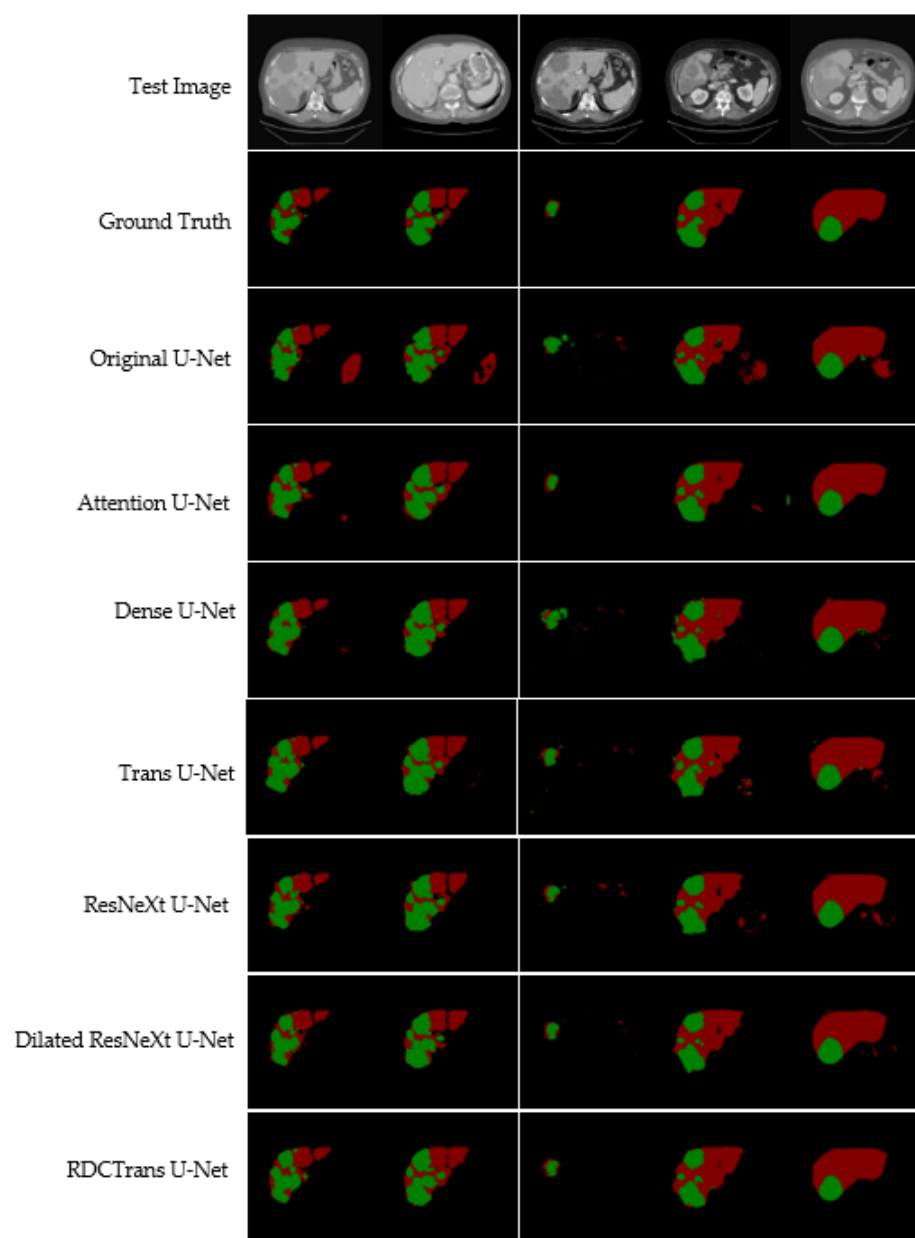


Figure 7. Illustrates the liver and tumor segmentation results of different methods on the test dataset. The red area represents the liver, and the green area represents the tumor.

Compared with liver segmentation, liver tumor segmentation is considered to be the most difficult segmentation task due to the variable shape and uncertain size of liver tumors. As can be seen from Figure 7, for the segmentation of liver and tumor, the segmentation images of the proposed model RDCTrans U-Net are closer to the Ground Truth map than those of other models. The original U-Net, Attention U-Net, and Dense U-Net over-segment or under-segment the liver, leading to poor segmentation results, which indicates that the Transformer-based model has stronger global context encoding ability and the ability to distinguish semantics. In addition, the Transformer-based Trans U-Net model also has a good segmentation effect because the Transformer encoder can learn the global contextual feature representation of the image, especially to encode the position information of the image, which definitely helps to improve the overall segmentation effect of the image. However, the segmentation accuracy of our proposed model is higher, which is attributed to the fact that the ResNeXt internal CNN and residual structure are more interested in some finer features on the image, such as tumor edge features. At the

same time, we replace the 3×3 convolution inside Bottleneck with dilated convolution to expand the receptive field, which can capture multi-scale context information and further improve the segmentation performance of the model. As can be seen from Figure 7, this paper combines the Transformer module with ResNeXt50 and dilated convolution, which effectively improves the segmentation accuracy of liver and tumor. In addition, Table 2 lists the detailed liver and tumor segmentation results. For liver and tumor segmentation, we achieve 93.38% Dice and 89.87% Dice, respectively, which is an ideal performance for liver and tumor segmentation. These observations demonstrate that RDCTrans U-Net is capable of finer segmentation and preserves detailed shape information.

5. Conclusions

In this work, we propose RDCTrans U-Net to meet the need for more accurate liver CT image segmentation tasks. The network structure proposed in this paper is designed based on the U-Net architecture, and only the encoder part is modified. First, down sampling with the ResNeXt50 encoder as the backbone, in which the 3×3 convolution in the last bottleneck layer of ResNeXt50 is replaced with a dilated convolution, increases the depth of the network and increases the receptive field. Then through the Transformers encoder integrated structure, the context global information can be effectively obtained, which greatly makes up for the shortcomings of U-Net in convolution. The reconstructed encoder has no pooling layer and fully connected layer, which maximizes the integrity of semantic information, and the results of liver image tumor segmentation are clearer and more accurate. In the training on the LiTS dataset, we not only confirmed the effectiveness of the network in this paper but also proved the coordination and efficiency of ResNeXt50, dilated convolution, and Transformer in down-sampling tasks through ablation experiments. However, RDCTrans U-Net also has some shortcomings. The RDCTrans U-Net proposed in this paper currently only performs segmentation on liver and tumor. In future work, we aim to further improve the structure of the proposed model so that it can be generalized to other medical image segmentation datasets and more flexibly applied to common medical image segmentation tasks to evaluate the segmentation performance of RDCTrans U-Net.

Author Contributions: Conceptualization, L.L. and H.M.; methodology, L.L.; software, L.L.; validation, L.L. and H.M.; formal analysis, H.M.; resources, L.L.; data curation, L.L.; writing—original draft preparation, L.L.; writing—review and editing, L.L. and H.M.; visualization, L.L.; supervision, H.M.; funding acquisition, H.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Cross-Media Intelligent Technology Project of Beijing National Research Center for Information Science and Technology (BNRist), grant number BNR2019TD01022.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: We evaluate our algorithm on the public dataset of the MICCAI 2017 Liver Tumor Segmentation Challenge (LiTS). The information link is: <https://competitions.codalab.org/competitions/17094> (accessed on 23 February 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ferlay, J.; Shin, H.R.; Bray, F.; Forman, D.; Mathers, C.; Parkin, D.M. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int. J. Cancer* **2010**, *127*, 2893–2917. [CrossRef] [PubMed]
2. Lu, R.; Marziliano, P.; Thng, C.H. Liver tumor volume estimation by semi-automatic segmentation method. In Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, Shanghai, China, 17–18 January 2006; pp. 3296–3299.
3. Bi, L.; Kim, J.; Kumar, A.; Feng, D. Automatic liver lesion detection using cascaded deep residual networks. *arXiv* **2017**, arXiv:1704.02703.
4. Dou, Q.; Chen, H.; Jin, Y.; Yu, L.; Qin, J.; Heng, P.-A. 3D deeply supervised network for automatic liver segmentation from CT volumes. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Athens, Greece, 17–21 October 2016; pp. 149–157.

5. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
6. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
7. Siddique, N.; Paheding, S.; Elkin, C.P.; Devabhaktuni, V. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access* **2021**, *9*, 82031–82057. Available online: <https://ieeexplore.ieee.org/abstract/document/9446143> (accessed on 13 March 2022). [[CrossRef](#)]
8. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Athens, Greece, 17–21 October 2016; pp. 424–432.
9. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
10. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
11. Li, X.; Chen, H.; Qi, X.; Dou, Q.; Fu, C.-W.; Heng, P.-A. H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Trans. Med. Imaging* **2018**, *37*, 2663–2674. [[CrossRef](#)] [[PubMed](#)]
12. Han, X. Automatic liver lesion segmentation using a deep convolutional neural network method. *arXiv* **2017**, arXiv:1704.07239.
13. Cai, Y.; Wang, Y. Ma-unet: An improved version of unet based on multi-scale and attention mechanism for medical image segmentation. *arXiv* **2020**, arXiv:2012.10952.
14. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Processing Syst.* **2017**, *30*, 6000–6010.
16. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
17. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
18. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J.A.; Van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [[CrossRef](#)] [[PubMed](#)]
19. Anwar, S.M.; Majid, M.; Qayyum, A.; Awais, M.; Alnowami, M.; Khan, M.K. Medical image analysis using convolutional neural networks: A review. *J. Med. Syst.* **2018**, *42*, 226. [[CrossRef](#)] [[PubMed](#)]
20. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [[CrossRef](#)] [[PubMed](#)]
21. Ibtehaz, N.; Rahman, M.S. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Netw.* **2020**, *121*, 74–87. [[CrossRef](#)] [[PubMed](#)]
22. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
23. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
24. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
25. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
26. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
27. Wei, Y.; Xiao, H.; Shi, H.; Jie, Z.; Feng, J.; Huang, T.S. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7268–7277.
28. Fu, J.; Liu, J.; Jiang, J.; Li, Y.; Bao, Y.; Lu, H. Scene segmentation with dual relation-aware attention network. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 2547–2560. [[CrossRef](#)]
29. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
30. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Loy, C.C.; Lin, D.; Jia, J. Psanet: Point-wise spatial attention network for scene parsing. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 267–283.
31. Yu, C.; Wang, J.; Gao, C.; Yu, G.; Shen, C.; Sang, N. Context prior for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12416–12425.
32. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
33. Parmar, N.; Vaswani, A.; Uszkoreit, J.; Kaiser, Ł.; Shazeer, N.; Ku, A.; Tran, D. Image transformer. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 4055–4064.

34. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the Icml, Haifa, Israel, 21–24 June 2010.
35. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
36. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [[CrossRef](#)]
37. Verma, D.; Kumar, M.; Eregala, S. Deep demosaicing using resnet-bottleneck architecture. In Proceedings of the International Conference on Computer Vision and Image Processing, Jaipur, India, 27–29 September 2019; pp. 170–179.
38. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
39. Bilic, P.; Christ, P.F.; Vorontsov, E.; Chlebus, G.; Chen, H.; Dou, Q.; Fu, C.-W.; Han, X.; Heng, P.-A.; Hesser, J. The liver tumor segmentation benchmark (lits). *arXiv* **2019**, arXiv:1901.04056.
40. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raskar, R.D. A challenge to parse the earth through satellite images. *arXiv* **2018**, arXiv:1805.06561.
41. Makridakis, S. Accuracy measures: Theoretical and practical concerns. *Int. J. Forecast.* **1993**, *9*, 527–529. [[CrossRef](#)]
42. Dice, L.R. Measures of the amount of ecologic association between species. *Ecology* **1945**, *26*, 297–302. [[CrossRef](#)]
43. Rubinstein, R. The cross-entropy method for combinatorial and continuous optimization. *Methodol. Comput. Appl. Probab.* **1999**, *1*, 127–190. [[CrossRef](#)]
44. Montavon, G.; Orr, G.; Müller, K.-R. *Neural Networks: Tricks of the Trade*; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7700.
45. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
46. Keskar, N.S.; Socher, R. Improving generalization performance by switching from adam to sgd. *arXiv* **2017**, arXiv:1712.07628.