IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON DEEP LEARNING FOR
FOR COMPUTER-AIDED MEDICAL DIAGNOSIS

Received January 16, 2019, accepted February 4, 2019, date of publication February 25, 2019, date of current version March 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2899608

AHCNet: An Application of Attention Mechanism and Hybrid Connection for Liver Tumor Segmentation in CT Volumes

HUIYAN JIANG, TIANYU SHI, ZHIQI BAI, AND LIANGLIANG HUANG
Software College, Northeastern University, Shenyang 110819, China
Corresponding author: Huiyan Jiang (hyjiang@mail.neu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61872075.

ABSTRACT The liver is a common site for the development of primary (i.e., originating from the liver, e.g., hepatocellular carcinoma) or secondary (i.e., spread to the liver, e.g., colorectal cancer) tumor. Due to its complex background, heterogeneous, and diffusive shape, automatic segmentation of tumor remains a challenging task. So far, only the interactive method has been adopted to obtain the acceptable segmentation results of a liver tumor. In this paper, we design an Attention Hybrid Connection Network architecture which combines soft and hard attention mechanism and long and short skip connections. We also propose a cascade network based on the liver localization network, liver segmentation network, and tumor segmentation network to cope with this challenge. Simultaneously, the joint dice loss function is proposed to train the liver localization network to obtain the accurate 3D liver bounding box, and the focal binary cross entropy is used as a loss function to fine-tune the tumor segmentation network for detecting more potentially malignant tumor and reduce false positives. Our framework is trained using the 110 cases in the LiTS dataset and extensively evaluated by the 20 cases in the 3DIRCADb dataset and the 117 cases in the Clinical dataset, which indicates that the proposed method can achieve faster network convergence and accurate semantic segmentation and further demonstrate that the proposed method has a good clinical value.

INDEX TERMS Liver tumor segmentation, deep convolutional neural network, feature fusion, attention mechanism.

I. INTRODUCTION

The liver is the largest solid organ in abdomen of the body, which has the complicated anatomical structure with vessels system. There are many kinds of liver tumors, and the incidence of liver disease is high, which has seriously threatened human health and life [1]. In recent years, Computed Tomography (CT) has become the most widely used medical imaging modalities in the diagnosis and treatment of liver tumors. The primary treatment methods include surgical resection, interventional therapy, locoregional ablation, *etc*. The size, shape, and location of the tumor are required in detail before therapy in order to develop a fine treatment program [2]. Therefore, accurate segmentation of liver tumor has become the primary task of liver tumor treatment, but there are several difficulties. Firstly, the size, shape, and location of the tumors are various among persons. Secondly, the boundaries between tumor and surrounding normal liver tissues are ambiguous, and some tumors may be adjacent to other organs and vessels. In addition, the immense diversity of tumors' appearances and inhomogeneous density make liver tumor segmentation become a challenging task. Therefore, research on automatic tumor segmentation algorithm has important value for guiding clinical diagnosis and treatment, not only can reduce the workload of manual segmentation, but more importantly, it can improve the accuracy of tumor segmentation, and form accurate preoperative prediction, introspective monitoring, and postoperative evaluation, which helps to develop a complete surgical treatment plan to improve the success rate of liver tumor surgery.

Researchers have attempted to address this challenge using graph cut [3], level set techniques [4], and machine learning [5], [6], however, limited by the robustness and generalization capabilities of these models, the interactive approach [7] remains the only acceptable liver tumor segmentation protocol.

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

2169-3536 © 2019 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

The latest developments in deep learning have revolutionized the field of artificial intelligence. Convolutional neural networks (CNNs), the representative of deep learning, have achieved great success in many areas of computer vision. Many researchers follow this trend and propose using various CNNs to achieve detection or segmentation of the liver and tumor.

Due to the high imbalance between liver tumors and background voxels in CT slices, it is difficult for Fully Convolutional Networks (FCNs) to accurately segment liver or tumor, particularly with small areas. Some work uses a cascaded approach to reduce training difficulty and computational costs while removing irrelevant information [8]. This approach is known as the hard attention mechanism [9]. Christ *et al.* [10] proposed to first divide the liver into regions of interest (ROI) through an FCN, and then train another FCN that is only used to segment tumor inside the liver ROI, and optimize segmentation result through dense 3D Conditional Random Field (CRF). Yuan [11] proposed a three-stage cascade FCN and used the Jaccard distance as a loss function. Bellver *et al.* [12] trained a detector to locate tumor, independent of a tumor segmentation cascade network and masked the results of segmentation networks with positive detection to reduce false positives. Due to the complexity of liver tumors, although hard attention mechanism can improve the accuracy of segmentation, it has not been effectively applied in the field of liver tumor segmentation.

Compared with hard attention mechanism, soft attention module can differentiate its input, which makes it possible to use the neural network to calculate the gradient and learn the weight of attention through back propagation [13], which also makes it possible for attention mechanism to be applied in the field of liver cancer segmentation. Xu *et al.* [14] proposed two models of image description generation based on attention mechanism, which was the earliest application of soft attention mechanism in CNN. Pian *et al.* [15] proposed a target classification model for visualizing the visual pathway of optic nerve information, which confirmed the importance of attention mechanism in CNN. Schlemper *et al.* [16] uses a self-gated soft-attention mechanism to generate a gating signal that is end-to-end trainable, which allows the network to contextualize local information useful for prediction. Oktay *et al.* [17] trained with Attention Gate (AG) implicitly learn to suppress irrelevant regions in an input image while highlighting salient features useful for a specific task, and demonstrate the implementation of AG in a standard UNet architecture (Attention UNet) and apply it to pancreas segmentation.

The short-range residual connection [19] and long-range copy and concatenation [20] can usually be used to improve CNN performance, and Drozdzal *et al.* [18] refer to them as short skip connections and long skip connections. Han [21] proposed a 24-layer FCN model to segment liver tumors, in which residual blocks were used as duplicate building blocks, and long skip connections were designed between the encoder and the decoder. Vorontsov *et al.* [22] also used ResNet-like residual blocks and long skip connections to connect 21 convolutional layers, which are shallower and have fewer parameters than those proposed by Han *et al.* [21]. Of course, the correct use of long and short skip connections will result in faster network convergence and retention of more discernable features that would otherwise lead to the opposite result. Therefore, it is necessary to formulate an appropriately long and short skip connections.

Extracting a single slice from the volume data and feeding it to the 2D FCNs [23] will limit the accuracy of the medical image segmentation by ignoring the context information on the z-axis. Some work uses adjacent slices as input to the network and uses 3D convolution kernels to probe spatial information along the third dimension. Dou *et al.* [24] uses the 3-D fully convolutional network structure to segment the liver while introducing a deep supervised mechanism during the learning process to combat potential optimization difficulties. Lu *et al.* [25] simultaneously perform liver and tumor segmentation by a 3D convolutional neural network, with graph cut refinement to automatically segment liver in CT scans. Li *et al.* [26] design a hybrid densely connected UNet to effectively probe hierarchical intra-slice features for liver and tumor segmentation, where the densely connected path and long skip connections are carefully integrated to improve the performance. In theory, more context information on the z-axis should lead to better network learning, but due to the size of the memory, improper selection of the number of slices may lead to network training difficulties.

In this work, we show a method for automatic segmentation of liver tumor in CT volume data. Experiments show that our method has high automatic segmentation accuracy and can effectively perform liver tumor segmentation. Our contributions are threefold. Firstly, we propose a 3D FCN structure, which is composed of multiple Attention Hybrid Connection Blocks (AHCBlocks) densely connected with both long and short skip connections and soft self-attention modules, to achieve fast and accurate semantics segmentation of medical images. Secondly, we use the hard attention mechanism in the form of the cascade network to complete the liver localization. At the same time, through the soft AG module, we also achieve the purpose of region proposal and eliminating the necessity of using explicit tumor localization module in liver tumor segmentation step. Our work combines fully differentiable soft attention mechanisms with cascaded hard attention mechanisms to demonstrate the importance of attention mechanisms in the segmentation of small organs or tumor in medical images. Third, we have designed two novel objective functions for different tasks of the cascade scheme. The joint dice loss function is used in the liver localization network to eliminate the need for class balance and improve the accuracy of the liver 3D bounding box. The focal loss is used to fine-tune the tumor segmentation network to reduce false positive, provide more potential tumor, and enhance the confidence of network output. It is worth mentioning that our work may be the first to apply soft-attention mechanism combined with CNN to tumor segmentation, and it has a

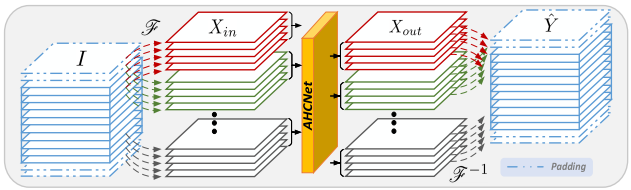


FIGURE 1. Schematic diagram of the data transformation process $\mathscr{F}$ and $\mathscr{F}^{-1}$.

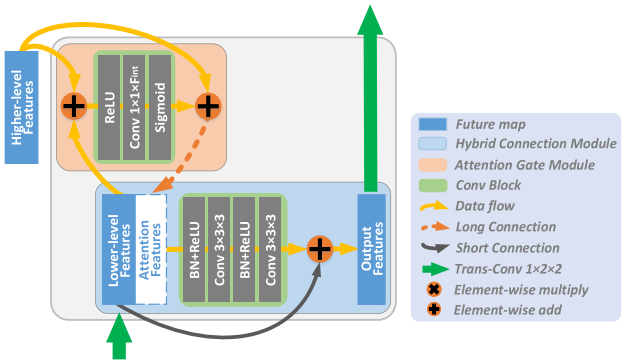


FIGURE 2. A block diagram of the proposed AHCBlock. It consists of two parts: the attention gate module and the hybrid connection block.

significant improvement in segmentation accuracy compared to other tumor segmentation models.

II. METHOD

A. DATA PREPROCESSING

Medical image volume data preprocessing was carried out in a slice-wise fashion. First, the Hounsfield unit values were windowed in the range $[-200]$, $[250]$ to enhance the contrast between the liver and the surrounding organs and tissues, so as to exclude irrelevant organs and tissues. Next, we crop the 512×512 scale data to 480×480 in order to reduce the amount of computation and increase the proportion of foreground regions. In addition to a series of geometric transformations, such as random clipping, flipping, shifting, scaling and tilting, Elastic Distortions [27] is used to extend the relatively real training data and further enhance the network generalization capability. We then performed a min-max normalization of every volume.

We use adjacent slices as the network input and use the 3D convolution kernel to detect volume data information along the third dimension (z-axis). Fig. 1 illustrates the detailed conversion process. Let the function $\mathscr{F}$ represent the transformation from the volume data to the network input data, and $\mathscr{F}^{-1}$ the transformation from the output data to the segmentation results:

$$X_{in} = \mathscr{F}(I), \quad X_{in} \in R^{n\times 5\times 480\times 480\times 1} \tag{1}$$

$$\hat{Y} = \mathscr{F}^{-1}(X_{out}), \quad \hat{Y} \in R^{n\times 480\times 480} \tag{2}$$

where X_{in} and X_{out} represent the input and output of the network respectively, $I \in R^{n\times 480\times 480}$ represents the volume data sample with the real label $Y \in R^{n\times 480\times 480}$, $\hat{Y}$ is the final prediction result of the network, and the batch size is n. This strategy of combining multiple sets of segmentation results are similar to the bagging method of ensemble learning, and can also play a role in data augment.

B. ATTENTION HYBRID CONNECTION BLOCK

In 2016, Ronneberger *et al.* [20] copied the encoder output of the same scale stage through a long skip connections and connected it in series with the decoder input to map the discriminable features into pixel space to achieve precise localization of the target details. In the same year, He *et al.* [19] performed matrix addition on the input and output of the residual module through a short skip connections, which solved the vanishing gradient problem without introducing additional parameters and complex calculations, and was more robust than the chain network [28]. In 2018, Oktay *et al.* [29] incorporate self-soft attention mechanism proposed AG module. AG module suppresses the unrelated regions in the input image by implicit learning through trainable models, and highlights the salient features useful for specific tasks, thus avoiding the over-use of computational resources and model parameters caused by hard-attention models such as cropping the ROI [8]. From the perspective of features, short skip connections are used to improve the existing features, long skip connections are used to explore new features to complement existing features [30], and AG prefers to select more efficient features [31]. We designed a network module called AHCBlock, which consists of the AG module and the hybrid connection module, as shown in Fig. 2.

The AG module performs feature selection on the coarse-scale context information obtained by the long skip connections to improve the sensitivity of the model to the foreground pixels. The input of the AG module takes the matrix addition of the higher-level feature map and the lower-level feature map and retains the positive activation feature through ReLU. The sigmoid function is used as the normalized function to calculate the attention coefficient (α), by the following formula:

$$\alpha = f_\alpha(x) = Sigmoid(W_1^{\mathrm{T}} \cdot ReLU(x_l \oplus x_h)) \tag{3}$$

where $f_\alpha(x)$ is the attention-guiding function, α is called the attention coefficient, W_1^{T} represents the linear transformation of the convolution operation of the current block, $\alpha \in \{0, 1\}$, x_l and x_h represent the lower-level features and higher-level features, respectively, $\oplus$ refers to the matrix addition, $Sigmoid(x) = 1/\left(1 + e^{-x}\right)$.

The output of the AG module is obtained from the element-wise product of the higher-level feature map and attention coefficient, and serially connected to the hybrid connection module as the input of the long skip connections. The soft attention mechanism enables the neural network to focus on selecting a specific feature subset. The AG module obtains the following attention features:

$$Z_\alpha = \alpha \odot x_h \tag{4}$$

where Z_α is the attention features, and $\odot$ refers to the element-wise product, which uses α as the attention guidance function of the higher-level features, and works with the lower-level features to achieve the attention mechanism. The AG module is fully differentiable, so no step-by-step training is required, and can be easily integrated into a standard CNN architecture while improving model sensitivity and prediction accuracy without the significant computational overhead.

In the hybrid connection module, the metric addition is performed by inputting the lower-level feature map and the output feature map of the module to implement a short skip connections, thereby improving the information flow in the block, making it easy to train the deep network, which may be formulated as follows:

$$I(x_l, x_h) = f_{plain}(x_l, Z_\alpha, \{W_i^{\mathrm{T}}\}) \oplus x_l \tag{5}$$

$$\sigma = ReLU(Batch\_norm(x)) \tag{6}$$

$$f_{plain}(x_l, Z_\alpha) = W_3^{\mathrm{T}} \cdot \sigma(W_2^{\mathrm{T}} \cdot \sigma(x_l + Z_\alpha)) \tag{7}$$

where $I(x_l, x_h)$ is feature integration function, σ represents the nonlinear transformation in the convolutional block, where *Batch_norm* [32] is used for reduce the impacts of earlier layers by keeping the mean and variance fixed, and $ReLU(x) = max(0, x)$. $f_{plain}(x_l, Z_\alpha)$ represents the simply stack layers. The output of the current hybrid connection module acts as the input to the next hybrid connection module:

$$x_{h+1} = v(f_{plain}(x_l + \alpha(x_l + x_h)) + x_l) \tag{8}$$

where v represents the upsampling operation. That is, we fuse higher-level and lower-level features through AG module, in which the 1×1 kernels achieves global average pooling of feature tensors to get the attention of channel domain [33]. In this case, the gating signal does not act on the local features but instead focuses on the overall information of the gating signals to calculate the correlation between the channels. Then we get the activation function of the average value of the feature tensor in the channel domain to get the attention of the spatial domain [34]. In this way, by controlling the feature intensity, the directivity of the extracted feature is enhanced, and the correlation in the spatial distribution is calculated by the information in the channel. The mixed attention mechanism [35] formed by the two attention mechanisms highlights the distinguishable features of the higher-level features and extracts context semantic features aggregates with lower-level features. Finally, the short skip connections prevents the vanishing gradient problem caused by the too little attention information and passes the feature to the next AHCBlock through the up-sampling or long skip connections.

C. ATTENTION HYBRID CONNECTION NET

The previous semantic segmentation work is usually the structure of the encoder-decoder. However, the fine-grained information lost by multiple pooling operations in the encoder structure is difficult to recover by gradually recovering the single-scale information. Some current work [36], [37] is devoted to solving how to learn multi-scale feature representations with strong semantic information. However, in the Atrous Spatial Pyramid Pooling (ASPP) modules proposed in [36], the sparse computation of atrous convolution may result in grid artifacts, which may interrupt the continuity between local information, and pyramid pooling module will largely lose pixel location information, resulting in gridding effect. However, the performance of the Pyramid Pooling module proposed in [37] is still unsatisfactory for high-dimensional feature representation, resulting in the loss of spatial resolution of context pixels in the original scene. Therefore, we propose the AHCNet, which integrates the higher-level semantic features and the lower-level location information by combining both long and short skip connections and soft attention mechanism to complete the fine-grained information recovery of the medical image. The network model is shown in Fig. 3.

Each AHCBlock takes the output of the previous AHCBlock of the same scale as higher-level features, and the output of the previous smaller-scale AHCBlock is up-sampled as lower-level features. The long skip connections and the self-soft AG module perform context semantic integration, and the short skip connections is used in the block to overcome the optimization difficulty, and finally, the integrated feature is output as higher or lower-level features input of the next AHCBlock. The topological connection mode of each AHCBlock is similar, which integrated from shallow to deep to make a progressively deeper and higher resolution decoder, thus achieving the features dense integration. We don't use AHCBlock at the initial scale because we wanted to focus on the lower-level features rather than wasting memory on the higher-level features.

The structure of AHCNet is inspired by the recurrent structure and attention mechanism [38]. It makes full use of AHCBlock structure to refined feature maps repeatedly and maximizes the flow of information needed to be concerned. These dense connection structure benefits feature re-usage and iterative learning. Each back propagation refines the features from top to bottom, realizes the spatial attention mechanism [39], while suppressing the image background, it concentrates more activation on the region representing the target object. This connection method enables AHCBlock to guide the selection of relevant lower-level features during the training process while acquiring and delivering higher-level semantic features [40]. At the same time, because AHCNet is an end-to-end network, different blocks participate in supervised learning at the same time, and because of the network topology structure of hybrid dense connections, it can effectively extract more discriminatory strong semantic information which takes into account both the semantic and spatial information while accelerating the network training speed.

D. CASCADE ARCHITECTURE

Medical images consist of regions of interest and background regions, in which the regions of interest contain

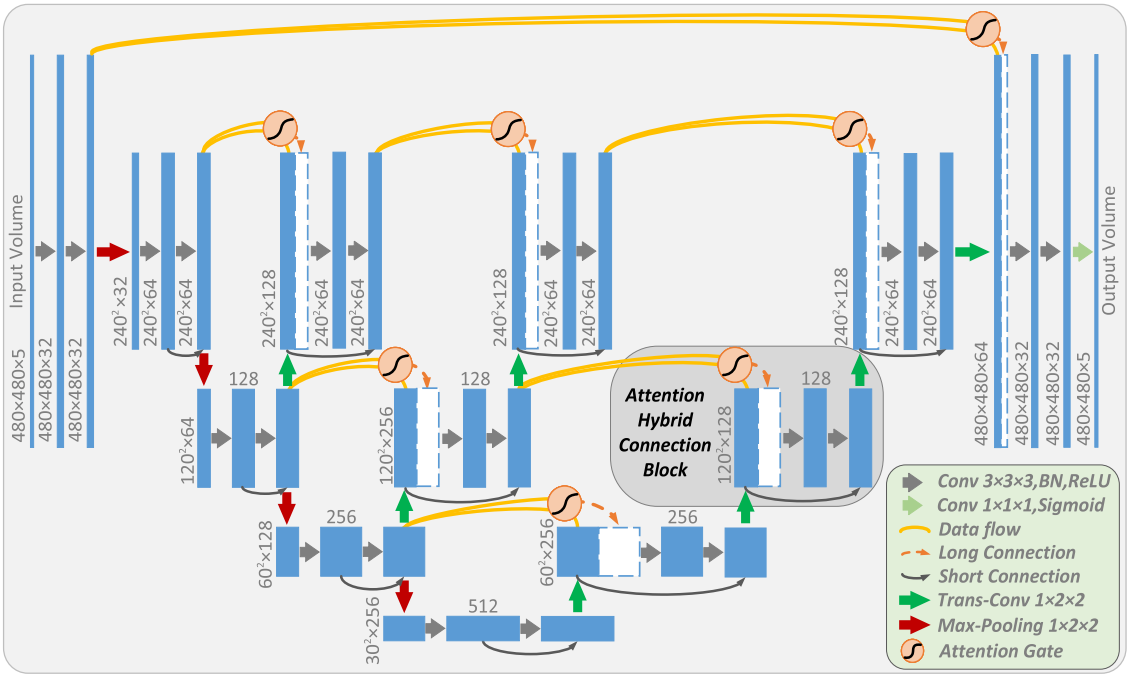


FIGURE 3. The Attention Hybrid Connection Net structure presented in this paper.

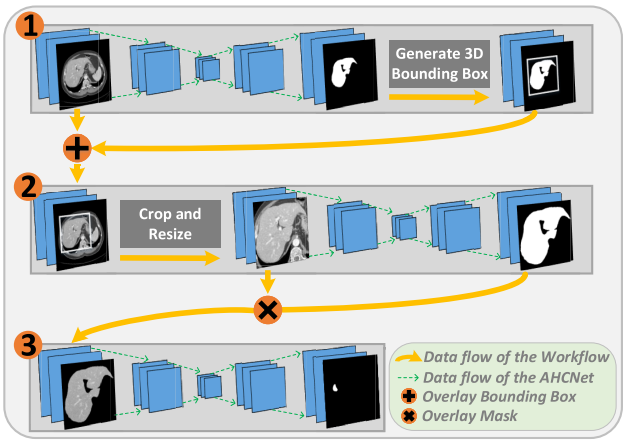


FIGURE 4. Overview of the proposed liver and tumor segmentation workflow. Stage 1, 2, and 3 represent the tasks of liver localization, liver segmentation, and liver tumor segmentation, respectively.

important diagnostic information and can provide a reliable basis for clinical diagnosis and pathological research. Usually, in order to prevent interference to the intraoperative inter-organizational, it is necessary to perform background segmentation to remove the redundant portion of the error description, and then further analyze and identify each sub-region within the ROI to determine the surgical plan. Similarly, our framework is hierarchical. Considering that precise segmentation of the liver is the premise of tumor segmentation, we design a cascade segmentation model for contrast enhancement abdominal CT images of liver and lesions, in order to reduce the false positives. The workflow of the cascaded network is shown in Fig. 4:

The first AHCNet learns liver-specific filters and performs liver coarse segmentation on the entire abdominal CT volume. Once all slices have segmented, apply the threshold of 0.5 to the output of AHCNet, and select the 3D connect-component labeling of the segmentation result, the largest connected component is defined as the initial liver region. Based on the largest boundary of the initial 3D liver region, expand 10 voxels in three directions to extract the 3D bounding box of the liver region, thereby increasing the area of the liver foreground region. This network is called a liver localization network.

The second AHCNet is called liver segmentation network, which focuses on learning the discriminative features of the liver and background in the 3D bounding box of the liver region. This network resizes the 3D bounding box of the liver to needed size as input and apply a threshold of 0.5 to the output as a mask and overlays onto the original CT image to generate the liver volume of interest (VOI), thus to exclude extra complex background of non-liver organ formation.

The last AHCNet performs a 3D regional histogram equalization to the liver VOI as input to enhance the contrast between the tumor and surrounding liver tissue [11]. The network only focuses only on learning the discriminant features of tumor and background and applies a threshold of 0.5 to the output to separating tumor from the liver tissue, which called a tumor segmentation network.

E. LOSS FUNCTION

A large area of background usually leads to the learning process to get trapped in local minima of the loss function yielding a network with strong background bias prediction [41],

so the selection of the loss function is very important. We designed two different loss functions for the liver localization network and the lesion segmentation network, respectively.

1) JOINT DICE SIMILARITY COEFFICIENT

Dice similarity coefficient (DSC) is a common evaluation index of the segmentation effect in medical image analysis, which is divided into global dice similarity coefficient and mean dice similarity coefficient [42]:

1) Global dice similarity coefficient is evaluated by counting the binary volume data of all slices, defined as:

$$DSC_{global} = \frac{2\sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2} \tag{9}$$

The total number of voxels i is N, the predicted binary segmentation voxel $p_i \in P$ and the ground truth binary voxel $g_i \in G$.

2) The mean dice similarity coefficient is evaluated by counting the binary volume data of each slice, and then counting the average score of all the slices, which is defined as:

$$DSC_{mean} = \frac{2}{m}\sum_j^m \frac{\sum_i^n p_i g_i}{\sum_i^n p_i^2 + \sum_i^n g_i^2} \tag{10}$$

The total number of slices j is represented by m, and n represents the total number of voxels per slice.

When choosing the global dice similarity coefficient to train the network, the mean dice score may be relatively low due to the over-attention to the overall index, that is, the segmentation effect of small liver or tumor may be poor. Correspondingly, when the mean dice similarity coefficient is selected, the liver or tumor segmentation effect of the larger area may not be fine enough due to excessive attention to the average index. Therefore, the joint dice similarity coefficient is proposed in this paper, the formula is as follows:

$$L_{jointDSC} = \frac{-2 \cdot DSC_{global} \times DSC_{mean}}{DSC_{global} + DSC_{mean}} \tag{11}$$

That is, the harmonic mean of the global dice similarity coefficient and the mean dice similarity coefficient is used as the loss function. Fusing a network output with different loss functions to produce the ultimate prediction is usually more accurate [43]. The joint dice similarity coefficient is used to emphasize the importance of the smaller of the global dice similarity coefficient and the mean dice similarity coefficient, which is inspired by the F1-mesure [44]. We aim to minimize it to ensure a more comprehensive segmentation effect. We apply the joint dice similarity coefficient to the liver localization network to reduce over-fitting and guide the network to learn more balanced features in different slices, rather than paying more attention to the larger liver area.

2) FOCAL LOSS FUNCTION

Classification tasks in the medical field usually require identification of normal tissue and tumor areas. The distribution of data for various tissue categories in each case is generally unbalanced, and most normal tissue and organ training samples are highly correlated and are often over-represented. Equal treatment of these data during the learning process can result in a lot of training iterations being wasted on non-information samples, making the network training consume unnecessary time [44]. Meanwhile, network training dominated by healthy tissue samples will also lead to problems in the CNN model trained. In the liver tumor segmentation task, the ratio of foreground pixels to background pixels is very different, so tumor segmentation is more challenging than liver segmentation. The use of dice similarity coefficient or weighted cross-entropy is limited in the extreme class imbalance problem.

Focul loss [45] provides another solution. This function can reduce the weight of easy-to-classify samples and make the model more focused on difficult-to-classify pixels in training, thus achieving good performance on small object segmentation. This loss function is modified based on the standard binary cross-entropy(BCE) loss function. Compared with the dice loss, the binary cross-entropy loss function can provide more suspicious range and results of liver cancer while ensuring high-precision prediction [18].

$$L_{Focal} = \begin{cases} -\alpha(1-g_i)^{\gamma} \log\ g_i & p_i = 1 \\ -(1-\alpha) g_i^{\gamma} \log\,(1-g_i) & p_i = 0 \end{cases} \tag{12}$$

where g_i represents the probability that voxel i belongs to the foreground, and p_i represents the ground truth, the weighting factor α adjusts the weight of the positive and negative samples, $(1-g_i)^{\gamma}$ is called modulation factor, which is used to control the weight of easy and hard examples. This makes focal loss equivalent to smooth weighted cross-entropy loss, which makes the loss curve very smooth and the drop is stable. We determined that $\alpha = 0.3$ and $\gamma = 1.4$ had the best effect in the experiment.

Focal loss slightly inhibits the contribution of incorrectly segmented pixels while severely suppressing the correct segmentation of pixels, thus automatically extending the loss gap between the correctly segmented and the incorrectly segmented pixels.

In the experiment, we choose to use the focal loss to fine-tune the network on the model which has been trained with BCE as the loss function. On the one hand, this avoids the need for careful initialization to avoid instability at the beginning of network training, otherwise, the network will not converge [45]. On the other hand, the focal loss only needs to eliminate false positives and reduce the rate of missed detection, which will speed up network training [46]. And the threshold of the lesion segmentation network using focal loss fine-tuning is usually set higher than that of the network trained using the BCE loss function, which indicates

that the result of the model fine-tuned by focal loss is more reliable [47].

III. EXPERIMENTS

A. DATASET

It is important to state that the 3DIRCADb dataset is a subset of the LiTS dataset with case numbers 28 to 47. Therefore, the way to directly use the LiTS dataset as additional data and verify on the 3DIRCADb dataset is not allowed. We trained the network on the LiTS dataset after 3DIRCADb dataset was removed, then tested the network on the 3DIRCADb dataset and Clinical dataset to facilitate comparison with other deep learning-based segmentation models.

1) LiTS DATASET

The LiTS dataset[1] consists of 130 contrast-enhanced 3D abdominal CT scans from six different clinical sites with different scanners and protocols, thus have largely varying spatial resolution and field-of-view (FOV). The data and segmentations are provided by several clinical sites, scanners, and protocols. The voxel dimensions are $[0.60-0.98, 0.60-0.98, 0.45-5.0]$ mm. Most CT scans are pathological, including tumors of different sizes, metastases, and cysts. The axial slices of all scans have an identical size of 512×512, but the number of slices in each scan differs greatly and varies between 42 and 1026.

2) 3DIRCADb DATASET

We evaluated our proposed method on the 3DIRCADb dataset[2] [51]. The 3DIRCADb dataset includes 20 venous phase enhanced CT volumes from various European hospitals with different CT scanners. The liver dimensions are $[16.3-24.9, 12.0-18.6, 11.0-20.2]$ cm, of which 15 volumes containing hepatic tumors in the liver, and the number of tumors varied from 1 to 46. The voxel dimensions are $[0.56-0.87, 0.56-0.87, 1.6-4.0]$ mm, and the number of slices in each scan varies between 74 and 260.

3) CLINICAL DATASET

The dataset is a real-life clinical CT dataset from multiple CT scanners, including 117 patients and corresponding tumor segmentation masks. The in-plane resolution ranges from 0.71 mm to 1.17 mm, and the slice spacing from 1.25 mm to 5.0 mm. The number of slices varied between 42 and 84 in each scan. The patients examined had different kinds of tumor in the liver. In addition, there are different contrast agents in this data set, so there are varying degrees of contrast enhancement. The scan was manually selected and labeled by two radiologists using the software ITK-SNAP.[3]

[1] https://competitions.codalab.org/competitions/17094
[2] https://www.ircad.fr/research/3dircadb/
[3] http://www.itksnap.org/

B. TRAINING DATA SELECTION

In the LiTS dataset, there is a problem of a significant long-tailed distribution in the real world, that is, due to the large difference in the number of slices, the intra-class data imbalance is caused, which leads to the model being more biased towards the number of slices. So we re-sampled all the training images to a fixed resolution of 0.69 mm $\times$ 0.69 mm $\times$ 1.0 mm when we constructed the liver training data set. The training data for the tumor segmentation network were collected using only slices belonging to the liver region so as to focus the training on the liver and liver tumor. Data preprocessing was performed as described in Section 2.1 and 90/20 patients were randomly divided for training and validation of the network, i.e., to determine when to stop training to avoid over-fitting, and the remaining 20 patients were used to test the network.

C. EXPERIMENTAL SETTING

The AHCNet model is implemented using the publicly available keras package [48], Xavier normal distribution initializer initializes the weights, and trained with an Adam optimizer with a learning rate of $1e^{-4}$. Since the models selected in the three stages of our cascaded network have structural consistency, the cascading fine-tuning method is adopted, that is, the latter stage network is fine-tuned with the model of the previous stage network to achieve convergence of the model in a shorter time. Purpose, this parameter-transfer learning may also lead to better segmentation effects [49]. Training of each model took about 30 hours using a single NVIDIA GTX1080Ti GPU with 3584 cores and 11GB memory. The total processing time for final tumor segmentation thus depends on the number of slices for each scan, which ranged from 50 seconds to 200 seconds for the clinical test data. Finally, the output of the network consists of a probability map of the background and the foreground, and the foreground voxels with a higher probability than the background (>0.5) are considered to be part of the anatomical structure.

IV. RESULTS AND ANALYSIS

A. LOSS CURVE

We first analyze the learning process of liver segmentation network. The control model included 3D UNet with a similar number of parameters as our suggested architecture (Weight UNet), AHCNet without Hybrid Connection (AHCNet w/o HC), AHCNet without Dense Connection (AHCNet w/o DC) and AHCNet without Attention Gate (AHCNet w/o AG). As can be seen from Fig. 5, the AHCNet convergence is much faster and smoother than the control model. These results demonstrate the attention mechanism and hybrid connection can effectively speed up the training procedure by overcoming optimization difficulties through managing the training of both upper and lower layers in the network.

Fig. 6 shows the loss curve of the liver localization network training process. The metrics used for training loss is the global dice coefficient, and the metrics verification accuracy

TABLE 1. Segmentation results by ablation study of our methods on the 3Dircadb dataset. Statistically significant results are highlighted in bold font.

Model	Liver		Tumor						
	Dice per case	Dice global	Dice per case	Dice global	VOE(%)	RVD(%)	ASSD(mm)	MSD(mm)	RMSD(mm)
AHCNet w/o HC	0.837	0.861	0.542	0.589	2.474	0.481	1.541	8.537	1.854
Weight UNet	0.923	0.937	0.562	0.618	1.643	0.338	1.373	7.359	1.594
AHCNet 1-i/o	0.938	0.945	0.553	0.598	1.710	0.378	1.437	7.684	1.596
AHCNet w/o DC	0.949	0.957	0.657	0.690	1.361	0.133	1.216	6.870	**1.394**
AHCNet w/o AG	0.945	0.953	0.659	0.725	1.378	0.136	1.182	6.463	1.547
AHCNet	**0.953**	**0.959**	**0.668**	**0.734**	**1.354**	**0.129**	**1.074**	**6.271**	1.412

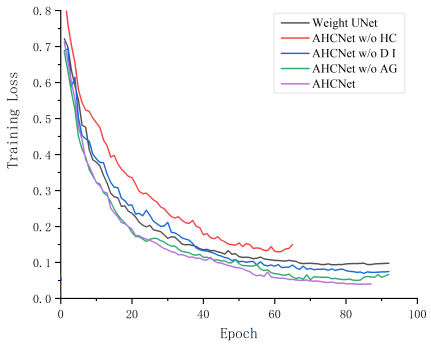


FIGURE 5. The learning process of liver segmentation network with different model. The x-axis is the number of training epochs. The y-axis is the training loss.

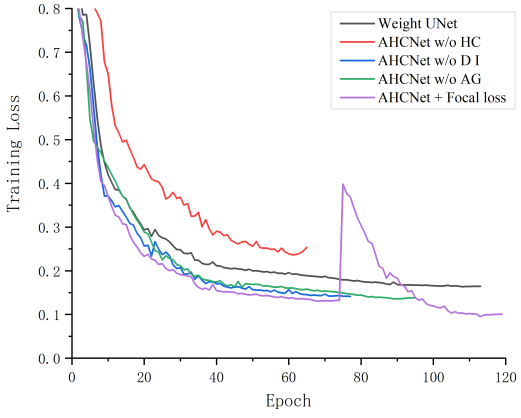


FIGURE 6. The learning process of liver localization network with different loss functions.

is binary cross entropy. The loss curve of the network trained with joint DSC is close to the network trained by global DSC, and the accuracy is higher. This indicates that the slice of the small area liver has a greater influence on the mean DSC, so the joint of the global DSC and the mean DSC can obtain more accurate localization results.

Furthermore, as shown in Fig. 7, it can be observed that AHCNet with focal loss fine-tune achieves the lowest training loss in the liver cancer segmentation network, which confirms that focal loss can improve the discriminating ability

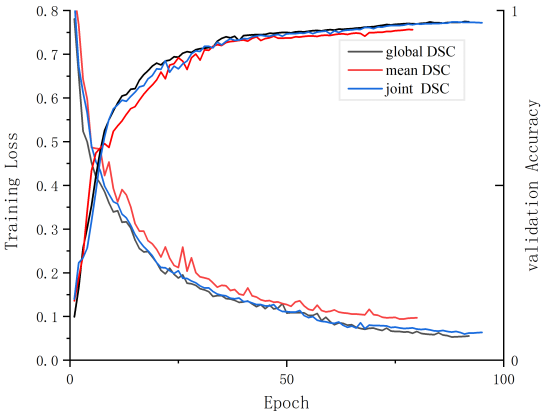


FIGURE 7. The learning process of liver localization network with different model, and fine tune AHCNet by the focal loss.

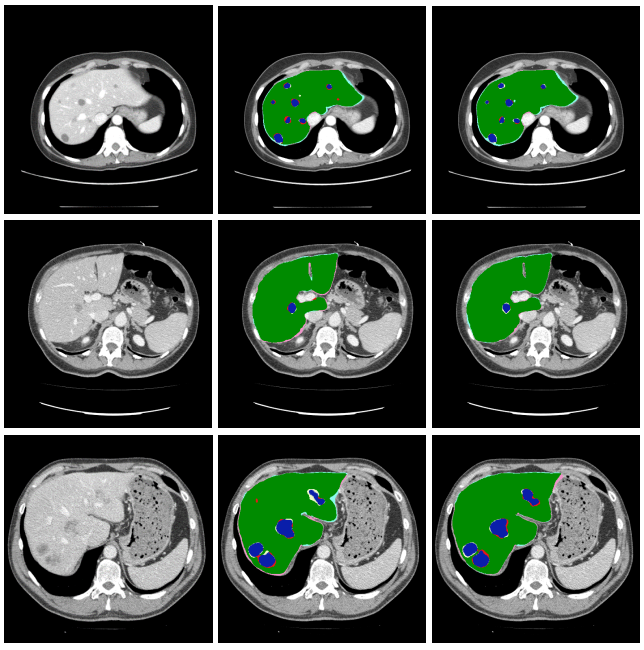
of the network. And AHCNet w/o AG performed poorly, indicating the importance of AG in small target segmentation. Because the tumor segmentation task is more complicated than liver segmentation, the superior effect of AHCNet is more obvious than the liver segmentation task. These results present the effectiveness of the voxel-to-voxel error back-propagation learning strategy benefiting from the AHCNet architecture.

B. PERFORMANCE EVALUATION

To validate the effectiveness and robustness of our approach, we conducted experiments on the clinic datasets and a subset of LiTS datasets. The subset of the LiTS dataset refers to the 3DIRCADb dataset, which is publicly available and provides a more accurate label. Table 1 and Table 2 show our method's performance on the 3DIRCADb dataset and Clinical dataset compared. We use seven metrics measure the accuracy of segmentation results, including the dice per case, dice global, volumetric overlap error (VOE), relative volume difference (RVD), average symmetric surface distance (ASSD), the maximum surface distance (MSD), and root mean square symmetric surface distance (RMSD). The smaller the value of the last five evaluation metrics, the better the segmentation results. Where Weight UNet refers to the baseline using the

TABLE 2. Segmentation results by ablation study of our methods on the Clinical dataset.

Model	Tumor						
	Dice per case	Dice global	VOE(%)	RVD(%)	ASSD(mm)	MSD(mm)	RMSD(mm)
AHCNet w/o HC	0.452	0.480	2.574	0.584	1.729	10.435	1.957
Weight UNet	0.502	0.539	2.102	0.471	1.515	8.842	1.567
AHCNet w/o DC	0.507	0.547	1.751	0.347	1.463	7.730	1.534
AHCNet 1-i/o	0.516	0.547	2.260	0.451	1.537	8.524	1.628
AHCNet w/o AG	0.533	0.562	1.747	0.335	1.484	7.971	1.547
AHCNet	**0.574**	**0.591**	**1.507**	**0.329**	**1.462**	**7.538**	**1.515**



(a) (b) (c)

FIGURE 8. Automatic liver and tumor segmentation using CDNN and AHCNet with 3DIRCADb dataset. The green regions denote correctly liver segmentation while the light blue ones for liver false positive pixels and the pink depicts true negative pixels; dark blue shows correctly predicted tumor segmentation while the red ones for tumor false positive pixels and light yellow depicts true negative pixels. (a) Test image. (b) CDNN. (c) AHCNet.

cascaded architecture, AHCNet 1-i / o is the one that inputs one slice in the network, and AHCNet is the one that inputs 5 consecutive slices in the network. From Table 1 and Table 2, we can see that our method achieved better performance than baseline on tumor segmentation accuracy, with 11.6% and 7.2% improvement on dice global, respectively. Statistically significant results are highlighted in bold font.

Table 3 shows the comparison of liver and tumor segmentation results on 3DIRCADb dataset with other recent segmentation works. All other experimental results are reported in paper [26], except experiment [11]. For a fair comparison, we ran experiments with methods of CDNN architecture, where the training setting keeps the same with Yuan [11]. And all experiment trained and evaluated model using the 15 volumes containing hepatic tumors in the liver with

TABLE 3. Comparison of liver and tumor segmentation results on 3DIRCADb dataset with other recent segmentation works (Dice: %).

Model	Liver	Tumor
Unet [54]	0.923 ± 0.03	0.51 ± 0.25
Christ *et al.* [10]	0.943	0.56 ± 0.26
Yuan *et al.* [11]	0.940 ± 0.03	0.56 ± 0.08
ResNet [21]	0.938 ± 0.02	0.60 ± 0.12
AHCNet	$\mathbf{0.945 \pm 0.02}$	$\mathbf{0.62 \pm 0.07}$

2-fold cross-validation. And our method still outperforms Unet [10] and ResNet [21], with 11% and 2% improvement on dice global for tumor segmentation respectively. Fig. 8 and Fig. 9 shows some examples of liver and tumor segmentation results achieved by CDNN and AHCNet on the 3DIRCADb dataset and Clinical dataset respectively. We can observe that AHCNet can achieve much better results than CDNN owing to AHCNet detected the complex and heterogeneous structure of the liver and all tumors in the qualitative results. Experimental comparisons verify the superiority of our proposed method compared to other methods.

C. FEATURE VISUALIZATION

In Fig. 10, we further visualize the set of corresponding feature maps produced by these 3D convolution kernels in the 41st convolutional layer, which are explicitly displayed in a slice-wise manner. We can see that the features extracted by AHCNet have less correlation with Weight UNet, which indicates that AHCNet has superior representative features [50].

Fig. 11 shows the attention coefficient of the AG module in the liver and tumor segmentation network, which well agree with the manually labeled bounding boxes. This shows that AG consistently focuses on the ROI of the target, which helps AHCNet learn the most important features of the target class, and only requires a very low computational cost. The attention map of the AG module is obtained by standardizing and averaging the maximum attention values of all AG modules.

To show that AHCNet can extract more distinguishable features, we visualize the convolution kernels of liver segmentation network by gradient ascend method [52]. As shown in Fig. 12, by comparing the maximum activation maps of the first layer convolution kernels between CDNN

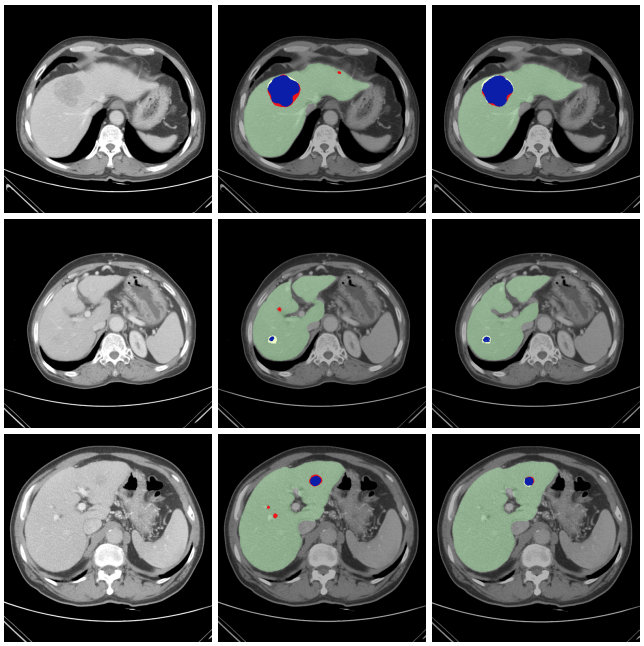


FIGURE 9. Automatic liver and tumor segmentation using CDNN and AHCNet with Clinical dataset. The transparent green (convenient display) depicts liver segmentation results, dark blue shows correctly predicted tumor segmentation while the red regions denote tumor false positive pixels, and pale yellow depicts true negative pixels. (a) Test image. (b) CDNN. (c) AHCNet.

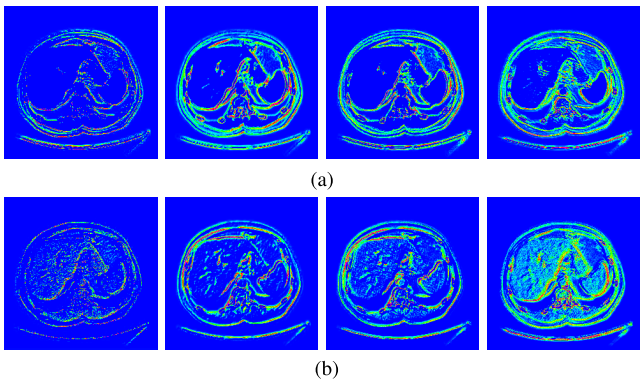


FIGURE 10. Visualization of typical features in the first layer of (a) UNet and (b) AHCNnet.

and AHCNet, we find that the maps obtained by AHCNet are more distinct and elementary than those obtained by CDNN (we performed pseudo-color processing on the obtained input image for the convenience of observation). And compared to CDNN, AHCNet will have fewer similar or blank maximum activation maps, indicating that the proposed method makes fewer redundant convolution kernels and more convolution kernels work effectively for subsequent operations [53].

D. DISCUSSION

Automatic liver and tumor segmentation can provide precise contours of the liver and any tumors inside the anatomical segments of the liver, and how to design an effective segmentation model is a fundamental and challenging problem in

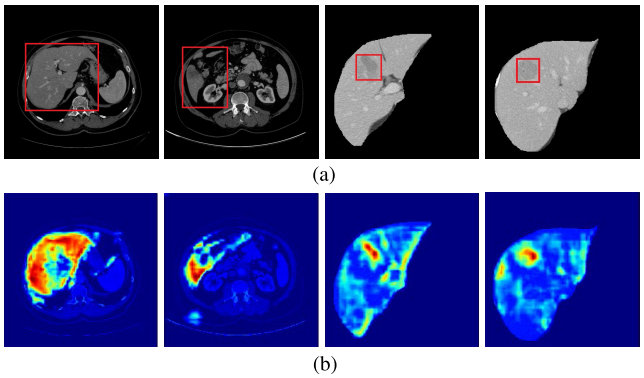


FIGURE 11. Examples of obtained attention map in the 41st layer from AHCNet. Red lines in (a) represent manually labeled bounding boxes. (a) The attention region of liver and tumor by manually labeled. (b) The attention map of liver and tumor by AG of AHCNnet.

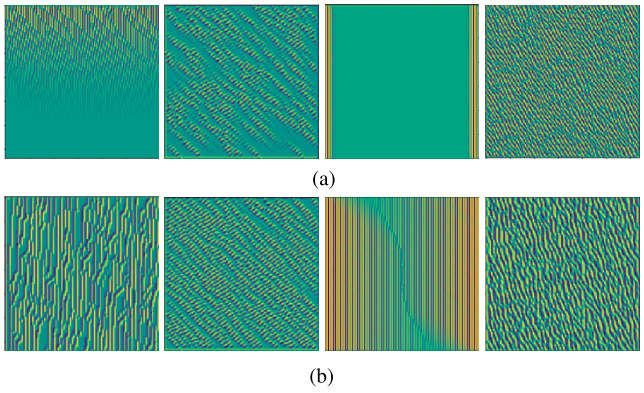


FIGURE 12. Comparing of the maximum activation input map of the first layer convolution kernels of (a) CDNN and (b) AHCNet.

medical image analysis. In recent years, deep learning techniques, especially deep convolution neural networks, have become competitive alternatives to complex medical image analysis tasks. Compared with traditional methods, these techniques utilize highly representative features learned from the labeled training datasets. This data-driven learning process can be easily generalized between different datasets and different imaging modes. In this regard, the development of a more general-purpose algorithm for medical image analysis tasks based on deep learning techniques is promising, and the successful application of the proposed cascaded AHCNet to liver and tumor segmentation tasks in volume data proves the striking features of these techniques. AHCNet is a 3D FCN model, which can fully and efficiently integrate contextual feature information. The network consists of multiple AHCBlocks with both long and short skip connections and self-soft attention models is a dense connection that enables precise medical image semantic segmentation.

One of the challenges of medical image segmentation is how to perform effective regional feature extraction on small targets. We propose a combination of hard and soft attention mechanism to gradually extract the characteristics of small-area targets. The hard attention mechanism completes liver localization, liver segmentation, and tumor

segmentation through the form of three-level convolutional neural network, and the fully differentiable AG module implements a soft attention mechanism. The AG module uses the 1×1 kernel implement global average pooling of feature tensors to obtain the attention of channel domain, activating the feature averages on a channel-by-channel basis to get the attention of the spatial domain. And the mixed attention of the two domains forms a high-efficiency information selection and betting mechanism, enabling the fine-grained feature learning. As far as we can tell, this work combines soft attention mechanisms with CNN for the first time and applies to tumor segmentation in CT volumes, embodying the importance of the attention mechanism in medical image processing, especially in small organs or tumor segmentation. In addition, we propose a joint dice loss function in the liver location network to eliminate the need for class balance, balance the various indicators and improve the accuracy of the liver bounding box; and use focal loss fine-tuning in the tumor segmentation network to eliminate false positive tumor as much as possible. These two novel loss functions are trained for different tasks of the cascade architecture respectively, reflecting the importance of the choice of the loss function.

Another challenge with medical image segmentation is that it often causes optimization problems when the network is deepened, although we hope to use a relatively deep network to capture more representative features for more accurate segmentation. We propose hybrid connection to reduce the training difficulty of low- and medium-level features through a tighter structure, thus stimulating the network to learn more representative features from the information flow. This compact structure not only effectively integrate contextual feature information, but also further alleviate the problem of limited training data by mining the potential of data, resulting in stronger and more general features [40].

Finally, we tested our trained model from the LiTS dataset on the Clinical dataset, and the dice score for tumor segmentation is 59.1%, indicating the generalization capability of our method in clinical practice. The segmentation evaluation scores we obtained on the Clinical dataset are lower than the 3DIRCADb dataset, and we attribute the reasons to the following two points. First, the image quality of the 3DIRCADb dataset is similar to the LiTS dataset, while the Clinical dataset is quite different from the LiTS dataset. Second, cases in public datasets usually have larger size of tumors (larger than 10mm in diameter) and are more inclined to evaluate the advantages and disadvantages of segmentation algorithms, while there are quite a number of small-scale tumors (about 5mm in diameter) in Clinical dataset, and more inclined to the verification of clinical diagnosis. The promising results achieved on the Clinical dataset also validated that our method is effective to generalize to different dataset under different data collection conditions.

V. CONCLUSION

In this paper, we proposed that a cascaded deployment of AHCNet can produce competitive results for liver tumor segmentation on a clinical CT dataset while being efficiently deployed on a single GPU. The architecture combines the soft and hard attention mechanism and the short and long skip connections to achieve efficient feature extraction and fusion, and the network architecture is essentially general, can be easily extended to other applications. Extensive experiments on the dataset of 3DIRCADb and Clinical dataset demonstrated the superiority of our proposed AHCNet. However, the accuracy of tumor segmentation is still low and needs further improvement.

In future work, we also plan to improve the accuracy of lesion segmentation by difference silhouette. Post-processing methods such as level set [55] and CRF [10] can also potentially improve our model segmentation performance. Finally, graph convolution [56] is a hot research direction in recent years, and how to contact self-attention mechanism and graph convolution, as well as a deeper understanding of self-attention mechanism, are important directions for the future.

REFERENCES

[1] J. Ferlay, H.-R. Shin, F. Bray, D. Forman, C. Mathers, and D. M. Parkin, "Estimates of worldwide burden of cancer in 2008: Globocan 2008," *Int. J. Cancer*, vol. 127, no. 12, pp. 2893–2917, 2010.
[2] P. Campadelli, E. Casiraghi, and A. Esposito, "Liver segmentation from computed tomography scans: A survey and a new algorithm," *Artif. Intell. Med.*, vol. 45, nos. 2–3, pp. 185–196, 2009.
[3] G. Li, X. Chen, F. Shi, W. Zhu, J. Tian, and D. Xiang, "Automatic liver segmentation based on shape constraints and deformable graph cut in CT images," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5315–5329, Dec. 2015.
[4] C. Li *et al.*, "A likelihood and local constraint level set model for liver tumor segmentation from CT volumes," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 10, pp. 2967–2977, Oct. 2013.
[5] R. Vivanti, A. Ephrat, L. Joskowicz, N. Lev-Cohain, O. A. Karaaslan, and J. Sosna, "Automatic liver tumor segmentation in follow-up CT scans: Preliminary method and results," in *Proc. Int. Workshop Patch-Based Techn. Med. Imag.* Cham, Switzerland: Springer, 2015, pp. 54–61.
[6] A. Ben-Cohen, E. Klang, I. Diamant, N. Rozendorn, M. M. Amitai, and H. Greenspan, "Automated method for detection and segmentation of liver metastatic lesions in follow-up CT examinations," *J. Med. Imag.*, vol. 2, no. 3, 2015, Art. no. 034502.
[7] Y. Häme and M. Pollari, "Semi-automatic liver tumor segmentation with hidden Markov measure field model and non-parametric distribution estimation," *Med. Image Anal.*, vol. 16, no. 1, pp. 140–149, 2012.
[8] H. R. Roth *et al.*, "An application of cascaded 3D fully convolutional networks for medical image segmentation," *Comput. Med. Imag. Graph*, vol. 66, p. 90, Jun. 2018.
[9] Y. Zhu, C. Zhao, H. Guo, J. Wang, X. Zhao, and H. Lu, "Attention CoupleNet: Fully convolutional attention coupling network for object detection," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 113–126, Jan. 2018.
[10] P. F. Christ *et al.*, "Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent.*, 2016, pp. 415–423.
[11] Y. Yuan. (2017). "Hierarchical convolutional-deconvolutional neural networks for automatic liver and tumor segmentation." [Online]. Available: https://arxiv.org/abs/1710.04540
[12] M. Bellver *et al.* (2017). "Detection-aided liver lesion segmentation using deep learning." [Online]. Available: https://arxiv.org/abs/1711.11069
[13] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified visual attention networks for fine-grained object classification," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1245–1256, Jun. 2017.
[14] K. Xu *et al.* (2015). "Show, attend and tell: Neural image caption generation with visual attention." [Online]. Available: https://arxiv.org/abs/1502.03044

[15] Z. Pian, T. Shi, P. Yuan, L. Hu, and D. Wang, "Application of hierarchical visual perception in target recognition," *J. Comput.-Aided Des. Comput. Graph.*, vol. 29, no. 6, pp. 1093–1102, 2017. [Online]. Available: http://en.cnki.com.cn/Article_en/CJFDTotal-JSJF201706015.htm

[16] J. Schlemper *et al.* (2018). "Attention-gated networks for improving ultrasound scan plane detection." [Online]. Available: https://arxiv.org/abs/1804.05338

[17] O. Oktay *et al.* (2018). "Attention U-Net: Learning where to look for the pancreas." [Online]. Available: https://arxiv.org/abs/1804.03999

[18] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep Learning and Data Labeling for Medical Applications*. Cham, Switzerland: Springer, 2016, pp. 179–187.

[19] K. He, X. Zhang, S. Ren, and J. Sun. (2015). "Deep residual learning for image recognition." [Online]. Available: https://arxiv.org/abs/1512.03385

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[21] X. Han. (2017). "Automatic liver lesion segmentation using a deep convolutional neural network method." [Online]. Available: https://arxiv.org/abs/1704.07239

[22] E. Vorontsov, A. Tang, C. Pal, and S. Kadoury, "Liver lesion segmentation informed by joint liver segmentation," in *Proc. IEEE 15th Int. Symp. Biomed. Imag.*, Apr. 2018, pp. 1332–1335.

[23] A. Ben-Cohen, I. Diamant, E. Klang, M. Amitai, and H. Greenspan, "Fully convolutional network for liver segmentation and lesions detection," in *Proc. Deep Learn. Data Labeling Med. Appl.* Cham, Switzerland: Springer, 2016, pp. 77–85.

[24] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng, "3D deeply supervised network for automatic liver segmentation from CT volumes," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2016, pp. 149–157.

[25] F. Lu, F. Wu, P. Hu, Z. Peng, and D. Kong, "Automatic 3D liver location and segmentation via convolutional neural network and graph cut," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 2, pp. 171–182, 2017.

[26] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2017.

[27] A. Sato, M. Mori, P. D. Funkenbusch, and T. Mori, "Microscopic observation of elastic distortions caused by orowan loops," *Acta Metallurgica*, vol. 34, no. 9, pp. 1751–1758, 1986.

[28] A. Arnab, O. Miksik, and P. H. S. Torr. (2017). "On the robustness of semantic segmentation models to adversarial attacks." [Online]. Available: https://arxiv.org/abs/1711.09856

[29] O. Oktay *et al.* (2018). "Attention U-net: Learning where to look for the pancreas." [Online]. Available: https://arxiv.org/abs/1804.03999

[30] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng. (2017). "Dual path networks." [Online]. Available: https://arxiv.org/abs/1707.01629

[31] J. Schlemper *et al.* (2018). "Attention-gated networks for improving ultrasound scan plane detection." [Online]. Available: https://arxiv.org/abs/1804.05338

[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[33] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. (2017). "Squeeze-and-excitation networks." [Online]. Available: https://arxiv.org/abs/1709.01507

[34] M. Jaderberg, K. Simonyan, K. Kavukcuoglu, and A. Zisserman, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[35] F. Wang *et al.* (2017). "Residual attention network for image classification." [Online]. Available: https://arxiv.org/abs/1704.06904

[36] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. (2017). "Rethinking atrous convolution for semantic image segmentation." [Online]. Available: https://arxiv.org/abs/1706.05587

[37] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6230–6239.

[38] F. Yu, D. Wang, E. Shelhamer, and T. Darrell. (2017). "Deep layer aggregation." [Online]. Available: https://arxiv.org/abs/1707.06484

[39] Y. Yang, Z. Zhong, T. Shen, and Z. Lin. (2018). "Convolutional neural networks with alternately updated clique." [Online]. Available: https://arxiv.org/abs/1707.06484

[40] J. B. Hopfinger, M. H. Buonocore, and G. R. Mangun, "The neural mechanisms of top-down attentional control," *Nature Neurosci.*, vol. 3, no. 3, pp. 284–291, 2000.

[41] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Netw.*, vol. 106, pp. 249–259, Oct. 2017.

[42] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," *BMC Med. Imag.*, vol. 15, no. 1, p. 29, 2015.

[43] C. Xu *et al.*, "Multi-loss regularized deep neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 12, pp. 2273–2283, Dec. 2016.

[44] Z. Jianqiang, G. Xiaolin, and Z. Xuejun, "Deep convolution neural networks for twitter sentiment analysis," *IEEE Access*, vol. 6, pp. 23253–23260, 2018.

[45] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[46] X.-Y. Zhou, M. Shen, C. Riga, G.-Z. Yang, and S.-L. Lee. (2017). "Focal FCN: Towards biomedical small object segmentation with limited training data." [Online]. Available: https://arxiv.org/abs/1711.01506

[47] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. (2017). "On calibration of modern neural networks." [Online]. Available: https://arxiv.org/abs/1706.04599

[48] F. Chollet. (2015). *Keras*. [Online]. Available: https://github.com/fcholle/keras

[49] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[50] M. Danciu, M. Gordan, C. Florea, and A. Vlaicu, "3D DCT supervised segmentation applied on liver volumes," in *Proc. 35th Int. Conf. Telecommun. Signal Process.*, 2012, pp. 779–783.

[51] L. Soler *et al.* (2012). *3D Image Reconstruction for Comparison of Algorithm Database: A Patient-Specific Anatomical and Medical Image Database*. [Online]. Available: http://www-sop.inria.fr/geometrica/events/wam/abstract-ircad.pdf

[52] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2013, pp. 818–833.

[53] G. Papandreou. (2014). "Deep epitomic convolutional neural networks." [Online]. Available: https://arxiv.org/abs/1406.2732

[54] G. Chlebus, H. Meine, J. H. Moltz, and A. Schenk. (2017). "Neural network-based automatic liver tumor segmentation with random forest-based candidate filtering." [Online]. Available: https://arxiv.org/abs/1706.00842

[55] M. Pinheiro and J. L. Alves, "A new level-set-based protocol for accurate bone segmentation from CT imaging," *IEEE Access*, vol. 3, pp. 1894–1906, 2015.

[56] Z. Zhang, P. Cui, and W. Zhu. (2018). "Deep learning on graphs: A survey." [Online]. Available: https://arxiv.org/abs/1812.04202

Authors' photographs and biographies not available at the time of publication.