IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 23, 2019, accepted June 3, 2019, date of publication June 6, 2019, date of current version June 24, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2921434

HSN: Hybrid Segmentation Network for Small Cell Lung Cancer Segmentation

WEI CHEN[1], HAIFENG WEI[2], SUTING PENG[1], JIAWEI SUN[1], XU QIAO[1], AND BOQIANG LIU[1]
[1]Department of Biomedical Engineering, Shandong University, Jinan 250061, China
[2]First Clinical Medical College, Shandong University of Traditional Chinese Medicine, Jinan 250012, China

Corresponding authors: Xu Qiao (qiaoxu@sdu.edu.cn) and Boqiang Liu (bqliu@sdu.edu.cn)

This work was supported in part by the Department of Science and Technology of Shandong Province under Grant 2017CXGC1502, in part by the Natural Science Foundation of Shandong Province under Grant ZR2014HQ054, and in part by the National Natural Science Foundation of China under Grant 61603218 and Grant U1806202.

ABSTRACT Small cell lung cancer (SCLC) is one of the most common types of malignant tumors, characterized by rapid growth and early metastasis spread. Early and accurate diagnosis of SCLC is vital for improved survival. Accurate cancer segmentation helps doctors understand the location and size of cancer and make better diagnostic decisions. However, manual segmentation of lung cancers from large amounts of medical images is a time-consuming and challenging task. In this paper, we propose a hybrid segmentation network (referred to as HSN) based on convolutional neural network (CNN) to automatically segment SCLC from computed tomography (CT) images. The design philosophy of our model is to combine a lightweight 3D CNN to learn long-range 3D contextual information and a 2D CNN to learn fine-grained semantic information, which is essential for accurate cancer segmentation. We propose a hybrid features fusion module to effectively fuse the 2D and 3D features and to jointly train these two CNNs. We utilize a generalized Dice loss function to tackle the severe class imbalance problem in data. A dataset consists of 134 CT scans was constructed to evaluate our model. Our model achieved high performances with a mean Dice score of 0.888, a mean sensitivity score of 0.872 and a mean precision of 0.909, outperforming the other state-of-the-art 2D and 3D CNN methods by a large margin.

INDEX TERMS Small cell lung cancer, CT, deep convolutional neural network, hybrid features fusion.

I. INTRODUCTION

Lung cancer is one of the most common types of malignant tumors in the world. It is the second most common cancer among both men and women in the United States [1], the first most common cancer among men and the second most common cancer among women in China [2]. Moreover, lung cancer is the leading cause of cancer death among both men and women in both countries. Small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) are the two main types of lung cancers. SCLC accounts for 15–20% of all lung cancer cases. Compared with NSCLC, SCLC is characterized by high malignancy, rapid progress, early metastasis spread, and poor prognosis, and has a severe impact on the physical and mental health of patients [3]. SCLC can be staged into two categories: limited stage and extensive stage. Being at a specific stage indicates how much cancer has spread through the body. In the limited stage, cancer is limited to one side of the chest, while in the extensive stage, cancer has spread throughout the lung to the lymph nodes or has metastasized to other parts of the body [4]. Unfortunately, about two in three patients with SCLC are in extensive stage upon the first diagnosis and require systemic chemotherapy [5].

Computed tomography (CT) is the primary imaging modality used to evaluate the tumor and determine the stage of the disease [6]. Contrast-enhanced CT can provide cancer images with high resolution and clear boundaries, which is useful to reveal cancer characteristics. Since SCLC is an aggressive malignancy characterized by rapid growth and early metastatic spread, fast localization and delineation of cancer regions in CT scans are very important for diagnosis and treatment planning. Accurate segmentation of NCLS can also assist clinicians in predicting the prognostic in SCLC patients. In our recent study [7], it is shown that the radiomics features extracted from cancer regions can help to predict the clinical response of SCLC patients to

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

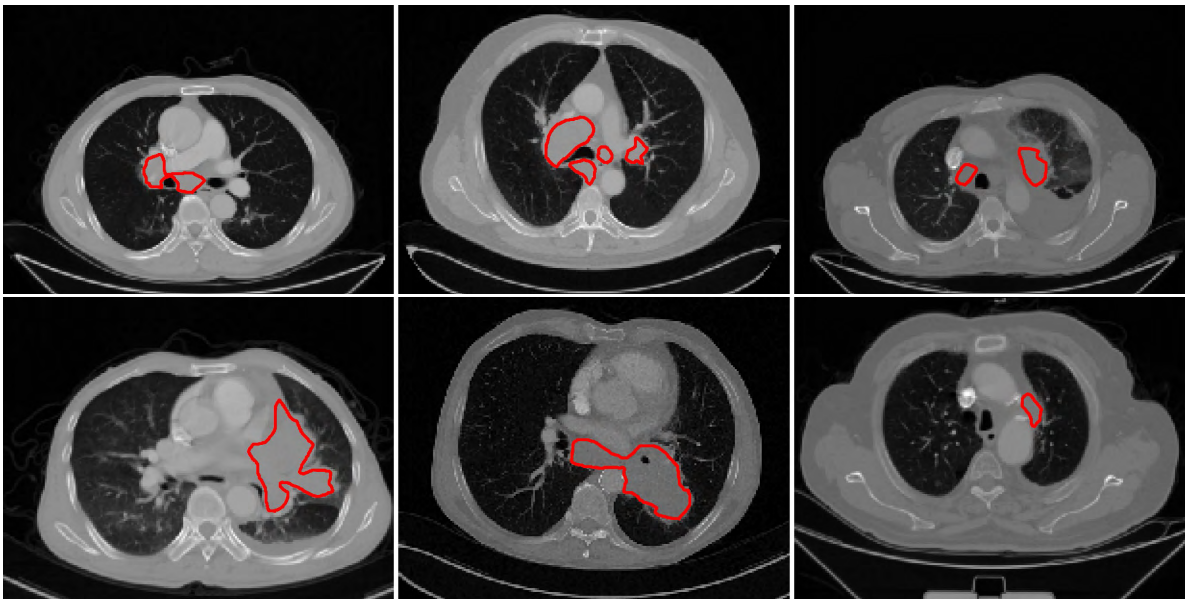

FIGURE 1. Illustration of the challenges in SCLC segmentation.The red contours correspond to cancers produced by experienced radiologists.

first-line chemotherapy. However, manual segmentation is a time-consuming and tedious task and is subject to significant inter- and intra-observer variations, which limit its value in the clinical settings. Therefore, automatic segmentation methods are highly demanded.

Approximately 90%–95% of SCLCs derive from lobar or main bronchi and are characterized by mediastinal or hilar lymphadenopathy. In a minority of cases, SCLCs appear as peripheral nodules or relatively small bronchial tumor [4]. However, the CT features of SCLC have not been fully investigated [8]. The complexity of CT characteristics makes automatic segmentation of SCLC a challenging task. We show some examples of labeled SCLC CT scans to demonstrate the challenges in Fig. 1. We can see from this figure that cancers may appear with different shapes and sizes.

Some methods have been proposed for the automatic segmentation of lung lesions, such as pulmonary nodule [9], [10] and NSCLC [11]. However, to the best of our knowledge, no previous study has dedicated to the automatic segmentation of SCLC. Existing methods for medical image segmentation can be divided into two major categories: (1) methods based on hand-crafted features and (2) methods based on data-driven deep features [12]. The first type of approach is usually related to machine learning, where a discriminative model is trained using hand-crafted features extracted from the raw data. After extracting different global and local features, a classifier is trained to determine which class each voxel belongs to. On the contrary, the methods based on data-driven deep features can learn more robust features specific to the task at hand, resulting in better segmentation performance. In this paper, we focus on the convolutional neural network (CNN), a data-driven deep features based method that has been widely used in the field of medical image analysis [13].

Although initially developed for image classification, CNNs can be used for semantic segmentation after some modifications. The straightforward approach is to train the model based on image patches and classify the labels of the center pixels [14]. An obvious limitation of this approach is that the feature representation is restricted to each patch, ignoring the global features that are very important for segmentation. The FCN proposed by Long et al. [15] can take arbitrary size image as input and produce the output with the same size as input. The approach solves the limitation of the method based on central pixel classification by replacing fully connected layers with 1×1 convolutions and using upsampling layers to restore the original size. Later, U-Net [16] modified FCN by adding multiple upsampling layers and concatenating multi-scale features for medical images segmentation.

In the field of medical image analysis, especially for MRI and CT images, a significant difference compared to natural images is the inherent volumetric nature of the data. A slice-by-slice learning strategy is inefficient and cannot capture inter-slice correlations. One direct way to learn volumetric information representations in medical images is to extend the convolution kernels from 2D to 3D, such as 3D U-Net [17] and V-Net [18]. In this way, the networks can take full advantage of the 3D context for better performance. Despite giving encouraging performances, 3D CNN has more parameters than 2D CNN, and the training of 3D CNN is computationally expensive, which limits the construction of very deep networks. Some works use downsampled images or 3D patches to train 3D models, but this can lead to loss of image information and zero-derivative problem [19]. In addition, volumetric medical images are usually anisotropic [20]. Take our data as an example; the voxel scale in depth (the z-axis, 5mm) is much larger than that in the xy plane (0.58–0.98mm).

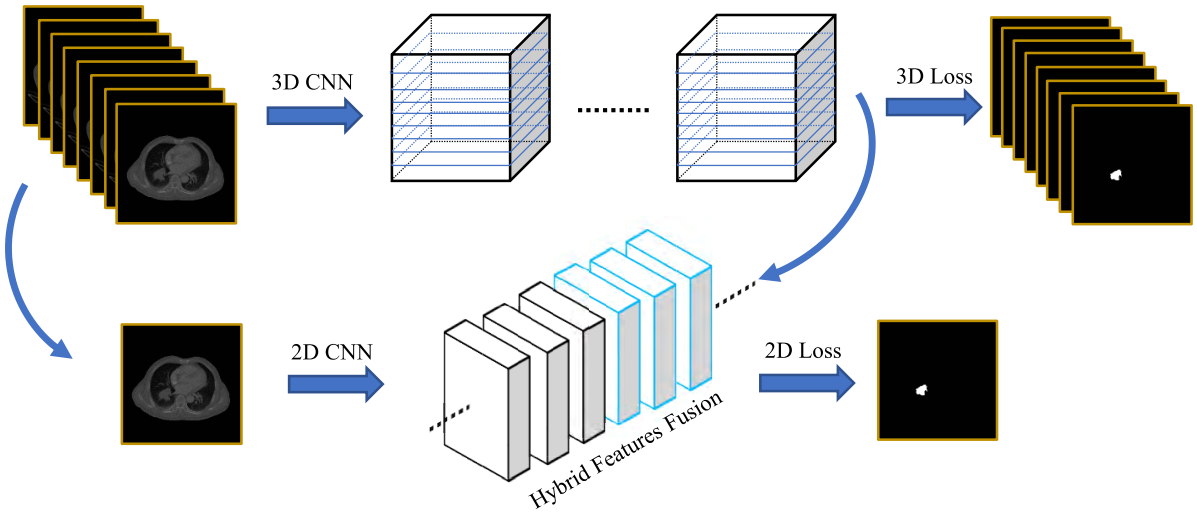

FIGURE 2. This figure shows the schematic structure of our proposed HSN.

Directly performing 3D convolutions with isotropic kernels on these anisotropic volumetric images could be problematic. A simple approach to address this problem is to re-sample all images to isotropic resolution, but this will result in much larger images, further increasing the computational cost and memory demand.

To solve the above problems, we propose a novel method called hybrid segmentation network (HSN) for SCLC segmentation from 3D CT images that combines the advantages of both 2D and 3D CNN. The design philosophy of HSN is clear. For 3D CNN, we build a lightweight network similar to 3D U-Net but use downsampled images and separable 3D convolutions to reduce the memory requirement and the computational cost. For 2D CNN, we use dilated convolutions to learn fine-grained semantic information while at the same time maintaining high spatial resolution. We then propose a hybrid features fusion module (HFFM) to fuse the 2D and 3D features effectively. In this way, the 2D CNN can leverage the 3D context extracted from 3D CNN. In addition, we utilize both 2D loss and 3D loss to optimize the network jointly.

In summary, the main contributions of this work are summarized as follows:

- We propose a hybrid segmentation network (HSN), which consists of a lightweight 3D CNN to learn long-range 3D contextual information and a 2D CNN to capture fine-grained semantic information.
- We propose a multiscale separable convolution (MSC) block to capture multiscale 3D context from anisotropic dimensional CT images.
- We propose a hybrid features fusion module (HFFM) to effectively fuse the 3D and 2D features and jointly train the hybrid network.
- We apply the proposed model in a CT dataset containing 134 SCLC patients. Results show that the proposed model achieves high performance in this challenging dataset.

The remainder of this paper is organized as follows. In section II, we describe the technical details of our proposed HSN model. In section III, we present the experimental results and discussions. This paper is finally concluded in Section IV.

II. METHODS

The proposed segmentation framework is shown in Fig. 2. Our model is an end-to-end trainable neural network that combines a lightweight 3D CNN to learn long-range 3D contextual information and a 2D CNN to capture fine-grained intra-slice semantic information. We employ spatiotemporal-separable 3D (S3D) convolutions to deal with the anisotropic dimensions of CT volumes and reduce the computational cost of the 3D CNN. In order to enlarge the receptive field while preserving high resolution to retain a large amount of semantic information about smaller objects, we employ dilated convolutions in 2D CNN. We design a hybrid features fusion module (HFFM) to fuse the 2D and 3D features effectively. We also utilize generalized Dice loss function to tackle the problem of data imbalance during training. In this section, we first introduce the key components of our network. Then we describe our model in detail.

- Spatiotemporal-separable 3D (S3D) convolution factorizes a standard 3D convolution into two consecutive convolutional layers: one 2D convolution to learn spatial features and one 1D convolution to learn temporal features.
- Multiscale separable convolution (MSC) Block is an Inception-ResNet-like architecture with S3D convolutions to effectively capture multiscale 3D contextual information.
- Dilated convolution, also known as ‘‘atrous convolution’’, can be used to enlarge the receptive field while preserving the resolution of feature maps.

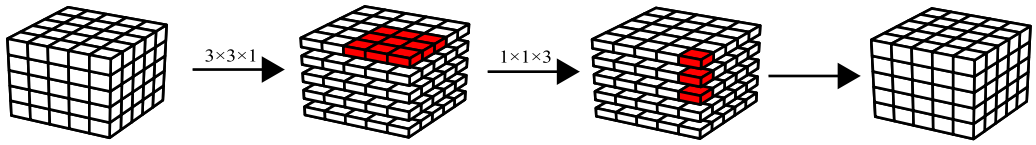

FIGURE 3. The S3D convolution. The kernel of size 3 × 3 × 1 does a 2D convolution and the kernel of size 1 × 1 × 3 does a 1D convolution. Combining them together we have the S3D convolution.

- Hybrid features fusion Module (HFFM) is designed to effectively fuse 3D and 2D features to allow the network to jointly train the 3D CNN and 2D CNN.

A. SPATIOTEMPORAL-SEPARABLE 3D CONVOLUTION

In order to build a lightweight 3D CNN model under anisotropic images, we employ spatiotemporal-separable 3D (S3D) convolution [21]. S3D convolution, as shown in Fig. 3, also called pseudo-3D convolution [22], has been widely used in 3D video tasks. Briefly, it splits one $k \times k \times k$ convolution into a $k \times k \times 1$ convolution and a $1 \times 1 \times k$ convolution. Given a standard 3D convolutional layer that takes a feature map F of size $D_F \times D_F \times D_F \times M$ as input, where D_F denotes the spatial width, height, and depth of a cubic input feature map, M is the number of input channels. Here we assume the feature map has the same spatial dimension. Through the convolution operation, a feature map $\mathbf{G}$ of size $D_G \times D_G \times D_G \times N$ is produced, where D_G denotes the spatial width, height and depth of a cubic output feature map, N is the number of output channels. The convolution operation is implemented by a kernel $\mathbf{K}$ of size $D_K \times D_K \times D_K \times M \times N$ where D_K is the spatial dimension of the kernel assumed to be cubic. The output feature map of standard 3D convolution with stride and padding can be computed as:

$$\mathbf{G}_{x,y,z,n} = \sum_{i,j,k,m} \mathbf{K}_{i,j,k,m,n} \mathbf{F}_{x+i-1,y+j-1,z+k-1,m} \tag{1}$$

In the context of S3D convolution, the full 3D convolution can be replaced by two consecutive convolutional layers: one 2D convolution to learn spatial features and one 1D convolution to learn temporal features. The first stage of S3D convolution can be computed as:

$$\hat{\mathbf{G}}_{x,y,z,m} = \sum_{i,j,m} \hat{\mathbf{K}}_{i,j,m,m} \mathbf{F}_{x+i-1,y+j-1,z,m} \tag{2}$$

here we assume that the input and output have the same number of M channels. The second stage of S3D convolution is a 1D convolution which can be computed as:

$$\hat{\mathbf{G}}'_{x,y,z,n} = \sum_{k,m} \hat{\mathbf{K}}_{k,m,n} \mathbf{F}_{x,y,z+k-1,m} \tag{3}$$

The total computational cost of the S3D convolution is:

$$\underbrace{D_K \cdot D_K \cdot M \cdot M \cdot D_F \cdot D_F \cdot D_F}_{\text{the cost of first stage}} + \underbrace{D_K \cdot M \cdot N \cdot D_F \cdot D_F \cdot D_F}_{\text{the cost of second stage}} \tag{4}$$

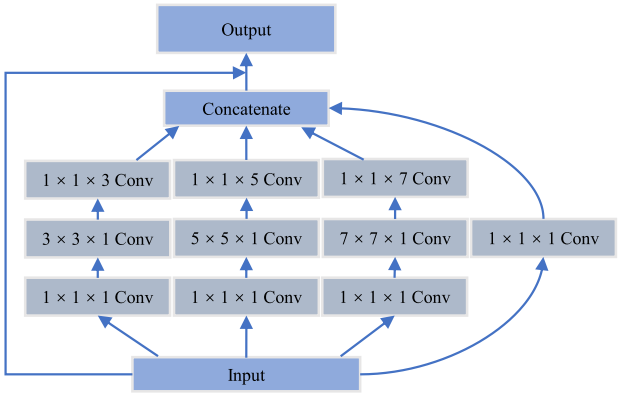

FIGURE 4. The figure illustrates the multiscale separable convolution (MSC) block. It contains four cascaded branches with different sizes of filters (1 × 1 × 1, 3 × 3 × 3, 5 × 5 × 5, and 7 × 7 × 7) and a residual connection.

By replacing standard 3D convolution with S3D convolution we get a reduction in computational cost of:

$$\frac{D_K \cdot D_K \cdot M \cdot M \cdot D_F \cdot D_F \cdot D_F + D_K \cdot M \cdot N \cdot D_F \cdot D_F \cdot D_F}{D_K \cdot D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F \cdot D_F} = \frac{M}{D_K \cdot N} + \frac{1}{D_K^2} \tag{5}$$

From the above, it can be seen that we can get a significant reduction in computational cost by using S3D convolutions. The S3D convolution is initially proposed for video understanding tasks, where the $k \times k \times 1$ filter is performed on the spatial domain and the $1 \times 1 \times k$ filter is performed on adjacent feature maps in time. In the context of volumetric medical images, we employ S3D convolutions to separately learn inter- and intra-slice features, which is beneficial to address the problem of anisotropic dimensions.

B. MSC BLOCK

Inception [23] and ResNet [24] are two powerful architectures widely used in deep learning. Inception structure adopts convolutional layers with different kernel sizes in a parallel way to efficiently recognize details at different extents. ResNet employs identity mapping to accelerate the speed of the training process and reduce the effect of vanishing gradient problem. Inception-ResNet [25], a hybrid architecture that combines the Inception and ResNet, has become a baseline component in many state-of-the-art CNNs. Motivated by the Inception-ResNet architecture and the S3D convolution, we proposed multiscale separable convolution (MSC) block to extract 3D contextual information from anisotropic images. As shown in Fig. 4, MSC block contains four

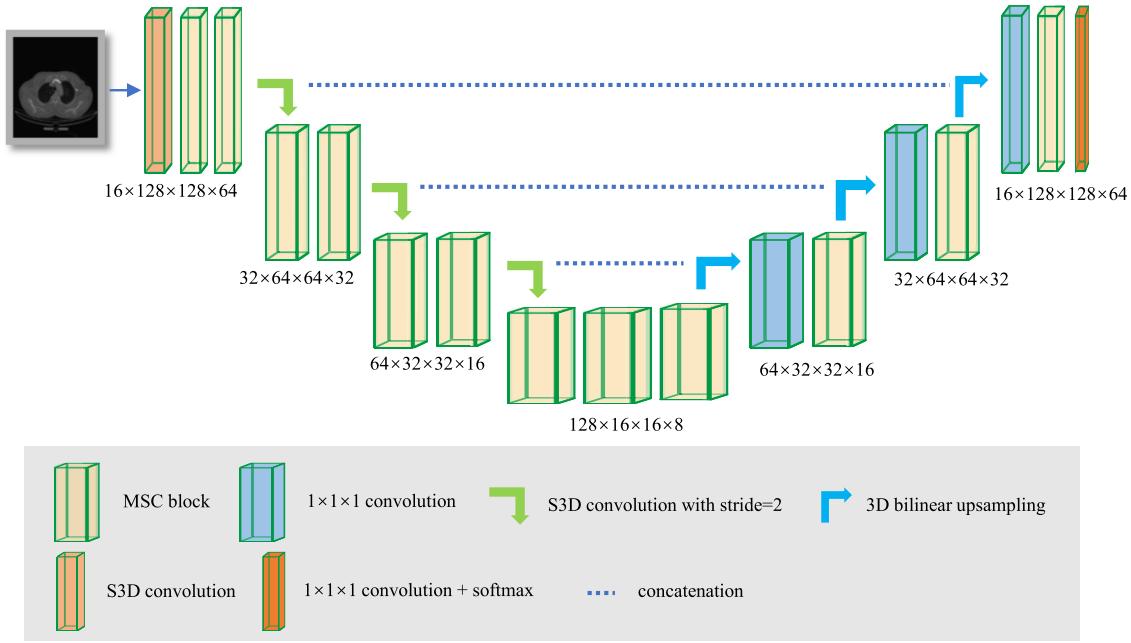

FIGURE 5. Schematic visualization of the 3D CNN part of our HSN model.

cascaded branches with different sizes of filters ($1 \times 1 \times 1$, $3 \times 3 \times 3$, $5 \times 5 \times 5$ and $7 \times 7 \times 7$) and a residual connection. To reduce the computational cost, we add an extra $1 \times 1 \times 1$ convolution before each S3D convolution to limit the number of feature maps.

C. 3D CNN

The architecture of our proposed 3D CNN is shown in Fig. 5. It is an encoder-decoder structure similar to 3D U-Net [17], but we carefully make some modifications to make it suitable for the lung tumor segmentation task. Our 3D CNN follows the standard U-Net architecture with an encoder to progressively extract image features and a decoder to generate the segmentation maps. In the encoder part, we use MSC blocks instead of simply 3D convolutions to handle the highly anisotropic dimensions and reduce the computational cost. In more detail, in the original 3D U-Net, each level contains two $3 \times 3 \times 3$ convolutions. In our proposed method, we replace it with two MSC blocks. To reduce the size of the feature maps, we use striding S3D convolutions. The goal of the decoder is to generate high-resolution feature maps with semantic information from the encoded features. To achieve it, we first upsample the low-resolution feature maps using 3D bilinear upsampling. We then concatenate the upsampled features with the features from the corresponding level of the encoder. Following the concatenation, an MSC block is used to adjust the number of feature maps. Compared with the original 3D U-Net, the proposed lightweight 3D CNN has fewer parameters and computational cost, which is essential for our hybrid segmentation network.

D. 2D CNN

In a typical CNN, the use of consecutive pooling operations or striding convolutions significantly reduces the feature resolution in order to learn global information. The highly abstract features with low resolution have great advantages for image classification task but will impede the semantic segmentation task where an output with full resolution labeled image is essential. This situation is likely to be an even bigger problem in the context of medical image segmentation. Therefore, we must address the following question: how to learn fine-grained semantic information for identifying small cancer regions. In order to answer this question, we adopt dilated convolutions in the design of our 2D CNN. Compared with standard convolution, dilated convolution introduces a parameter called dilation rate to insert zeros between the values in a kernel. Mathematically, considering a feature map $y[i]$ of dilated convolution of the input $x[i]$ with filter $w[k]$, the dilated convolution is computed as follows:

$$\mathbf{y}[i] = \sum_{k} x[i + rk]\, w[k] \tag{6}$$

where r corresponds to the dilation rate, which indicates the stride with which we sample the input signal. When using dilated convolution, it is equivalent to upsampling the filter by inserting $r - 1$ zeros between two adjacent filter values for each spatial dimension and then performing the convolution using the upsampled filter. When rate $= 1$, dilated convolution degenerates to standard convolution. Dilated convolution allows us to achieve a large receptive field while at the same time maintaining high spatial resolution at the output to capture fine-grained features for identifying small objects and blurred boundaries. We also propose a dilated unit block (DUB), which consists of two sequential dilated convolutions with the residual connection. Fig. 6 shows the architecture of the proposed 2D CNN. For an input image of 512×512, we first use a 3×3 convolution to generate 16 feature maps.

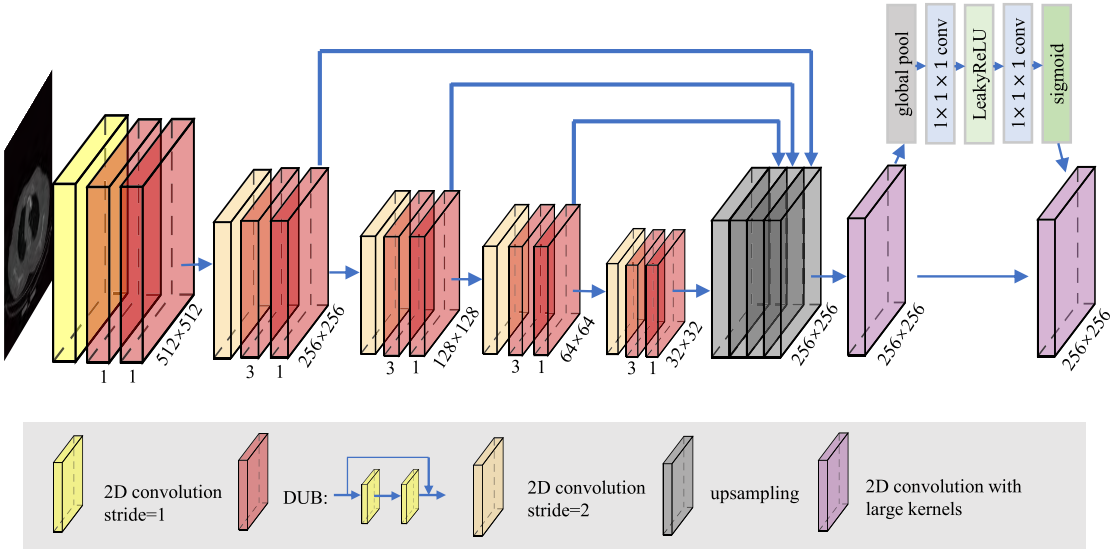

FIGURE 6. Schematic visualization of the 2D CNN part of our HSN model.

Then we alternately use DUBs and striding S3D convolutions to extract features until the feature responses are 16 times smaller than the input dimension. To combine features at different scales, we upsample the feature maps at different scales to the resolution of 256 × 256 and then concatenate them. After concatenation, a Squeeze-and-Excitation (SE) block is used to effectively select and combine features. First proposed in SENet [26],the SE block can explicitly model the interdependencies between the channels of feature maps, which can be used to recalibrate features. Unlike the original SE block, we replace the 3 × 3 kernels with large kernels of size 7×7 to further increase the valid receptive field for better performance [27].

E. HYBRID FEATURES FUSION MODULE

The feature maps produced by the two CNNs are different in size and level of feature representation. Therefore, we cannot simply sum or concatenate these features. The feature maps produced by 3D CNN are volumetric, whereas the feature maps produced by 2D CNN are two-dimensional. Moreover, the 3D features mainly encode 3D contextual information, and the 2D features encode fine-grained semantic information of 2D slices. In other words, the features of 3D CNN and the features of 2D CNN are at a different level of feature representation. Therefore, we propose a hybrid features fusion module (HFFM) to fuse these features effectively, as shown in Fig. 7. We first upsample the feature maps before the last 1 × 1 × 1 convolutional layer in the 3D CNN to the size of 256 × 256 × 64. Let $\mathbf{X}_{3d} \in R^{m \times c_1 \times 256 \times 256 \times 64}$ be the feature maps, where m denotes the batch size and c_1 denotes the channels. Let $\mathbf{X}_{2d} \in R^{n \times c_2 \times 256 \times 256}$ be the feature maps from 2D CNN, where n denotes the batch size and c_2 denotes the channels. For simplicity, we set the batch size of 3D CNN equal to 1, that is, $\mathbf{X}_{3d} \in R^{c_1 \times 256 \times 256 \times 64}$. For 2D

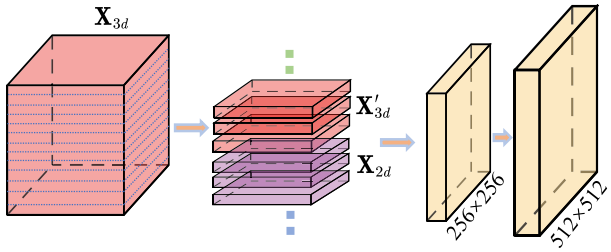

FIGURE 7. Hybrid features fusion module (HFFM).

CNN, we sample a stack of adjacent slices from only one CT volume along the z-axis. To fuse the 3D and 2D features, we first crop a stack of adjacent slices from $\mathbf{X}_{3d}$ with the same slices indices as we sampled for 2D CNN, then permute the dimensions to have the form $\mathbf{X}'_{3d} \in R^{n \times c_1 \times 256 \times 256}$. Then $\mathbf{X}'_{3d}$ and $\mathbf{X}_{2d}$ can be concatenated,

$$\mathbf{H} = Cat\left(\mathbf{X}'_{3d}, \mathbf{X}_{2d}\right), \quad \mathbf{H} \in R^{n \times (c_1 + c_2) \times 256 \times 256} \tag{7}$$

By doing so, the hybrid features $\mathbf{H}$ can be optimized in the context of 2D CNN, while at the same time the 3D information is integrated for accurate lung tumor segmentation. It is worth noting that this fusion strategy can be easily extended to large batch size. After the fusion, we add a convolution to refine $\mathbf{H}$, followed by an upsampling layer to increase the size of feature maps to the original input size. Then a 1 × 1 × 1 convolution is used to generate two channels of feature maps. Finally, a softmax layer is applied to generate the final segmentation.

F. LOSS FUNCTION

To reduce the impact of data imbalance during training, we utilize generalized Dice loss (GDL) [28] to optimize our

network, which is defined as:

$$GDL = 1 - \frac{2}{K}\sum_{k\in K}\frac{w_k \cdot \sum_n p_{nk} r_{nk}}{w_k \cdot \left(\sum_n p_{nk} + \sum_n r_{nk}\right)} \tag{8}$$

where p is the softmax output of the network and r is the one-hot encoding of the ground truth segmentation maps, each with K classes and N voxels. $w_k = 1/\left(\sum_{n=1}^{N} r_{kn}\right)^2$ is the weight to provide invariance for different label set properties. We use both 2D loss GDL_{2D} and 3D loss GDL_{3D} to train our model jointly.

III. RESULTS AND DISCUSSION

A. DATASET

Our dataset consists of 134 contrast-enhanced CT images, which collected from Shandong Cancer Hospital Affiliated to Shandong University under the approval of the institutional review board. All CT images used in this study were acquired under pulmonary CT examination using a Philips Brilliance 128i CT scanner (Philips Healthcare, Amsterdam, Netherlands) with a standard clinical protocol of 120 kV voltage, 220 mA current, 1.0 helical pitch, 64×0.625 mm collimation, and 5-mm reconstruction interval. Using an imaging matrix of 512×512 pixels, the pixel size associated with the scans ranged from 0.58 to 0.98 mm. All scans were annotated by two radiologists with more than ten years of experience in CT imaging of thoracic malignancies. They outlined the boundaries of the primary tumors on a transversal plane using Itk snap software (version 3.4; www.itksnap.org) [29]. Each radiologist reviewed the segmented images and any discrepancies were resolved by discussion until a consensus was reached.

B. IMPLEMENTATION

The dataset was randomly split into three subsets, with 84, 20, and 30 subjects for training, validation, and testing respectively. The proposed CNN was trained on an NVIDIA 1080Ti GPU, with 11GB of RAM for 100 epochs. Approximate training time was 12 hours. The model was trained using Adam Optimizer [30] with following hyperparameters: learning rate = 0.001, beta1 = 0.9, beta2 = 0.999 and epsilon = 1e-8. Learning rate was reduced by a factor of 5 whenever the validation loss has not reduced in the last 20 epochs. The code was written in PyTorch Library [31] using Python. We did not use data augmentation techniques such as rotation, scaling, elastic deformations, mirroring, etc., to focus on the discussion of network structure. Instead of standard ReLU, we use the LeakyReLU [32] as activation function in our proposed HSN. Compared to standard ReLU, LeakyReLU can retain the negative part of the feature information, thus preventing the optimization from getting trapped into the local minimum. For 2D CNN, we use batch normalization [33] to normalize the inputs to reduce the internal covariate shift problem. For 3D CNN with small batch size, we use instance normalization [34] due to the superior performance of instance normalization in the case of small batch size [35].

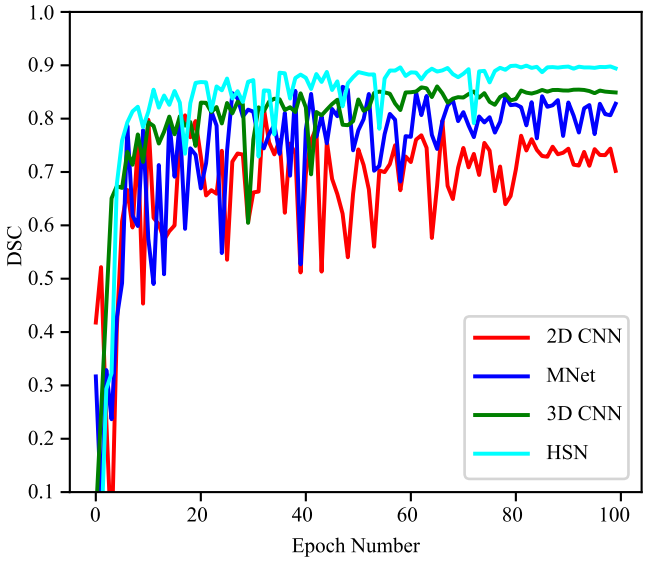

FIGURE 8. Mean DSC on the validation set along with the training progress.

C. EVALUATION METRICS

We quantitatively evaluated the segmentation accuracy using Dice similarity score (DSC), sensitivity, and precision. The DSC measures the similarity between the segmentation results and ground truth. The DSC is defined as follows:

$$DSC = \frac{2TP}{FP + 2TP + FN} \tag{9}$$

where TP is the number of true positives, FP is the number of false positives and FN is the number of false negatives. Sensitivity is defined as:

$$Sensitivity = \frac{TP}{TP + FN} \tag{10}$$

Precision is defined as:

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

DSC, sensitivity, and precision are all measures of voxel-wise overlap between the segmentation results and ground truth. The higher the values, the better the segmentation performance.

D. COMPARED METHODS

1) 3D CNN

We performed a pure 3D CNN similar to 3D U-net [17]. Briefly, each CT volume was resampled to $256 \times 256 \times 64$ to fit the 11GB GPU memory maximumly. Each layer in the encoder consists of two $3 \times 3 \times 3$ convolutions with instance normalization and LeakyReLU. We adopted $3 \times 3 \times 3$ convolutions with strides of 2 to gradually downsize image dimensions by a factor of 2 and simultaneously double the numbers of feature maps. The initial number of filters was 16, and the endpoint feature size was eight times spatially smaller than the input volume. In the decoder, each layer consists of a 3D bilinear upsampling layer with a factor of 2, followed by

TABLE 1. Evaluation results of various methods on the testing set.

		2D CNN	M-net	3D CNN	HSN
DSC	median	0.751	0.840	0.844	0.898
	mean ± std	0.692 ± 0.190	0.789 ± 0.123	0.840 ± 0.049	0.888 ± 0.033
Sensitivity	median	0.690	0.0.849	0.863	0.889
	mean ± std	0.690 ± 0.193	0.819 ± 0.125	0.830 ± 0.076	0.872 ± 0.059
Precision	median	0.845	0.845	0.863	0.925
	mean ± std	0.766 ± 0.201	0.781 ± 0.154	0.856 ± 0.060	0.909 ± 0.048

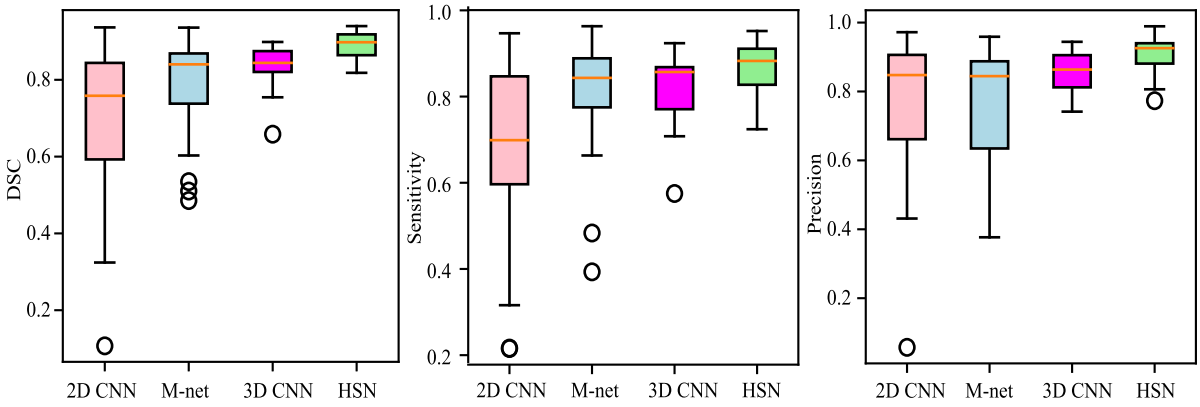

FIGURE 9. Boxplots of evaluation metrics on the testing set with various methods.

two 3 × 3 × 3 convolutions with instance normalization and LeakyReLU. Shortcut connections were used to provide the decoder with detailed information from the encoder. In the end, the output segmentation was resampled to the original resolution using nearest-based sampling.

2) 2D CNN

Compared with the 3D model, the 2D model allows larger resolution images as input. Therefore, full resolution slices were used to leverage detailed spatial context. We performed the pure 2D CNN similar to 3D CNN except that the 3D convolutions were replaced by 2D convolutions and the 3D bilinear upsampling layers were replaced by 2D bilinear upsampling layers.

3) M-NET

M-net [38] is a CNN based method that originally proposed for brain structures segmentation from Magnetic Resonance Images (MRI). Briefly, it is an end-to-end trainable network that takes a stack of consecutive slices as input to leverage 3D information and adopt a large 3D kernel to output a 2D slice, and the followed convolutions are operated only on 2D information.

All of these networks were trained for 100 epochs using generalized Dice loss function. The mean validation DSC along 100 training epochs of different models is presented in Fig. 8. The results show that the 2D CNN achieves the lowest DSC. M-net achieves better performance than the 2D model, thanks to its ability to combine a slice and its neighbors to learn a wider range of information. However, they both yield inferior performance compared with the 3D model. Our proposed HSN achieves the highest and most robust DSC among all models. We attribute this performance improvement to the ability of HSN to fuse 2D and 3D features into a single model.

We selected the trained models with highest validation DSC for each network to test on the testing set. The segmentation results for all networks on the testing set are reported in Table 1. Fig. 9 presents the boxplot of each evaluation metric for different networks. Based on these results, we can observe that HSN outperforms all other networks in terms of all metrics. The comparison results demonstrate the efficacy of our proposed HSN, indicating that the combination of 3D and 2D features is beneficial to CNN models in lung cancer segmentation.

Fig. 10 shows representative segmentation results of different networks on the testing set. These segmentation results demonstrate that HSN has better performance with high spatial and appearance consistency

E. ABLATION ANALYSIS OF HSN

1) COMPARE OF LOSS FUNCTION

When dealing with severe class imbalance problems, the choice of the loss function is crucial to obtaining accurate segmentation results. We employed generalized Dice loss (GDL) function to address the class imbalance problem in CT images. Since many works have demonstrated Dice loss can achieve more robust results than Cross Entropy loss [19], [36], we focus on the difference in performance between Dice loss and generalized Dice loss.

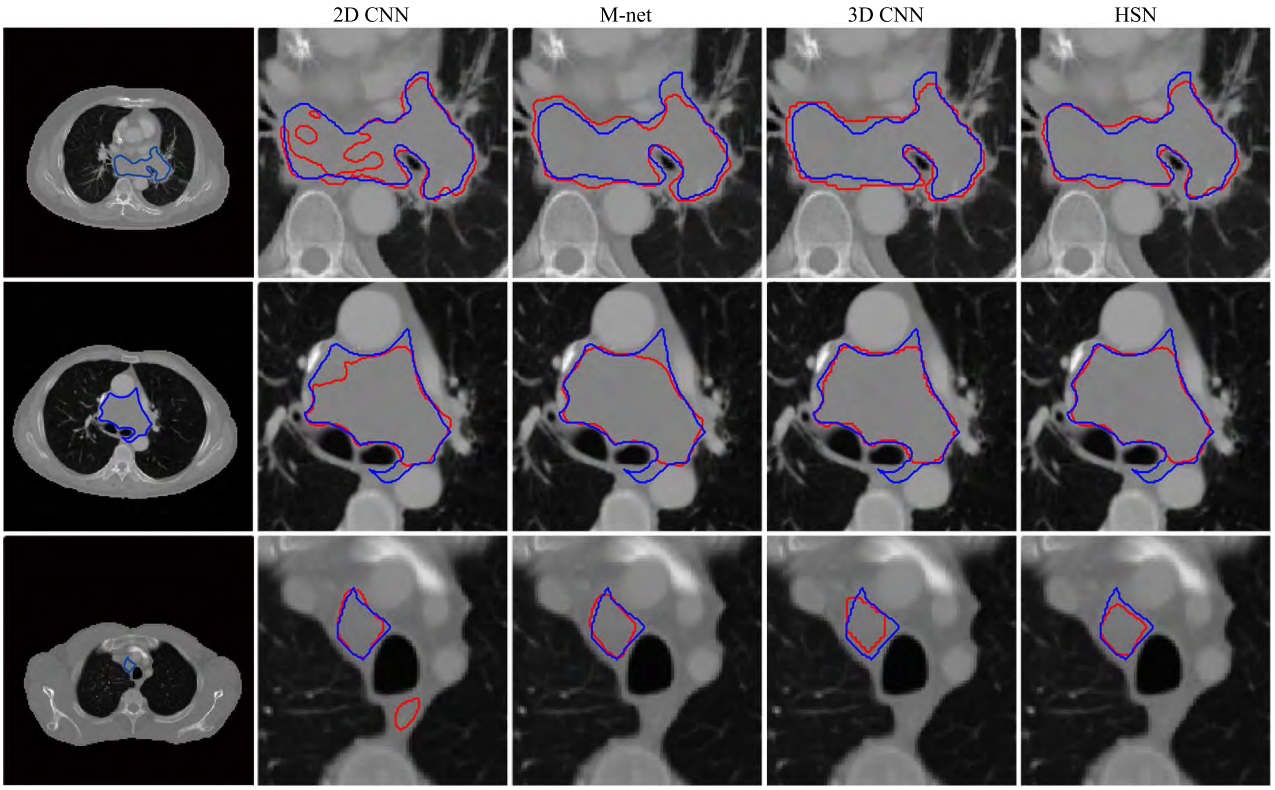

FIGURE 10. This figure shows the qualitative segmentation results of different compared methods on the testing set. The rows represent three slices from different CT scans, and the columns represent the segmentation results produced by various methods. The blue contour corresponds to ground truth, while red contour corresponds to the segmentation results.

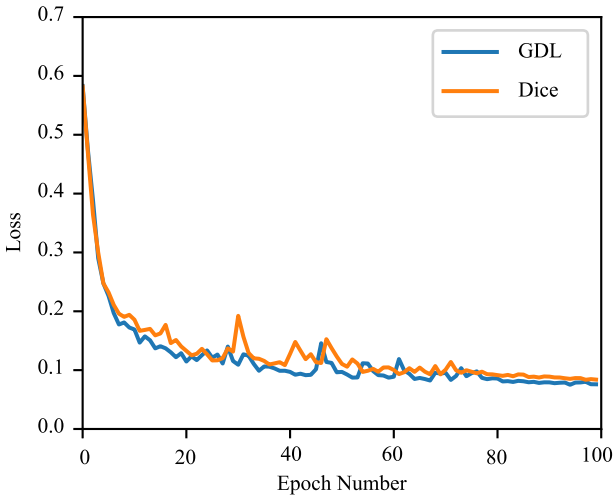

FIGURE 11. The learning process of HSN with different loss functions.

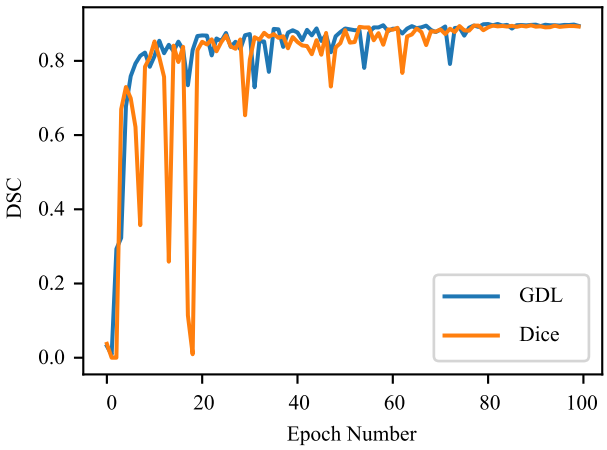

FIGURE 12. Mean DSC on the validation set with different loss functions.

The Dice loss is defined as:

$$Dice = 1 - \frac{2}{K}\sum_{k\in K}\frac{\sum_n p_{nk} r_{nk}}{\sum_n p_{nk} + \sum_n r_{nk}} \tag{12}$$

where p is the softmax output of the network, and r is the one-hot encoding of the ground truth segmentation maps, each with K classes and N voxels. Compared with the Dice loss, GDL introduces a weight w_k to provide invariance to the different label set properties.

We retrained the proposed HSN with Dice loss function and tested it on the test set. Fig. 11 shows the learning process of HSN with different loss functions and Fig. 12 shows the mean DSC on the validation set along with the training progress. In Fig. 11, the loss curve of the model trained with GDL is lower than that trained with Dice loss and is more robust. In Fig. 12, the validation DSC curve of

TABLE 2. Evaluation results of various ablation experiments on the testing set.

		HSN	HSN_Dice	HSN-3D	HSN-N	HSN-L
DSC	median	0.898	0.893	0.889	0.882	0.885
	mean± std	0.888 ± 0.033	0.881 ± 0.046	0.877 ± 0.055	0.872 ± 0.064	0.869 ± 0.068
Sensitivity	median	0.889	0.898	0.893	0.871	0.901
	mean± std	0.872 ± 0.059	0.880 ± 0.068	0.874 ± 0.073	0.850 ± 0.083	0.878 ± 0.075
Precision	median	0.925	0.896	0.888	0.918	0.881
	mean± std	0.909 ± 0.048	0.886 ± 0.059	0.884 ± 0.062	0.901 ± 0.069	0.866 ± 0.088

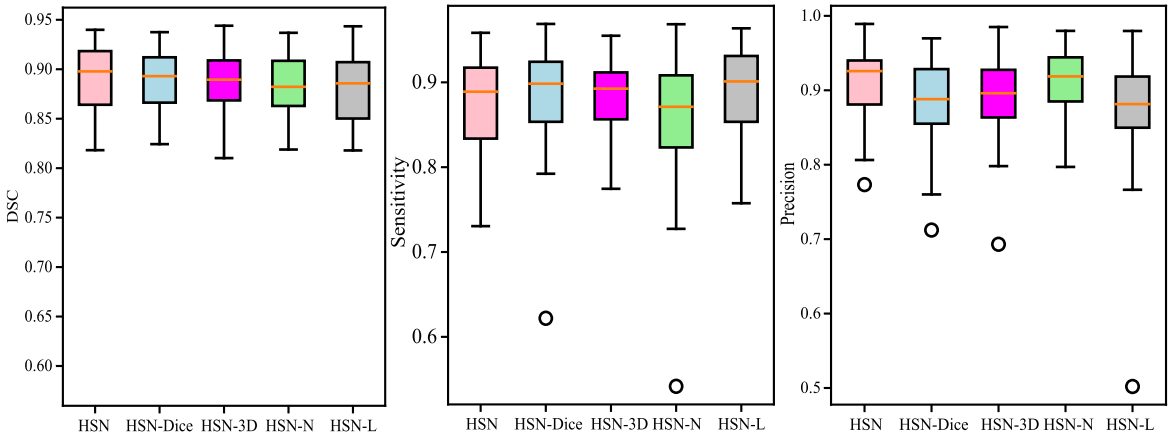

FIGURE 13. Boxplots of evaluation metrics on the testing set with various ablation experiments.

the model trained with GDL is smoother, indicating that GDL is more robust to challenging CT image segmentation. The quantitative results for the testing set are presented in Table 2 and Fig. 13. Compared with the original HSN, the model trained with Dice loss has a lower DSC, a lower precision, and a higher sensitivity, but all have higher deviations. The results demonstrate that GDL is an effective objective function to solve the class imbalance problem and achieve robust results in our lung cancer segmentation task.

2) THE EFFECTIVE OF S3D CONVOLUTION

Our data set consists of CT images of highly anisotropic dimensions. Therefore, we proposed to use S3D convolutions to tackle this problem. To investigate the impact of S3D convolutions on the segmentation performance, we trained HSN with standard 3D convolutions. In particular, we replaced each S3D convolution with standard 3D convolution in HSN while preserving the entire network architecture. The evaluation results are summarized in Table 2. Boxplots of all metrics on the testing set are shown in Fig. 13.

The results shown in Table 2 and Fig. 13 indicate that the use of S3D convolution in our model leads to a performance boost against the standard 3D convolutional version by 1.1% in terms of mean DSC. The results show the advantages of employing S3D convolutions for anisotropic CT images segmentation by decomposing 3D learning into 2D convolutions to learn intra-slice features and 1D convolutions to learn inter-slice features.

3) THE EFFECTIVE OF DILATION RATE

The goal of 2D CNN is to provide fine-grained semantic information about smaller and less salient objects for accurate segmentation. Therefore, it is essential to preserve high spatial resolution on the output feature maps. Simply reducing pool layers or striding convolutional layers lead to a reduction in the receptive field. Thus, We proposed to use dilated convolutions to enlarge the receptive field to learn long-distance contextual information while maintaining the resolution on the output feature maps.

To investigate whether the proposed dilated CNN helps to learn fine-grained semantic information, we compared it with its standard 2D convolutional version, i.e., HSN-N, which had the same architecture as HSN but all of its 2D convolutional layers were performed without dilation. As stated in [37], a convolutional kernel with large dilation rate is too sparse to capture any local information, leading to "gridding issue". To investigate the impact of large dilation rates on the segmentation performance, we also built and evaluated another HSN model with larger dilation rate, i.e., 2D HSN-L, which increased the dilation rate of 3 in the original HSN to 5.

The evaluation results are shown in Table 2 and Fig. 13. We can observe that HSN with dilation rate of 3 achieved the best DSC. The results suggest that dilated convolution can boost the segmentation performance, but the rate of dilation should be carefully considered.

Fig. 14 presents qualitative segmentation results of different ablation experiments on the testing set. These

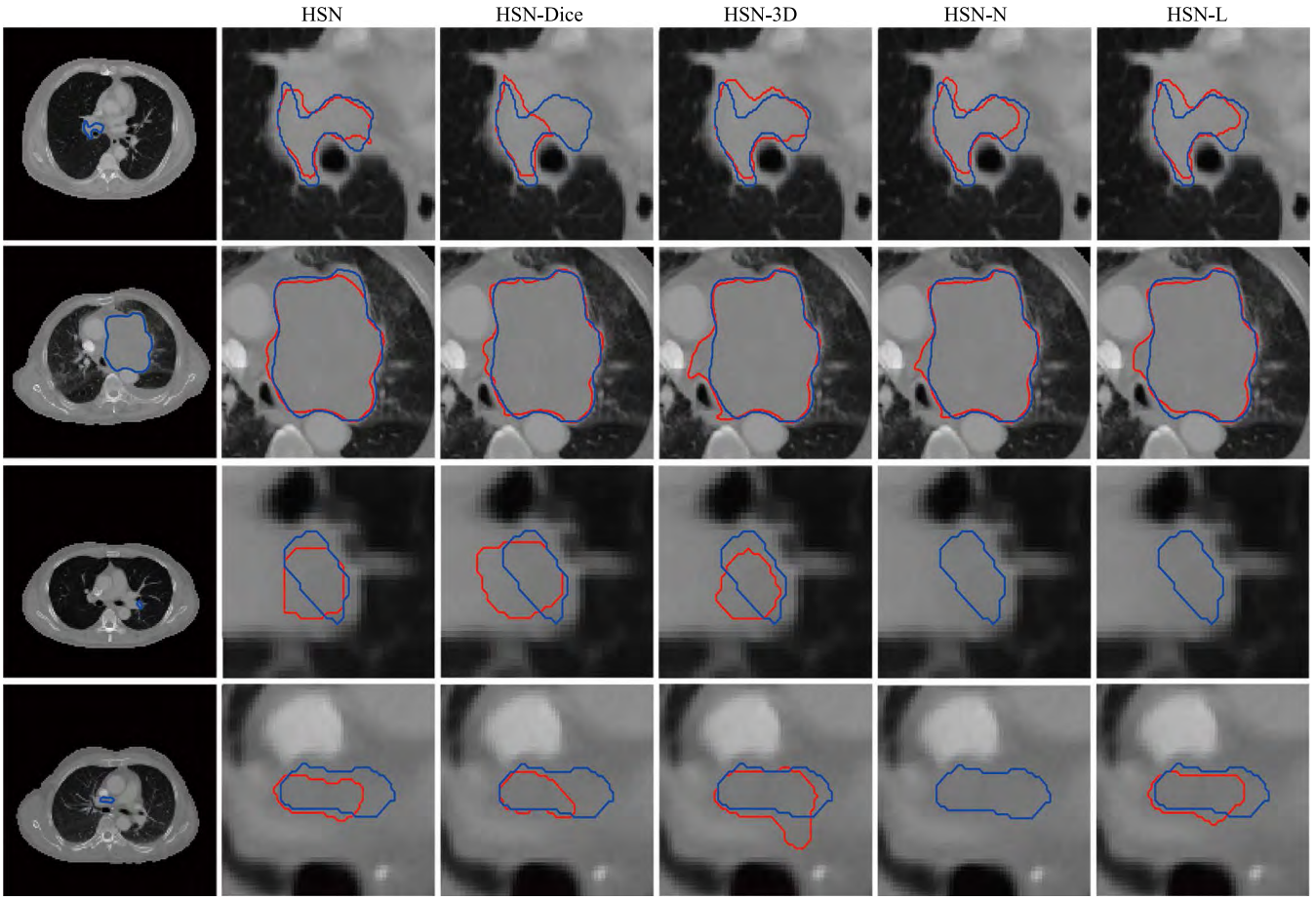

FIGURE 14. This figure shows the qualitative segmentation results of different ablation experiments on the testing set. The rows represent four slices from different CT scans, and the columns represent the segmentation results produced by various methods. The blue contour corresponds to ground truth, while red contour corresponds to the segmentation results.

segmentation results demonstrate that HSN behaves very well in segmenting lung cancers, even for small regions, which we attribute to the long-range 3D contextual information and fine-grained 2D semantic segmentation learned by our model.

IV. CONCLUSION

Automatic non-small cell lung cancer segmentation of CT image can provide precise cancer contours and contribute to the construction of computer-aided diagnostic systems. How to effectively design a model for accurate cancer segmentation is a hot topic in the field of medical image analysis. Over the past few years, deep learning techniques, especially deep convolution neural networks, have been widely used for medical image segmentation. However, due to the complexity of medical images, there is no general-purpose method for obtaining accurate segmentation. In this paper, we proposed a novel end-to-end deep convolutional network for small cell lung cancer segmentation of CT image, capable of fusing 2D and 3D features for better segmentation performance. Particularly, we developed a light-weight 3D CNN to learn long-range 3D contextual information and developed a 2D CNN to capture fine-grained semantic information. Then we proposed a hybrid features fusion module to fuse the 2D and 3D features effectively. Our approach combined the advantages of 2D and 3D CNN to learn sufficient information, while at the same time being efficient in terms of computation and memory requirement. Our experiments showed that the HSN achieves better performances than other state-of-the-art methods.

Some limitations of our study should be acknowledged. First, the CT scans were acquired at a single center, and more scans from different institutions would be needed to improve the generalization performance of our model. Second, scans of healthy people were not included in the dataset. Incorporating the data of healthy population into the training process would further improve the performance of the model. Future work will focus on these limitations and investigate whether the segmentation results might help doctors treat cancer in the clinical setting.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, Cancer J. Clin.*, vol. 68, no. 6, pp. 394–424, 2018.

[2] W. Chen, R. Zheng, P. D. Baade, S. Zhang, H. Zeng, F. Bray, A. Jemal, X. Qin, and J. He, "Cancer statistics in China, 2015," *CA, Cancer J. Clin.*, vol. 66, no. 2, pp. 115–132, Jan. 2016.

[3] J. P. van Meerbeeck, D. A. Fennell, and D. K. De Ruysscher, "Small-cell lung cancer," *Lancet*, vol. 378, no. 9804, pp. 1741–1755, 2011. [Online]. Available: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(11)60165-7/fulltext
[4] B. W. Carter, B. S. Glisson, M. T. Truong, and J. J. Erasmus, "Small cell lung carcinoma: Staging, imaging, and treatment considerations," *Radiographics*, vol. 34, no. 6, pp. 1707–1721, 2014.
[5] S. E. Schild, L. Zhao, J. A. Wampfler, T. B. Daniels, T. Sio, H. J. Ross, H. Paripati, R. S. Marks, J. Yi, H. Liu, Y. He, and P. Yang, "Small-cell lung cancer in very elderly (≥ 80 years) patients," *Clin. Lung Cancer*, to be published. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1525730419301111
[6] S. J. Swensen, J. R. Jett, T. E. Hartman, D. E. Midthun, J. A. Sloan, A.-M. Sykes, G. L. Aughenbaugh, and M. A. Clemens, "Lung cancer screening with CT: Mayo clinic experience," *Radiology*, vol. 226, no. 3, pp. 756–761, Mar. 2003.
[7] H. Wei, F. Yang, Z. Liu, S. Sun, F. Xu, P. Liu, H. Li, Q. Liu, X. Qiao, and X. Wang, "Application of computed tomography-based radiomics signature analysis in the prediction of the response of small cell lung cancer patients to first-line chemotherapy," *Exp. Therapeutic Med.*, vol. 17, no. 5, pp. 3621–3629, 2019.
[8] D. Lee, J. Y. Rho, S. Kang, K. J. Yoo, and H. J. Choi, "CT findings of small cell lung carcinoma: Can recognizable features be found?," *Medicine*, vol. 95, no. 47, Nov. 2016, Art. no. e5426.
[9] R. Roy, T. Chakraborti, and A. S. Chowdhury, "A deep learning-shape driven level set synergism for pulmonary nodule segmentation," *Pattern Recognit. Lett.*, vol. 123, pp. 31–38, May 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016786551830641X
[10] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian, "Multi-scale convolutional neural networks for lung nodule classification," in *Proc. Int. Conf. Inf. Process. Med. Imag.* Cham, Switzerland: Springer, 2015, pp. 588–599.
[11] J. Jiang, Y.-C. Hu, C.-J. Liu, D. Halpenny, M. D. Hellmann, J. O. Deasy, G. Mageras, and H. Veeraraghavan, "Multiple resolution residually connected feature streams for automatic lung tumor segmentation from CT images," *IEEE Trans. Med. Imag.*, vol. 38, no. 1, pp. 134–144, Jul. 2018.
[12] M. Havaei, N. Guizard, H. Larochelle, and P.-M. Jodoin, "Deep learning trends for focal brain pathology segmentation in MRI," in *Machine Learning for Health Informatics*. Cham, Switzerland: Springer, 2016, pp. 125–148.
[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[14] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2843–2851.
[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.
[16] 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.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[17] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2016, pp. 424–432.
[18] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.
[19] R. Zhang, L. Zhao, W. Lou, J. M. Abrigo, V. C. Mok, W. C. Chu, D. Wang, and L. Shi, "Automatic segmentation of acute ischemic stroke from DWI using 3-D fully convolutional densenets," *IEEE Trans. Med. Imag.*, vol. 37, no. 9, pp. 2149–2160, Sep. 2018.
[20] J. Chen, L. Yang, Y. Zhang, M. Alber, and D. Z. Chen, "Combining fully convolutional and recurrent neural networks for 3D biomedical image segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3036–3044.
[21] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 305–321.
[22] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5533–5541.
[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.
[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[25] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 4278–4284.
[26] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[27] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—Improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4353–4361.
[28] W. R. Crum, O. Camara, and D. L. G. Hill, "Generalized overlap measures for evaluation and validation in medical image analysis," *IEEE Trans. Med. Imag.*, vol. 25, no. 11, pp. 1451–1461, Nov. 2006.
[29] P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *NeuroImage*, vol. 31, no. 3, pp. 1116–1128, Jul. 2006.
[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[31] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *Proc. NIPS*, 2017, pp. 1–4.
[32] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," 2015, *arXiv:1505.00853*. [Online]. Available: https://arxiv.org/abs/1505.00853
[33] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: https://arxiv.org/abs/1502.03167
[34] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*. [Online]. Available: https://arxiv.org/abs/1607.08022
[35] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, "No new-net," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum, Eds. Cham, Switzerland: Springer, 2019, pp. 234–244.
[36] L. Wang, S. Wang, R. Chen, X. Qu, Y. Chen, S. Huang, and C. Liu, "Nested dilation networks for brain tumor segmentation based on magnetic resonance imaging," *Frontiers Neurosci.*, vol. 13, p. 285, Apr. 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2019.00285
[37] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1451–1460.
[38] R. Mehta and J. Sivaswamy, "M-net: A convolutional neural network for deep brain structure segmentation," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, 2017, pp. 437–440.

WEI CHEN received the B.S. degree in electrical engineering and its automation and the M.S. degree in control science and engineering from Qufu Normal University, China, in 2012 and 2015, respectively. He is currently pursuing the Ph.D. degree in biomedical engineering with Shandong University. His current research interests include deep learning and medical image analysis.

HAIFENG WEI received the bachelor of medicine degree in medical imaging from Taishan Medical College, China, in 2003, the M.S. degree in imaging and nuclear medicine from Dalian Medical University, China, in 2006, and the Ph.D. degree in imaging and nuclear medicine from Shandong University, China, in 2018. Since 2009, he has been a Lecturer with the Department of Radiology, Affiliated Hospital, Shandong University of Traditional Chinese Medicine, China. His research interests include imaging diagnosis and medical image analysis.

SUTING PENG received the B.S. degree from the School of Control Science and Engineering from Shandong University, Jinan, China, in 2017, where she is currently pursuing the M.S. degree. Her research interests include medical image segmentation and machine learning.

JIAWEI SUN received the B.S. degree in biomedical engineering from Shandong University, China, in 2017, where she is currently pursuing the M.S. degree. Her current research interests include deep learning and medical image processing.

XU QIAO received the B.S. degree in mathematical statistics and the M.S. degree in probability and mathematical statistics from Shandong University, China, in 2004 and 2007, respectively, and the Ph.D. degree in information science from Ritsumeikan University, Japan, in 2010. From 2012 to 2018, he was a Lecturer with the School of Control Science and Engineering, Shandong University. Since 2018, he has been an Associate Professor of biomedical engineering with the School of Control Science and Engineering, Shandong University. His research interests include imaging diagnosis and medical image analysis. From 2010 to 2012, he was a Research Fellow with the Japan Society for the Promotion of Science (JSPS).

BOQIANG LIU received the B.S. degree in automation and the M.S. degree in control science and engineering from Shandong University, China, in 1982 and 1986, respectively, and the Ph.D. degree in biomedical engineering from Tianjin University, China, in 2005. Since 2000, he has been a Professor with the Biomedical Engineering Department, School of Control Science and Engineering, Shandong University. His research interests include the development of medical instrumentation and medical image analysis.