IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 8, 2019, accepted December 22, 2019, date of publication December 26, 2019, date of current version January 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2962608

S^{3}egANet: 3D Spinal Structures Segmentation via Adversarial Nets

TIANYANG LI[1,2,3], BENZHENG WEI[2,3], JINYU CONG[2,3], XUZHOU LI[4], AND SHUO LI[5,6]

[1]College of Science and Technology, Shandong University of Traditional Chinese Medicine, Jinan 250355, China
[2]Center for Medical Artificial Intelligence, Shandong University of Traditional Chinese Medicine, Qingdao 266112, China
[3]Qingdao Academy of Chinese Medical Sciences, Shandong University of Traditional Chinese Medicine, Qingdao 266112, China
[4]School of Information Engineering, Shandong Youth University of Political Science, Jinan 250103, China
[5]Department of Medical Imaging and Medical Biophysics, Western University, London, ON N6A 5W9, Canada
[6]Digital Imaging Group of London, London, ON N6A 4V2, Canada

Corresponding authors: Benzheng Wei (wbz99@sina.com) and Shuo Li (slishuo@gmail.com)

This work was supported in part by the Natural Science Foundation of China under Grant 61872225, in part by the Natural Science Foundation of Shandong Province under Grant ZR2015FM010, in part by the Project of Shandong Province Higher Educational Science and Technology Program under Grant J15LN20, and in part by the Project of Shandong Province Medical and Health Technology Development Program under Grant 2016WS0577.

ABSTRACT 3D spinal structures segmentation is crucial to reduce the time-consumption issue and provide quantitative parameters for disease treatment and surgical operation. However, the most related studies of spinal structures segmentation are based on 2D or 3D single structure segmentation. Due to the high complexity of spinal structures, the segmentation of 3D multiple spinal structures with consistently reliable and high accuracy is still a significant challenge. We developed and validated a relatively complete solution for the simultaneous 3D semantic segmentation of multiple spinal structures at the voxel level named as the S^{3}egANet. Firstly, S^{3}egANet explicitly solved the high variety and variability of complex 3D spinal structures through a multi-modality autoencoder module that was capable of extracting fine-grained structural information. Secondly, S^{3}egANet adopted a cross-modality voxel fusion module to incorporate comprehensive spatial information from multi-modality MRI images. Thirdly, we presented a multi-stage adversarial learning strategy to achieve high accuracy and reliability segmentation of multiple spinal structures simultaneously. Extensive experiments on MRI images of 90 patients demonstrated that S^{3}egANet achieved mean Dice coefficient of 88.3% and mean Sensitivity of 91.45%, which revealed its effectiveness and potential as a clinical tool.

INDEX TERMS Adversarial nets, spine, segmentation, magnetic resonance imaging, multi-stage, multi-modality, computer-aided detection and diagnosis.

I. INTRODUCTION

Spinal diseases and the associated pain are critical issues of body health, such as: (1) disc degeneration [1] is a common cause of back pain and stiffness for adults [2]; (2) patients with neural foraminal stenosis may develop symptoms include: pain in the back, muscle weakness, tingling, burning sensations, etc [3], [4]; (3) vertebral fractures cause pain, functional disability and decreased quality of life, which may last for several years [5]. The clinical segmentation methods on spinal diseases were mainly done by means of manual annotation, which were rather tedious, time-consuming, and often subject to inter- and intra- observer variabilities caused by the grading criterion and expertise [6]–[8]. In this regard, 3D spinal segmentation of multiple structures can assist in the disease treatment by providing quantitative parameters, which improves the efficiency and accuracy for spine pathologies diagnosis.

Due to the superior ability to distinguishing soft tissue contrast, MRI images have emerged as the modality of choice for some spinal diseases such as intervertebral disc degeneration [9]. The existing work for MRI image segmentation of the spine mostly uses 2D single structure segmentation method. However, 3D multiple spinal structures segmentation has more merits compared with 2D segmentation for single structure: (1) The value of simultaneous segmentation for multiple structures has been shown in [10] in the scope of spinal diseases diagnosis. Foraminal stenosis, intervertebral disc

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

(a) A segmentation of 3D spinal structures (b) A slice of T1WI (c) A slice of T2WI

FIGURE 1. (a) The solution of 3D multiple spinal structures segmentation features advantages of 3D segmentation and simultaneously segmentation of multiple spinal structures; (b)-(c) Feature extraction from multi-modality MRI images can combine the structural characteristics of different modalities.

degeneration, and lumbar vertebral deformities have strong pathological correlations and automated semantic segmentation of the three structures can significantly promote clinical pathogenesis-based diagnosis [8]. (2) Compared with 2D, 3D spinal MRI image segmentation shows more details, such as 3D structures and the relative position between 3D structures, as shown in Fig. 1(a). (3) 3D spinal MRI image segmentation has wide application in image-guided surgery. Its segmentation results can make the orthopedic surgeon real-time control surgical instruments relative to the patient the precise location of anatomical structures, and thus improve the accuracy of the operation [2].

Segmentation of 3D multiple spinal structures with consistently reliable and high accuracy is still a great challenge in MRI images. Although a large amount of studies has been focused on spinal segmentation, the solution of 3D multiple spinal structures segmentation from MRI images lacks relevant experience due to three primary challenges. Firstly, 3D spinal structures are challenges to distinguish due to the high complexity of spinal structures and pathological MRI images. In addition to identifying complex spinal structures (For example, the structure of L5 quite different from L1, L2, L3, and L4), normal and abnormal structures have a high degree of visual similarity and subtle differences. The segmentation results of the intervertebral disc structure are often affected by fat and degeneration. The fine-grained segmentation of the spinal structures is a great challenge due to the tissue synechia, which often appears in spinal MRI images. Secondly, structural information is not comprehensive on single modality MRI images but the popular segmentation solutions are difficult to apply in different MRI modalities images. Fig. 1(b)-(c) shows that compared with T2WI slice, the edge features of T1WI slice are more clearly presented. Specifically, for L3, L5, IVD L2-L3, and IVD L3-L4. But for L1 and IVD L5-S1, the local structural features on T2WI are better than that on T1WI. Finally, simultaneously segmentation for multiple spinal structures with high accuracy and reliability results must be satisfied. The spatial correlations between intervertebral foramen (IVF), intervertebral discs (IVD), and vertebrae increase the difficulty of multiple spinal structures segmentation.

A. RELATED WORK

The most common traditional approach are based on intensity [11] and shape information [12]. Nevertheless, these models do not explicitly cope with the individual differences of the vertebrae anatomical structure and often suffer from serious boundary leaking problems when the objects have weak boundaries.

Machine learning methods have gained increasing interest in the field of spinal structures segmentation. For example, marginal space learning (MSL) [13] was proposed to detect the spine in CT and MRI images. Unified data-driven estimation framework [14] was proposed to estimate the image displacements to localize IVD and then segment IVD by predicting the foreground and the background probability of each pixel in which the neighborhood intensity vector was used as visual features. A sparse kernel machine [15] based regression method taking hand-crafted features including texture and shape as input to segment disc and vertebral structures from both MRI and CT modalities. However, these methods do not completely solve the problems encountered in 3D multiple spinal structure segmentation due to the difficulty of design. While considering the segmentation accuracy, these methods also need to consider the influence of different modal data on equipment and environment in order to obtain more feature information.

As existing deep learning methods, the representation power of CNNs often leads to successful image prediction results. But the expressive power of regular CNN is usually limited, especially for 3D semantic segmentation. Kinds of literature have proposed deep learning methods to segment spinal structures [2], [16]. These methods indicate the potential for 3D CT or MRI images segmentation of one spine

component solely. A common property across various types of regular CNNs approaches is that all label variables are predicted independently from each other. To assess the joint configuration of many label variables, the method of semantic segmentation using adversarial networks (GAN) [17] was proposed to produce label maps that cannot be distinguished from ground-truth. The adversarial theory has extended to prostate cancer detection [18], brain MRI segmentation [19], and quantification of myocardial infarction [20]. All these works have gained different levels of improvement, which proves the effectiveness of adversarial learning. In addition, GAN has been applied to achieve 2D multiple spinal structures segmentation [10].

Despite the improvements, the solution of 3D multiple spinal structures segmentation with high accuracy and reliability is still a challenge worth exploring. In this regard, we proposed a relatively complete solution for the simultaneous 3D semantic segmentation of spinal structures at the voxel level named as the S^{3}egANet, which shows effective performance for distinguishing of 3D multiple spinal structures.

B. OUR CONTRIBUTIONS

We propose a novel adversarial model based on multi-stage learning approach to segment 3D multiple spinal structures from multi-modality MRI images. Experimental results on 90 sets of 3D multi-modality MRI images demonstrated the superiority of our proposed method.

Our primary contribution of this study include:

1) For the first time, a relatively complete solution was achieved to implement the simultaneous 3D semantic segmentation of multiple spinal structures from multi-modality MRI images.
2) We proposed a novel adversarial model to distinguish the high complexity of spinal structures.
3) We proposed a multi-stage learning strategy that encourages the model to predict right voxel-wised class labels and achieve high accuracy and reliability segmentation simultaneously for multiple spinal structures.
4) We leverage a cross-modality voxel fusion module (CMVF) to establish the relationship of multiple modalities MRI images. The CMVF module can assign different weights to each modality and fuse the feature values in the feature maps.

The rest of this paper is organized as follows: Section 2 describes in detail the proposed S^{3}egANet, including the segmentation network, the discriminative network, the cross-modality voxel fusion module, the multi-stage adversarial learning strategy (MSAL), and the global optimization. Section 3 introduces the details about the dataset, configuration, and experiments. In Section 4, we report our results, conduct a comprehensive performance analysis of our S^{3}egANet on clinical data and discuss the significant influence of our work. Then, conclusions are given in Section 5.

II. METHODOLOGY

Our adversarial nets are implemented by two competing networks: the segmentation network (Subsection A) and the discriminative network (Subsection B). The segmentation network generates predicted masks and then the discriminative network receives either the predicted masks or ground truth masks. Finally, our model outputs a scalar of 0 or 1 for each voxel, which shows whether the segmentation network can generate sufficiently accurate predicted masks to cheat the discriminative network. After completing the feature extraction encode (Subsubsection 1), we deployed a cross-modality voxel fusion module (Subsubsection 2) to combine the features from different modality MRI images. In order to improve the expressivity of the model, our adversarial nets were trained by a multi-stage adversarial learning strategy (Subsection C), each of which is carefully designed for the characteristics of 3D spinal structure segmentation.

A. SEGMENTATION NETWORK

The segmentation network consists of a multi-modality autoencoder module (MMAE) and a cross-modality voxel fusion module (CMVF), as shown in Fig. 2. The network builds the spatial feature extraction layers firstly based on stacks of 3D convolution [21], [22] to learn a comprehensive representation of the 3D data. And then builds a CMVF module to share the representation of the feature from multi-modality MRI images. Finally, the expansive path based on stacks of deconvolution (Deconv) [23] to propagate context information to higher resolution layers. The detailed configuration parameters of the segmentation network are delicately design.

1) MULTI-MODALITY AUTOENCODER MODULE (MMAE)

The MMAE module consists of multiple contracting paths as an encoder to receive different modalities input and an expansive path as a decoder to generate the predicted results. Each contracting path comprises nine convolutions with kernel size $3 \times 3 \times 3$ to produce a set of feature maps, which are further applied by a batch normalization layer [24] and a rectified linear unit (ReLU) [25]. After each three continuous convolution + batch normalization + ReLU operations, a $2 \times 2 \times 2$ max pooling layer with stride 2 for down-sampling is deployed. The batch normalization operation causes the activation input to fall in a region where the nonlinear function is sensitive to the input, such that the input a small change will result in a significant change in the loss function. The max pooling can reduce the deviation of the estimated mean caused by the convolutional layer parameter error and preserve more spinal structures information.

Each contracting path comprises an up-sampling process of the feature map with a $2 \times 2 \times 2$ deconvolution that halves the number of feature channels, each followed by a ReLU. At the final two $1 \times 1 \times 1$ convolutional layers are used to map each component feature vector to the desired number of classes. For one thing, compared to the single convolutional

layer, the cascade of two layers reduce the number of input channels and modify the algorithm complexity. For the other, with the size of the feature map is unchanged (Without change the resolution), $1 \times 1 \times 1$ convolutional layers making the segmentation network deep and significantly increase the non-linear characteristics.

2) CROSS-MODALITY VOXEL FUSION MODULE (CMVF)

In order to combine the imaging advantages of multi-modality MRI images, a cross-modality voxel fusion module is deployed after the feature extraction of different contracting paths. Our cross-modality voxel fusion module consists of a cross-modality convolution [26] and convolutional LSTM layers.

After the contracting paths, each modality generates feature maps of size $C \times H \times W \times c$, where C, W and H are feature dimensions, and c is the number of channels. We stack the features of the same channels from modalities number n into one stack. After reshaping, we have $c \times n \times H \times W$ 2D feature maps. Our cross-modality convolution performs 3D convolution with the kernel size $n \times 1 \times 1$, followed by a $n \times 1 \times 1$ average pooling. The repeated application of two convolutional LSTM (ConvLSTM) layers is deployed after the cross-modality convolution to better exploit the spatial and sequential correlations of consecutive slices. ConvLSTM captures the sequential dependencies and further mixes structural features from multi-modality data. Given that i_t, f_t, $\tilde{c}_t$, o_t, and h_t represent the input gates, forget gates, cell, output gates, and final state respectively, the ConvLSTM is defined as following:

$$\begin{cases} i_t = \sigma(x_t * Wxi + h_{t-1} * W_{hi} + b_i) \\ f_t = \sigma(x_t * Wxf + h_{t-1} * W_{hf} + b_f) \\ \tilde{c}_t = tanh(x_t * W_{x\tilde{c}} + h_{t-1} * W_{h\tilde{c}} + b_{\tilde{c}}) \\ c_t = \tilde{c}_t \odot i_t + c_{t-1} \odot f_t \\ o_t = \sigma(x_t * Wxo + h_{t-1} * Who + b_o) \\ h_t = o_t \odot tanh(c_t) \end{cases} \tag{1}$$

where σ and $tanh$ are the sigmoid and hyperbolic tangent functions; $*$ and $\odot$ represent the convolution operation and Hadamard product respectively.

B. DISCRIMINATIVE NETWORK

The discriminative network learns the ground truth in the adversarial training. During training, the adversarial nets receive the predicted maps from the segmentation network firstly and manual maps from ground truth, then output a single scalar representing whether the inputs are from the segmentation network or ground truth. When a strong confrontation occurs, the discriminative network prompts the segmentation network to detect mismatches in a wide range of higher-order statistics between predicted segmentation maps and ground truth.

We propose a practical and straightforward discriminative network. It consists of the repeated application of three $5\times5\times5$ convolutions, each followed by a batch normalization operation, a ReLU, and a $2 \times 2 \times 2$ average pooling operation with stride 2 for down-sampling. In order to ensure the computing capacity of the neural network, we double the number of feature channels at each down-sampling step. At the final layer, a $1 \times 1 \times 1$ convolution is used to map each component feature vector to the desired number of classes. A dropout function is set between the two $1 \times 1 \times 1$ convolutional layers to eliminates the joint adaptability and enhances the generalization ability.

C. MULTI-STAGE ADVERSARIAL LEARNING STRATEGY (MSAL)

The number of voxels from spinal MRI images in different classes varies greatly. In particular, the ratio of the class with the most voxels (background) to the class with the least voxels (IVF) is about 850: 1. Besides, simultaneously segmentation for multiple spinal structures with high accuracy and reliability results is still not completely solved. To solve these challenge and obtain highly reliable and accurate segmentation result of multiple spinal structures, we design a novel strategy of adversarial learning which empirically has stable performance.

Let us consider the learning strategy of primary GAN. The objective function of primary GAN adjusts parameters for segmentation network G to minimize the probability of the samples from G to be recognized and adjust parameters for discriminative network D to maximize the probability of assigning the correct label to both training examples and samples from G, as if they are following the two-player min-max game with value function $V(G, D)$:

$$\min_G \max_D V(G, D) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))], \tag{2}$$

where $x \sim p_{data}(x)$ denotes the real data samples, $D(x)$ denotes the probability that x came from the real data. $z \sim p_z(z)$ denotes the fake data samples. $G(z)$ denotes the newly generated data.

In connection with the characteristics of adversarial nets, we design an high efficiency multi-stage mode to training our adversarial network. Our learning strategy contains two training stages for the segmentation network, corresponding to solve the imbalance of different spinal structures and high accuracy and reliability segmentation simultaneously for multiple spinal structures respectively. The first stage aims at distinguishing spinal major structure and building the spatial relationship between different structures. Relatively, the second stage deals explicitly with preserving accurate structural information to segment the edge of the spinal structure.

1) MULTI-STAGE CROSS-ENTROPY LOSS OF SEGMENTATION NETWORK

Our segmentation network $S(x)$ is a function parametrized as a network predicting the confidences for K classes of image voxels and softmax is employed to obtain the probability of sample x belonging to each class. After obtaining the

segmentation result, the predicted map and ground truth map are input to a discriminative network for further regularization in the training. The $S(x)$ is trained with a multi-stage learning strategy. In the first stage, a weighted cross-entropy learning strategy (WCEL) is designed and the segmentation loss $\mathcal{L}(S_1)$ is:

$$\mathcal{L}(S_1) = -\sum_{i=1}^{C\times H\times W} w_k y_i \log p_i, \tag{3}$$

where y denotes real label and p_i denotes the probability. $w_k = \sum_{m=1}^{M}(C\times H\times W)/A_k$ denotes the k-th class's weight, where A_k is the voxel amounts of the k-th class in training dataset. Our S^{3}egANet optimizes the network with different weights. This method reduced the training difficulty of GAN and improve the stability of our Seg^{3}ANet.

During the second stage, the effect of imbalance in the training samples has been greatly reduced due to the weighted learning strategy which is balanced by the voxel amounts of different classes. And corresponding to that, the prediction accuracy of different classes would inevitably prohibit the segmentation network to predict right voxel-wised class labels. To solve this issue, we used a focal cross-entropy learning strategy (FCEL) and adopt the focal loss [27] as following:

$$\mathcal{L}(S_2) = -\sum_{i=1}^{C\times H\times W} (1-p_i)^{\gamma} y_i \log p_i. \tag{4}$$

The term $(1-p_i)^{\gamma}$ controls the contrast of loss value. The larger γ become, the larger effect between loss value of easy and hard classes become (we found $\gamma = 1$ to work best in our experiments).

2) THE ADVERSARIAL LOSS OF DISCRIMINATIVE NETWORK

The discriminative network of S^{3}egANet $D(x)$ is an auxiliary adversarial convolutional network. The main function of $D(x)$ is to distinguish whether the input data is from ground truth or the prediction of the segmentation network.

We introduce the loss functions of adversarial network starting from the cross-entropy loss for binary classification:

$$\mathcal{L}(y,\hat{y}) = -y\log(\hat{y}) - (1-y)\log(1-\hat{y}), \tag{5}$$

where y is the label and $\hat{y}$ is the prediction. For the adversarial network, y is marked as 1 or 0. If sample x_r came from the real ground truth, the value of y is marked as 1, as following:

$$\begin{aligned}\mathcal{L}(D_r) &= -(1 * \log D(x_r) + (1-1) * \log(1-D(x_r))) \\ &= -\log D(x_r),\end{aligned} \tag{6}$$

where the $D(x_r)$ denotes the predicted value of the discriminative network. If sample x_f came from the fake data, the value of y is marked as 0, as following:

$$\begin{aligned}\mathcal{L}(D_f) &= -(0 * \log D(x_f) + (1-0) * \log(1-D(x_f))) \\ &= -\log D(1-D(x_f)),\end{aligned} \tag{7}$$

where the $D(x_f)$ denotes the predicted value of the discriminative network. For each i on the total number of batch M, the loss function of the adversarial network is:

$$\mathcal{L}(D) = -\frac{1}{M}\Big(\sum_{x_r}^{M}\log(D(x_r)) + \sum_{x_f}^{M}\log(1-D(x_f))\Big). \tag{8}$$

D. GLOBAL OPTIMIZATION

The optimizer employed in our model is RMSProp algorithm [28] to update our network with dynamic learning rate based on exponential decay, as following:

$$E[g^2]_t = \alpha E[g^2]_{t-1} + (1-\alpha)g_t^2, \tag{9}$$

$$W_{t+1} = W_t - \frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} \odot g_t, \tag{10}$$

where g_t denotes the gradient of the cost function and $E[g^2]_t$ denotes the gradient's mean value of t times square. α, η and ϵ are the moving average parameter (good default value 0.9), learning rate and a parameter added to prevent division by zero, respectively. The initial learning rates of the segmentation network and the discriminative network are set to 0.001 and 0.0001, respectively.

Three types of updates are possible based on the above settings during the training procedure, i.e., an update of only the segmentation network, an update of only the discriminative network, and an update of the whole network (segmentation and discriminator network). The implementation of update the whole network is shown in Algorithm 1. The updates based on segmentation loss (Cross-entropy of multiple classes) and the updates based on discriminative loss (Two cross-entropy functions of binary classification) were performed by a separate optimizer using separate learning rates. Given an adversarial network, the training of the segmentation network minimizes the cross-entropy loss and reduces the performance of the discriminative model. This encourages the segmentation network to generate accurate predictions and these predictions are difficult to distinguish from ground truth.

III. DATA AND EXPERIMENTS

A. DATA AND PRE-PROCESSING

We evaluated our proposed method on the dataset which consists of 90 sets of clinical patients' lumbar scans. Lumbar spine scan was performed in patients, and the neural foramen, intervertebral disc and vertebral were presented simultaneously in the sagittal direction. Due to these patients are examined by different models of vendors, their MRI scans have different parameters. The range of repetition time is from 380 ms to 4,000 ms. Slice thickness from 0.879 mm to 2.0 mm. Pixel spacing from 0.38 mm × 0.38 mm to 0.86 mm × 0.86 mm. We unified the scale of our data and made the number of slices are 32 with and each slice is 256 × 256 pixels. Each set of data has two modalities MRI images, i.e., T1 weighted images (T1WI) and T2 weighted images (T2WI). Each set of multi-modality MRI images was derived from two consecutive scans in a patient's clinical

Algorithm 1 The Training Procedure of the S^{3}egANet

Require:
A dataset of P training MRI images x;
Ground truth maps y;
Minibatch size p;
Maximum epochs Q;
The discriminative network's parameters ω;
The segmentation network's parameters θ;
The learning rate η.

Ensure:
Initialize all parameters $\{\omega, \theta, \eta\}$;
for $\frac{QP}{p}$ **do**
  $x_p \Leftarrow$ fed minibatch p from training dataset
  $g_\theta \Leftarrow \nabla_\theta \mathcal{L}(S)$
  Update the segmentation network:
  $\theta \Leftarrow \theta + \eta \cdot RMSProp(\theta, g_\theta)$
  $x_p, y_p \Leftarrow$ fed minibatch p from training dataset
  $g_\omega \Leftarrow \nabla_\omega \mathcal{L}(D)$
  Update the discriminative network:
  $\omega \Leftarrow \omega + \eta \cdot RMSProp(\omega, g_\omega)$
end for

diagnosis and thus are aligned with each other. Through the selection and repeated confirmation by two radiologists, the spatial position differences of the spine structures of the multi-modality MRI images we used reached an indistinguishable level. Therefore, our segmentation results can be applied to both modalities MRI images simultaneously. According to the reports of spinal surgery, our data includes, but is not limited to, disc degeneration disease, schmorl snode, ankylosing spondylitis, osteoporosis, and spinal stenosis.

B. EXPERIMENTS

We use the standard five-fold cross-validation for performance evaluation and comparison [29]. We divided the data into five groups, each time selecting four groups as a training dataset and one group as a test dataset. After the testing, we calculated the average value as the model metrics. The advantage of this method is that in the case of small amounts of data, all observations are used for training and testing, and each observation is used for testing once. We implemented the proposed method with Python 3.6 based on TensorFlow 1.2 library [30] on a workstation equipped with GPUs of NVIDIA TESLA P100.

C. PERFORMANCE EVALUATION

We employed the challenge evaluation metrics to evaluate the performance of our S^{3}egANet. Dice coefficient is adopted for measuring the accuracy of segmentation results and standard deviation (STD) quantifies the degree of variation. Let us consider TP_n and FP_n are the true positives and false positives of class n, while TN and FN are true negatives and false negatives of background respectively. Dice coefficient and Sensitivity of one class n are defined as:

$$Dice_n = \frac{2TP_n}{2TP_n + FP_n + FN_n}, \tag{11}$$

$$Sensitivity_n = \frac{TP_n}{TP_n + FN_n}, \tag{12}$$

Dice coefficient and Sensitivity are standard metrics calculated on voxel-level confusion matrix.

IV. RESULTS AND ANALYSIS

The S^{3}egANet is a novel adversarial method and achieves 3D semantic segmentation of multiple spinal structures. S^{3}egANet combines the advantages of adversarial nets and multi-stage learning strategy for distinguishing the 3D multiple spinal structures, and leverages the cross-modality voxel fusion module to share the fine-grained representation from multi-modality MRI images. Experimentation on 90 sets of clinical patients' lumbar scans, S^{3}egANet achieved mean Dice coefficient of 88.30%, which verified consistently reliable and high accuracy segmentation for 3D spinal structures. S^{3}egANet also achieved 91.45% mean Sensitivity on multiple spinal structures simultaneously, which demonstrated that S^{3}egANet held the potential to improve the diagnosis efficiency.

A. EFFECTIVENESS OF MULTI-MODALITY INPUT IMAGES

In order to quantitatively analyze the effectiveness of multi-modality input, we conducted comparative experiments by using single modality images and training different networks respectively. All networks adopt identical network architecture (Fig. 2) and single stage training strategies (Equation 2). As shown in Table 1, the experimental results produced by the network with multi-modality input outperform the single modality input by approximately 2-4 percent in Dice coefficient. Compared to single modality data, multi-modality can provide richer complementary information. It is worthy pointing out that the segmentation results generated from networks trained with T1WI image input have higher segmentation accuracy than that with T1WI image input. Table 1 presents the segmentation accuracy of IVD, IVF, and vertebrae. Networks trained with T2WI image improve the accuracy of IVF. One of the main reasons is that the intensity contrast around the IVF and its neighboring regions of the T2WI image is larger than the T1WI image, which reduces the difficulty of IVD recognition.

For the sake of analysis, we use the 2D ground truth slice as the background and color the 2D segmentation slice to cover the background. Fig. 3 presents several examples of segmentation results for different experimental settings, including training with T1WI image, T2WI image, and multi-modality images. It is observed that all the experimental settings can segment vertebrae and IVD with reasonable accuracy. However, different modalities have positive effects on the extraction of local features. Specifically, for L3, L5, IVD L2-L3 and IVD L3-L4 in Fig. 3(b), L1 and IVD L4-L5 in Fig. 3(c). As shown in Fig. 3(d), the prediction boundaries

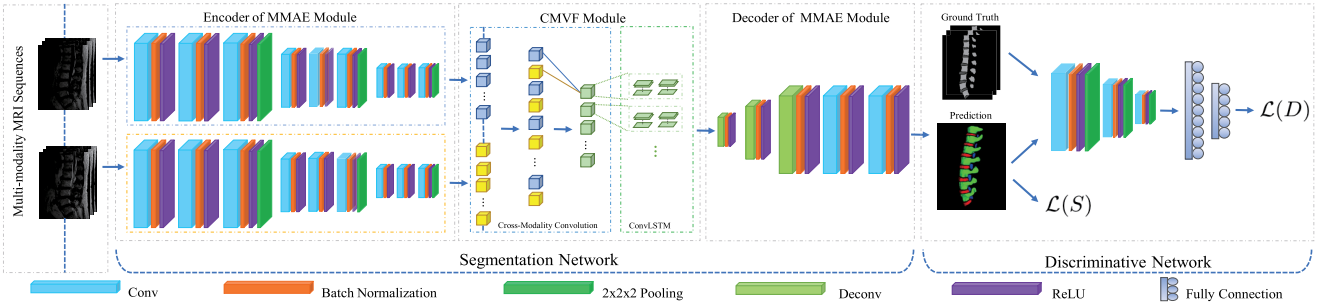
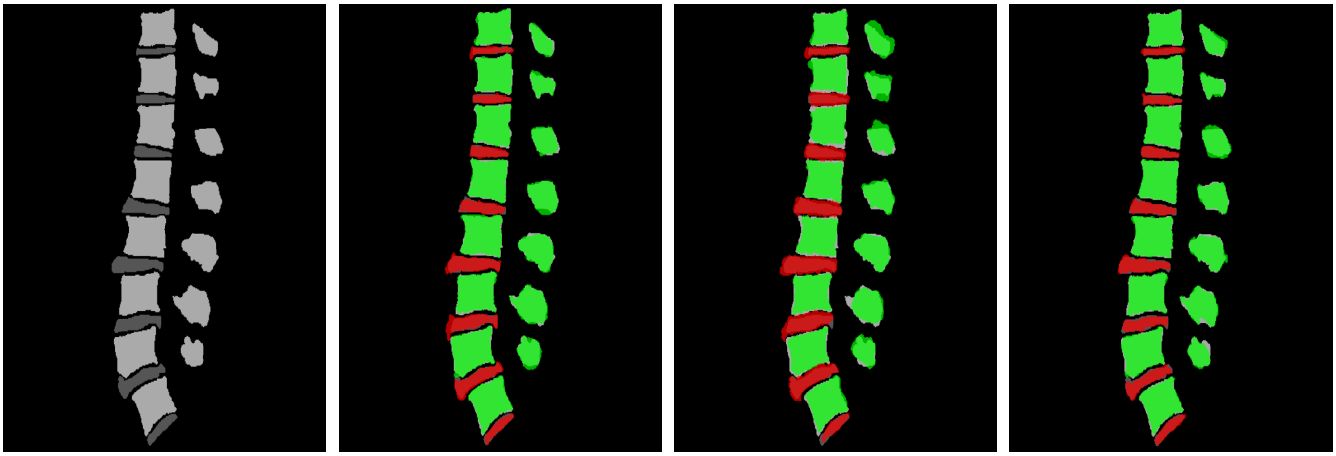


FIGURE 2. The architecture of S3egANet: the segmentation network combines multi-modality autoencoder module and cross-modality voxel fusion module for multi-modality features extraction; the discriminative network determines the source of input and achieves adversarial learning.

TABLE 1. Comparison of segmentation results produced by the network with single modality data input and multi-modality images input.

Modality	Dice_{Mean}	Dice_{IVD}	$\text{Dice}_{Vertebrae}$	Dice_{IVF}	Sensitivity
T1WI	83.61 ± 2.19	85.08 ± 1.78	$\mathbf{83.84 \pm 2.92}$	76.97 ± 1.79	88.23
T2WI	81.29 ± 2.97	80.66 ± 3.25	83.17 ± 2.89	78.83 ± 1.89	90.10
T1WI+T2WI	$\mathbf{85.37 \pm 1.96}$	$\mathbf{86.90 \pm 1.72}$	82.96 ± 2.76	$\mathbf{80.80 \pm 1.75}$	**91.14**



(a) Ground Truth (b) T1WI (c) T2WI (d) Multi-modality MRI

FIGURE 3. The application of multi-modality input images has improved the segmentation performance.

generated from multi-modality MRI images can achieve reasonable accuracy in these cases. These observations confirm that training network with multi-modality images can achieve better segmentation results than using the single modality MRI images as input.

B. EFFECTIVENESS OF LEARNING TECHNIQUES

To investigate the effectiveness of our major technological contributions including both MSAL and CMVF, we compare the segmentation results achieved by different architectures. The architecture of multi-channel input encoder is a traditional method to process the multi-modality data. This architecture places different modalities as initial inputs on different channels for training.

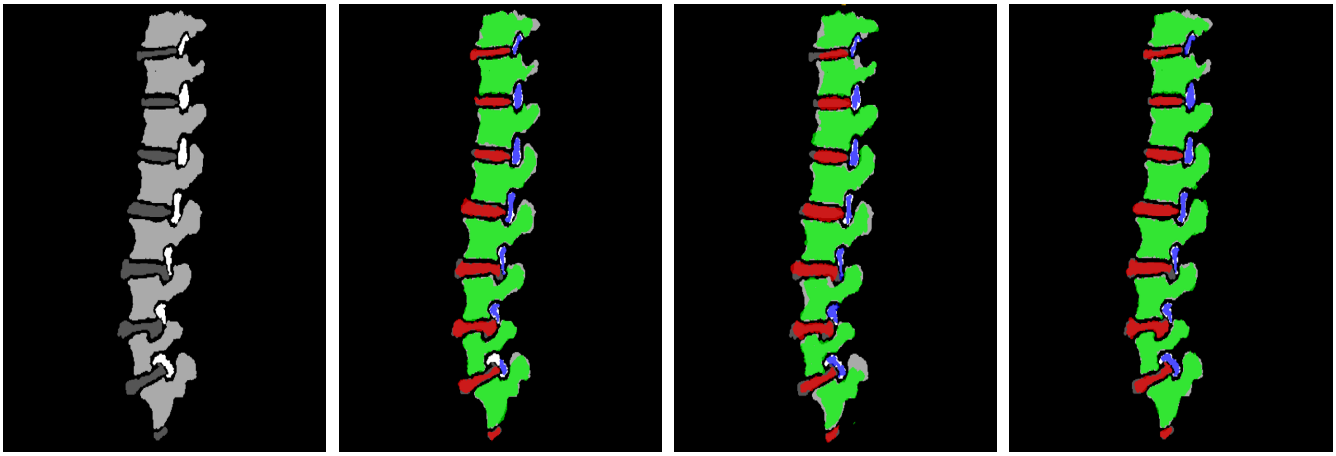
1) MODULES ANALYSIS BY INTRA-COMPARISON

To investigate the effectiveness of a multi-stage learning strategy, we compare the segmentation results achieved by the WCEL, FCEL, and MSAL. Table 2 present the results of these experimental configurations. MSAL can distinguish spinal structures much better than directly employing the single stage learning strategy, with the performance of our S3egANet outperforms the MMAE + CMVF + WCEL structure by approximately 3% percent on the Dice coefficient.

By comparison to the segmentation accuracy of each spine structure, the experimental results showed that a multi-stage strategy mainly improves the segmentation accuracy of IVF and achieves simultaneous and efficient segmentation of multiple classes. In addition, as is shown in Fig. 4(a)-(c), although the FCEL mainly focuses on the prediction accuracy of different classes, it still preserves fine details which accurately delineate the spinal structures. Moreover, S3egANet also achieves higher Sensitivity than its ablated versions. As shown in Fig. 4(d), the segmentation network corrected the voxels which are improperly identified after training by using the MSAL. As a result, performance

TABLE 2. Experimental results demonstrated that S^{3}egANet has superior segmentation effectiveness on the Dice coefficient and Sensitivity from inter-comparisons and intra-comparisons.

Method	Dice_{Mean}	Dice_{IVD}	$\text{Dice}_{Vertebrae}$	Dice_{IVF}	Sensitivity
3D FCN	80.93 ± 3.89	84.39 ± 3.68	79.10 ± 4.42	74.40 ± 2.91	76.12
3D U-Net	84.42 ± 2.46	83.90 ± 2.12	83.6 ± 2.86	77.70 ± 2.14	80.24
3D Spine-GAN [10]	85.65 ± 2.21	85.93 ± 2.87	84.61 ± 2.43	82.7 ± 1.76	87.43
Multi-channel input encoder + WCEL	82.30 ± 4.01	80.92 ± 3.15	80.96 ± 2.36	78.27 ± 3.72	77.93
MMAE + CMVF + WCEL	85.37 ± 1.96	86.90 ± 1.72	82.96 ± 2.76	80.80 ± 1.75	91.14
MMAE + CMVF + FCEL	86.11 ± 2.35	87.1 ± 2.55	85.61 ± 2.21	85.61 ± 2.32	88.90
MMAE + CMVF - ConvLSTM + WCEL	80.91 ± 2.55	83.10 ± 2.32	80.52 ± 2.71	74.60 ± 2.49	82.27
MMAE + CMVF - ConvLSTM + FCEL	82.90 ± 2.83	85.27 ± 2.71	84.90 ± 2.95	84.64 ± 2.67	82.50
MMAE + CMVF - ConvLSTM + MSAL	85.29 ± 2.36	90.23 ± 2.56	85.90 ± 2.87	84.71 ± 2.05	85.16
S^{3}egANet (MMAE + CMVF + MSAL)	**88.30 ± 1.64**	**92.59 ± 1.43**	**88.01 ± 1.89**	**86.38 ± 0.91**	**91.45**



(a) Ground Truth (b) WCEL (c) FCEL (d) MSAL

FIGURE 4. Multi-stage adversarial learning achieved reliable performance in the 3D semantic segmentation of intervertebral discs, vertebrae, and intervertebral foramen.

evaluations are significantly improved after using a multi-stage learning strategy.

The MMAE + CMVF architecture in our experiment is superior to the traditional multi-channel input encoder, as shown in Table 2. One possible reason is that extracting features separately for different modalities data makes the network adapt to the characteristics of each modality, and is not easily disturbed by the differences between different modalities. For example, the T2WI image may reduce the segmentation accuracy of the IVD boundary. After extracting the features of different modalities, the CMVF module assigns different weights to each modality and sums the feature values in the feature maps. This allows the features of different modalities to be well integrated.

We also compared the performance of ConvLSTM. As shown in Table 2, CMVF can distinguish spinal structures much better than directly employing the cross-modality convolution layer without ConvLSTM, with the performance outperforms the MMAE + CMVF - ConvLSTM + MSAL 3.01% percent on the Dice coefficient. Our experiments demonstrate that the CMVF, which combines cross-modal convolution and ConvLSTM to simultaneously mix structural features from multi-modality MRI data and extract the sequential dependencies, is the best choice among similar methods.

2) SUPERIOR ANALYSIS BY INTER-COMPARISON

In Table 2, we compare our S^{3}egANet with 3D FCN, 3D U-Net, and 3D Spine-GAN. S^{3}egANet significantly outperforms the 3D FCN by approximately 7.37% percent on the mean Dice coefficient. We also compare our method with 3D U-Net [31] for spinal structures segmentation, one of the most known frameworks in the medical image community. S^{3}egANet outperforms the 3D U-Net network by 3.88% mean Dice coefficient. Spine-GAN is the most related to our work. We built its 3D version according to the structure of Spine-GAN, which is the 3D Spine-GAN. It is observed that the results achieved by S^{3}egANet have higher segmentation accuracy than 3D Spine-GAN, with mean Dice coefficient of 2.65% improvement. Therefore, S^{3}egANet enjoys a strong superiority of segmentation performance of 3D spinal structures.

C. DISCUSSION

S^{3}egANet is an integrated adversarial net with a multi-stage learning strategy for 3D multiple spinal structures segmentation. Extensive experimental results demonstrate that our method is effective and achieves state-of-the-art Dice coefficient accuracy. Our solution provides a possibility for radiologists to solve the time-consuming, laborious and error-prone problems of manual labeling in the current clinical routine.

The limitation of S3egANet have mainly two aspects: 1) S3egANet not develops an effective method to acquire 3D MRI data with a high resolution on the third dimension. More detailed structure can be delineated on the third dimension with more slices than we collect now available, which would also help improve performance. In the future, we shall investigate how to utilize some image generation techniques to acquire smoother MRI images and further improve the performance. 2) S3egANet does not develop a reliable method to achieve accurate registration between different modalities data. Each set of multi-modality MRI images we used was derived from two consecutive scans in a patient's clinical diagnosis. The visible misalignment images have been manually removed. The process of data selection affects the clinical application of S3egANet. In the future, we can find more available data by finding a reliable registration method and achieve more accurate segmentation results.

V. CONCLUSION

In this study, S3egANet was developed and validated as a relatively complete solution for the simultaneous 3D semantic segmentation of multiple spinal structures at the voxel level. It combines the advantages of the powerful multi-stage adversarial learning strategy to achieve highly reliable and accurate segmentation of multiple spinal structures and leverages the cross-modality voxel fusion module to further effectively integrate the multi-modality information and improve the learning capability. Extensive experiments on MRI images of 90 patients have demonstrated the effectiveness of our S3egANet. S3egANet improves the existing work to the level of 3D segmentation with high accuracy and reliability for multiple spinal structures. S3egANet lays a good foundation for computer-aided automatic diagnosis of spinal diseases and can be applied to other organss' semantic segmentation.

REFERENCES

[1] H. S. An, P. A. Anderson, V. M. Haughton, J. C. Iatridis, J. D. Kang, J. C. Lotz, R. N. Natarajan, T. R. Oegema, Jr., P. Roughley, and L. A. Setton, "Introduction: Disc degeneration: Summary," *Spine*, vol. 29, no. 23, pp. 2677–2678, 2004.

[2] X. Li, Q. Dou, H. Chen, C.-W. Fu, X. Qi, D. L. Belavý, G. Armbrecht, D. Felsenberg, G. Zheng, and P.-A. Heng, "3D multi-scale FCN with random modality voxel dropout learning for intervertebral disc localization and segmentation from multi-modality MR images," *Med. Image Anal.*, vol. 45, pp. 41–54, Apr. 2018.

[3] L. G. Jenis and H. S. An, "Spine update: Lumbar foraminal stenosis," *Spine*, vol. 25, no. 3, pp. 389–394, Feb. 2000.

[4] J. F. Audette, E. Emenike, and A. L. Meleger, "Neuropathic low back pain," *Current Pain Headache Rep.*, vol. 9, no. 3, pp. 168–177, 2005.

[5] P. Gerdhem, "Osteoporosis and fragility fractures: Vertebral fractures," *Best Pract. Res. Clin. Rheumatol.*, vol. 27, no. 6, pp. 743–755, Dec. 2013.

[6] P. Violas, E. Estivalezes, J. Briot, J. S. de Gauzy, and P. Swider, "Objective quantification of intervertebral disc volume properties using MRI in idiopathic scoliosis surgery," *Magn. Reson. Imag.*, vol. 25, no. 3, pp. 386–391, 2007.

[7] R. Niemeläinen, T. Videman, S. Dhillon, and M. Battié, "Quantitative measurement of intervertebral disc signal using MRI," *Clin. Radiol.*, vol. 63, no. 3, pp. 252–255, Mar. 2008.

[8] S. Lee, J. W. Lee, J. S. Yeom, K.-J. Kim, H.-J. Kim, S. K. Chung, and H. S. Kang, "A practical MRI grading system for lumbar foraminal stenosis," *Amer. J. Roentgenol.*, vol. 194, no. 4, pp. 1095–1098, Apr. 2010.

[9] T. M. Emch and M. T. Modic, "Imaging of lumbar degenerative disk disease: History and current state," *Skeletal Radiol.*, vol. 40, no. 9, pp. 1175–1189, Sep. 2011.

[10] Z. Han, B. Wei, A. Mercado, S. Leung, and S. Li, "Spine–GAN: Semantic segmentation of multiple spinal structures," *Med. Image Anal.*, vol. 50, pp. 23–35, Dec. 2018.

[11] C. Chevrefils, F. Cheriet, C.-É. Aubin, and G. Grimard, "Texture analysis for automatic segmentation of intervertebral disks of scoliotic spines from MR images," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 4, pp. 608–620, Jul. 2009.

[12] T. Klinder, R. Wolz, C. Lorenz, A. Franz, and J. Ostermann, "Spine segmentation using articulated shape models," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2008, pp. 227–234.

[13] B. Michael Kelm, M. Wels, S. Kevin Zhou, S. Seifert, M. Suehling, Y. Zheng, and D. Comaniciu, "Spine detection in CT and MR using iterated marginal space learning," *Med. Image Anal.*, vol. 17, no. 8, pp. 1283–1292, Dec. 2013.

[14] C. Chen, D. Belavy, W. Yu, C. Chu, G. Armbrecht, M. Bansmann, D. Felsenberg, and G. Zheng, "Localization and segmentation of 3D intervertebral discs in MR images by data driven estimation," *IEEE Trans. Med. Imag.*, vol. 34, no. 8, pp. 1719–1729, Aug. 2015.

[15] Z. Wang, X. Zhen, K. Tay, S. Osman, W. Romano, and S. Li, "Deep regression segmentation for cardiac bi-ventricle MR images," *IEEE Trans. Med. Imag.*, vol. 34, no. 8, pp. 1640–1648, Jan. 2015.

[16] J. Chmelik, R. Jakubicek, P. Walek, J. Jan, P. Ourednicek, L. Lambert, E. Amadori, and G. Gavelli, "Deep convolutional neural network-based segmentation and classification of difficult to define metastatic spinal lesions in 3D CT data," *Med. Image Anal.*, vol. 49, pp. 76–88, Oct. 2018.

[17] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," 2016, 2016, *arXiv:1611.08408*. [Online]. Available: https://arxiv.org/abs/1611.08408

[18] S. Kohl, D. Bonekamp, H.-P. Schlemmer, K. Yaqubi, M. Hohenfellner, B. Hadaschik, J.-P. Radtke, and K. Maier-Hein, "Adversarial networks for the detection of aggressive prostate cancer," 2017, *arXiv:1702.08014*. [Online]. Available: https://arxiv.org/abs/1702.08014

[19] P. Moeskops, M. Veta, M. W. Lafarge, K. A. J. Eppenhof, and J. P. W. Pluim, "Adversarial training and dilated convolutions for brain MRI segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2017, pp. 56–64.

[20] C. Xu, L. Xu, G. Brahm, H. Zhang, and S. Li, "Mutgan: Simultaneous segmentation and quantification of myocardial infarction without contrast agents via joint adversarial learning," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2018, pp. 525–534.

[21] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[22] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, *Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction*, vol. 6791. Berlin, Germany: Springer, 2011, ch. 1, pp. 52–59.

[23] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2528–2535.

[24] 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

[25] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.

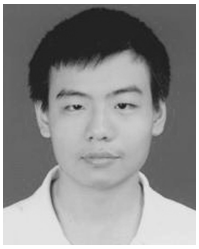
[26] K.-L. Tseng, Y.-L. Lin, W. Hsu, and C.-Y. Huang, "Joint sequence learning and cross–modality convolution for 3D biomedical segmentation," in *Proc. 2017 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3739–3746.

[27] 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.*, Oct. 2017, pp. 2980–2988.

[28] T. Tieleman and G. Hinton, "Lecture 6.5-RmsProp: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Netw. Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.

[29] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. Artif. Intell.*, Montreal, QC, Canada, 1995, vol. 14, no. 2, pp. 1137–1145.

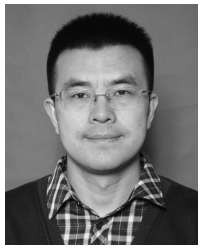
[30] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, and M. Isard, "Tensorflow: A system for large-scale machine learning," in *Proc. OSDI*, vol. 16, 2016, pp. 265–283.
[31] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2016, pp. 424–432.



TIANYANG LI is currently pursuing the master's degree with the Shandong University of Traditional Chinese Medicine. His major research interests include machine learning and medical image analysis.

BENZHENG WEI received the B.S. degree in computer science from the School of Computer Science, Shandong Institute of Light Industry, Jinan, China, in 2000, the M.S. degree in computer science from the School of Computer Science and Technology, Shandong University, Jinan, in 2007, and the Ph.D. degree in precision instrument and machinery from the College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2013. He is a Professor with the Shandong University of Traditional Chinese Medicine. He is also the Director of the Center for Medical Artificial Intelligence and the Computational Medicine Laboratory, Shandong University of Traditional Chinese Medicine. He has published over 80 articles in refereed international leading journals/conferences such as *Medical Image Analysis*, *Neuroinformatics*, and *Neurocomputing*. He has published over 80 articles in refereed international leading conferences such as MICCAI. His current research interests include artificial intelligence, medical image analysis, and computational medicine.

JINYU CONG received the B.E. and M.E. degrees from the Shandong University of Traditional Chinese Medicine, Jinan, China. She is currently pursuing the Ph.D. degree with Shandong Normal University, Jinan. Her research interests include the medical image processing and machine learning.



XUZHOU LI received the B.S. degree in computer science from the School of Computer Science, Shandong Institute of Light Industry, Jinan, China, in 2002, and the M.S. and Ph.D. degrees in computer science from the School of Computer Science and Technology, Shandong University, Jinan, in 2006 and 2018, respectively. He is an Associate Professor with the Shandong Youth University of Political Science. His current research interests include artificial intelligence, medical image analysis, and computational medicine.

SHUO LI received the Ph.D. degree in computer science from Concordia University, Montreal, QC, Canada in 2006. He is a Research Scientist and the Project Manager of GE Healthcare, London, ON, Canada. He is an Adjunct Research Professor with Western University and an Adjunct Scientist with the Lawson Health Research Institute. He is also the Scientific Director of the Digital Imaging Group, London. His current research interests include medial image analysis, with a main focus on automated medial image analysis and visualization.