Hindawi
Computational Intelligence and Neuroscience
Volume 2022, Article ID 4103524, 11 pages
https://doi.org/10.1155/2022/4103524

Research Article

A Novel U-Net Based Deep Learning Method for 3D Cardiovascular MRI Segmentation

Yinan Lu,[1] **Yan Zhao**,[1] **Xing Chen**,[2] **and Xiaoxin Guo**[1]

[1]*College of Computer Science and Technology, Jilin University, Changchun 130000, China*
[2]*College of Artificial Intelligence, Jilin University, Changchun 130000, China*

Correspondence should be addressed to Yinan Lu; luyn@jlu.edu.cn

Received 9 December 2021; Revised 25 January 2022; Accepted 16 February 2022; Published 20 May 2022

Academic Editor: Daqing Gong

Medical multiobjective image segmentation aims to group pixels to form multiple regions based on the different properties of the medical images. Segmenting the 3D cardiovascular magnetic resonance (CMR) images is still a challenging task owing to several reasons, including individual differences in heart shapes, varying signal intensities, and differences in data signal-to-noise ratios. This paper proposes a novel and efficient U-Net-based 3D sparse convolutional network named SparseVoxNet. In this network, there are direct connections between any two layers with the same feature-map size, and the number of connections is reduced. Therefore, the SparseVoxNet can effectively cope with the optimization problem of gradients vanishing when training a 3D deep neural network model on small sample data by significantly decreasing the network depth, and achieveing better feature representation using a spatial self-attention mechanism finally. The proposed method in this paper has been thoroughly evaluated on the HVSMR 2016 dataset. Compared with other methods, the method achieves better performance.

1. Introduction

In multiobjective segmentation, medical image segmentation aims to segment the images into multiple regions and extract the parts of interest based on the similar characteristics or single attributes of the image, such as edge contour, structure, and shape, which is of great significance for medical image analysis, disease diagnosis, and clinical applications (e.g., 3D computed tomography (CT) and magnetic resonance image (MRI)). Accurate segmentation can not only help precise diagnosis and prediction of prognosis but also benefit surgical planning and intraoperative guidance.

For example, in diagnosing congenital heart disease, segmenting the blood pool and myocardium from 3D cardiovascular magnetic resonance (CMR) images is a prerequisite before creating patient-specific heart models for preprocedural planning of children with complex congenital heart disease (CHD).

One of the main applications of deep learning is medical field, including biomedicine and MRI analysis [1]. However, the level of doctors is uneven, and some departments are labor-intensive, which has had a profound impact on the development of artificial intelligence in this field [2]. Currently, segmenting vital organs or structures from 3D medical images is an imperative preliminary action for a wide range of clinical treatments. The recognized standard segmentation results are obtained from experienced physicians and radiologists via visual inspection and manual delineations. On the one hand, there always are hundreds of images in an individual's cardiac MRI. It is tedious, time-consuming, and costly to annotate the 3D medical images in a slice-by-slice manner. On the other hand, the whole heart's manual labeling is subjective and suffers from low reproducibility. The results of the labeling could be seriously affected by the experience and knowledge of the observer. Consequently, automatic medical image segmentation with high accuracy is highly demanded. However, using deep learning for automatic medical image segmentation with high accuracy is also a huge challenge [3]. The reasons include (I) the missing borders or indefinite boundaries, with inadequate edge information, and (II) the too low quality of the cardiac images.

There are many deep learning models applied to image segmentation [4]. Convolutional neural network (CNN) based deep learning strategies especially achieved remarkable success in medical image segmentation methods. U-Net [5] is a semantic segmentation network based on fully convolutional networks (FCN) [6], which mainly applied CNN structure for the heart segmentation. U-Net can be quickly trained on small sample data in medical segmentation by data augmentation and achieve outstanding segmentation results. Different from the FCN structure, there is not any encoder or decoder in the U-Net. U-Net contains two paths, the downsampling contraction path, which extracts the high-level abstract features of pixels, and the extended upsampling path, which can reconstruct pixel information lost during downsampling. In the process adopted above, the parts from the comprehensive upsampling approach and the features extracted by downsampling are stitched to maximize the retention of low-level feature information lost by the pooling and convolution operations. Compared with FCN, U-Net can run more efficiently because there is no fully connected layer in the structure. The paper on DenseNet [7] was voted the best paper of CVPR in 2017, which has the same basic idea as ResNet [8]. However, it establishes the dense connection between all the previous layers and the latter layers. The dense block in the DenseNet is a densely connected network model between layers. In each dense block, the input of each layer is the union of the outputs of all the previous layers. DenseNet enhances feature representation with skip connections.However, the feature maps in DenseNet are relatively large, resulting in a large amount of computation in the convolution process, which affects the overall performance of the network.

This paper proposes a novel and efficient 3D sparse convolutional network named SparseVoxNet to comprehensively address these challenges, which can effectively carry out voxel-to-voxel learning and infer 3D medical images. Specifically, we develop a sparse convolutional network that aims to contribute the following ideas:

(1) The sparse network can eliminate redundant computation, reduce model parameters, and decrease the risk of overfitting small sample training data. (2) The full skip connection mechanism in the module can effectively solve the problem of gradient disappearance in 3D deep model training, accelerate the convergence speed, and improve recognition ability. (3) The self-attention mechanism is added to optimize the expression ability of feature maps and capture the long-range dependency between features better.

2. Related Work

Multiobjective image segmentation can be divided into supervised and unsupervised methods. Pham et al. [9] proposed a multiobjective optimization approach to segment the brain MRI using fuzzy entropy clustering and region-based active contour methods. Hongwei et al. [10] proposed a multiobjective clustering and toroidal model-guided tracking method to distinguish vascular structures from complicated structures in background regions. In recent years, deep learning has been successfully applied to medical image segmentation. Çiçek et al. [11] proposed a 3D U-Net network structure to realize the 3D image segmentation. Habijan et al. [12] proposed a framework consisting of two 3D U-Nets. In this framework, the first network was used for localizing the bounding box encompassing the heart, and the second network was employed to segment the different substructures. Ding et al. [13] incorporated attention mechanism within the gradient expanding process to enhance the coarse segmentation information with less computation expense. Furthermore, they extended the network's gradient flow and used the low-resolution feature information. Jeevakala et al. [14] proposed a Mask R-CNN approach driven with U-Net to detect and segment the Internal Auditory Canal (IAC) and its nerves. In this method, the U-Net segmented the structure related information of IAC and its nerves by learning its features.

However, the variants' structure of U-Net suffers from redundant information. More and more network structures have been proposed and applied to image analysis [15].

Fisher and Koltun [16] proposed a new convolutional network module which used dilated convolutions. This module could aggregate multiscale contextual information systematically without losing resolution. Recently, dilated convolution is increasingly applied to medical images. Wolterink et al. [17] proposed a method to segment the myocardium and blood automatically in CMR of patient who has CHD by CNN. In the same year, Fisher et al. [18] developed a convolutional network module specifically for intensive prediction which used extended convolution to systematically aggregate multiscale context information without loss of resolution. Residual network (ResNet) was proposed in 2016, which added skip connections to each convolution layer for 2D image classification tasks. In addition, this architecture has been extended to 3D volumetric segmentation [19–21]. Huang et al. proposed the DenseNet with $L(L+1)/2$ direct connections, which improved ResNet. It can strengthen feature propagation and reuse all features. After this improvement, Jégou [22] proposed a 2D fully convolutional DenseNet for semantic segmentation. In the same year, Yu et al. [23] proposed the DenseVoxNet; this network extended the deep residual learning in 2D image recognition tasks into 3D, which could simplify network training, reduce the parameters, and add auxiliary paths to enhance gradient propagation. However, there were no direct connections between the dense blocks and the final prediction layer. DenseVoxNet may not be able to appropriately capture multiscale contextual information useful for accurate segmentation. The correlation of adjacent images or frames should be effectively exploited for improving the accuracy of the target tasks which involves 3D volumetric data. Therefore, more and more methods have been proposed to use 3D features for biomedical volumetric data [24–29]; for example, Hosseini-Asl et al. [30] proposed a deep supervised adaptive 3D CNN, which could automatically extract and recognize the characteristics of Alzheimer's disease and capture the changes caused by Alzheimer's disease, such as the size of ventricle, the shape of the hippocampus, and the thickness of cortex. Dou et al. [31]

proposed a 3D fully convolutional network, called 3D Deeply Supervised Network (DSN), equipped with a deep supervision mechanism. This method has obtained good results in two tasks: liver segmentation of 3D CT scan, and whole heart and large blood vessels segmentation of 3D MRI. Previous CNN expresses dependencies between different image regions through convolution. Convolution operators have local receptive fields, so processing long-range dependencies goes through multiple convolutional layers, which may prevent learning about long-term dependencies. While it is possible to increase the representational capacity of the network by increasing the size of the convolutional kernels, the computational and statistical efficiency gained by using local convolutional structures are lost. However, self-attention [32–34] can exhibit a better balance between the ability to model long-range dependencies and computational efficiency. However, it is still a challenging task for CNNs to segment the important organs from 3D medical images due to the complexity of 3D structures, the difficulty of voxelized grid optimization, and the insufficiency in training samples.

Dou et al. [35] proposed Pnp-AdaNet using the method of adversarial learning, which could adapt to medical images of different modalities through plug-and-play modules. In another experiment, Dou et al. [36] constructed a domain adaptation module (DAM) to map the target region to features that were spatially aligned with the source domain region. The domain critic module(DCM) was responsible for distinguishing the feature spaces of the two domains. Then these two modules were optimized via an adversarial loss without using any target domain label. They trained the network using MRI, used it to segment CT images, and finally achieved certain results. The experiments done by Schlemper et al. [37] showed that using a grid-like attention mechanism in CT images might achieve better results. Shi et al. [38] proposed Bayesian VoxDRN for segmenting the entire heart from 3D MRI. Bayesian VoxDRN could predict voxel class labels by measuring the uncertainty of the model. During the test, it was realized by sampling based on Monte Carlo to generate a posteriori distribution of voxel labels. The attention mechanism was first applied to the text field. When the improved attention mechanism was applied to image processing, very good results were achieved. Liu et al. [39] proposed a novel medical image super-resolution method based on dense neural network and blended attention mechanism to address the problem that medical image would suffer from severe blurring caused by the lack of high-frequency details in the process of image super-resolution reconstruction. Kaul et al. [40] joined the attention tool to CNNs using feature maps generated by a separate convolutional autoencoder. This attention architecture was well suited for incorporation into deep convolutional networks. The results showed that this attention architecture was better than U-Net and residual variant.

3. Methods

3.1. The Architecture of SparseVoxNet. The architecture of SparseVoxNet proposed in this paper is shown in Figure 1. It improves U-Net which includes upsampling and downsampling processes to implement end-to-end training. The padding is used for keeping the feature-map sizes constant in every sparse block, because the sparse block is not applicable when the feature maps have different sizes. Therefore, in each sparse block, the first 4 layers use ordinary convolutions, and the last 3 layers use dilated convolutions. The hole sizes are 2, 3, and 5. The spatial self-attention mechanism is added after the original feature map of data to strengthen the more important features in the original feature map. In the final deconvolution layer, instead of using a fully connected layer, three $1 \times 1 \times 1$ convolution layers and softmax layer are used to obtain the segmented final label map. A dropout layer with a coefficient of 0.2 is added after each convolution layer to enhance the generalization ability of network.

Inspired by DenseNet, the black dotted line in SparseVoxNet in Figure 1 represents a skip connection. The image is segmented once by deconvolution on the skip connection. The network will converge faster and the accuracy rate will be higher due to the skip connection. The first segmented image will perform better on edge segmentation, because the shallow neural network loses less information through convolution and gets more edge information. The result is a fine grained segmentation. The result of the second segmented image is better in overall segmentation, which is coarse grained segmentation. Deep neural network features are high-level abstract features, which is really helpful when extracting the segmented central area of the entire tissue. The final segmentation result is determined by the voting of multiple segmentation results of different cropped input data on a single voxel point. The downsampling process of U-Net is replaced with sparse blocks, and the two deconvolutions are equivalent to the upsampling process.

Furthermore, we calculate the number of parameters for each layer in the SparseVoxNet shown in Table 1. Table 1 shows the parameters of 4 convolution layers, 2 deconvolution layers, 2 sparse blocks, a spatial attention mechanism layer, and a skip connection layer. Among them, the 4 convolution layers are represented by Conv_n, the 2 deconvolution layers are represented by Deconv_n, and the 2 sparse blocks are represented by Sparse Block_n. We also show the convolution kernel and stride of each layer in Table 1. Note that each row in Table 1 corresponds to each layer in Figure 1.

3.2. Sparse Block. DenseNet has denser connections compared to ResNet, which makes the consumption of hardware resources very high. Therefore, we propose a sparse network structure to change the way of feature reuse while keeping feature reuse and skip connection characteristics unchanged. The sparse block which we propose reduces the number of connections, just having direct connections between any two layers with the same feature-map size, referred to as full skip connection, but the effect of sparse block is similar to dense block. The input of transition layer is as follows:

$$[T_0, T_1, T_2, T_3, T_4] = [T_0, H_1(T_0), H_2(T_1), H_3(T_2), H_4(T_3)]. \quad (1)$$

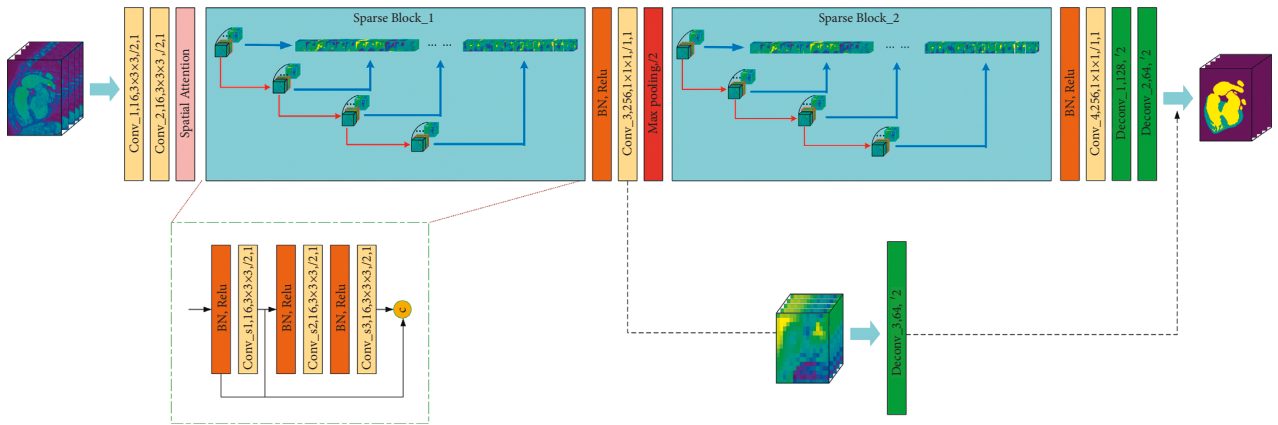

Figure 1: An overview of the proposed SparseVoxNet, with intermediate feature volumes. The light blue and dark blue areas of the slice represent the blood pool and myocardium. The dark blue and black areas belong to the background. The blue, yellow, and dark purple of segmented result represent the myocardium, blood pool, and background, respectively. There are two sparse blocks in this network. The black dotted line at the bottom right represents a skip connection.

Table 1: Our 3D convolutional model.

Input image				Output				Layer (type)	Stride	Kernel	Parameters
64	64	64	1	32	32	32	16	Conv_1 (convolution)	2	3	448
32	32	32	16	16	16	16	16	Conv_2 (convolution)	2	3	6928
16	16	16	16	16	16	16	16	Spatial attention	2	1	816
16	16	16	16	16	16	16	100	Sparse Block_1 (sparse block)	1	3	43300
16	16	16	100	16	16	16	100	Conv_3 (convolution)	1	1	10100
16	16	16	100	16	16	16	184	Sparse Block_2 (sparse block)	1	3	496984
16	16	16	184	16	16	16	64	Conv_4 (convolution)	1	1	11840
16	16	16	64	32	32	32	64	Deconv_1 (deconvolution)	2	4	262208
32	32	32	64	64	64	64	64	Deconv_2 (deconvolution)	2	4	262208
16	16	16	100	64	64	64	64	Skip connection	1	1	6464

where the input of H_1 is T_0, the input of H_2 is $T_0 + T_1$, and so on.

The feature maps of different receptive fields are referred to as different scales. It is found that the nonlinear combination of the features of different scales is not better than the linear combination. Inspired by the U-Net network structure, composite expression features are constructed by directly stacking feature maps of different scales.

DeletedUnlike U-Net, the improved network structure uses deconvolution to replace the upsampling process, which reduces the loss of information during the conversion process. In DenseNet, the network connections of the previous layer and the latter layer are too dense, which can easily cause overfitting. The sparse network can solve this problem. The network's feature expression ability is greatly enhanced, and there is no vanishing gradient.

3.3. 3D Dilated Convolution. Dilated convolution has one more hyperparameter than traditional convolution, called dilation rate. Dilated convolution adds holes to the standard convolution kernel. In this paper, we extend the dilated convolution to 3D data, and mix the traditional convolution and dilated convolution. Referring to the DenseVoxNet, we use 4 layers of $3 \times 3 \times 3$ traditional convolution and 3 layers of dilated convolution with 2, 3, and 5 holes. The 4 layers of traditional convolution can extract the local features of the image, and the 3 layers of dilated convolution expand the reception field of the feature exponentially to capture the potential relationship between long distance features. We only use 7-layer convolution to make the reception field reach $26 \times 26 \times 26$.

3.4. Spatial Self-Attention Mechanism. In both the computer vision tasks and the natural language processing tasks, the dependencies between long distance features are difficult to capture. In serialization tasks, recurrent and recursive neural networks are major means to capture long-range dependencies. In convolutional neural networks, large reception fields are formed by superposing multiple convolution operations. Currently, there are no specific methods to capture long-range features. Convolution and cyclic operators have the following disadvantages: (1) being too inefficient, (2) easily producing gradient disappearance, (3) difficulty of passing information back and forth between long ranges.

Inspired by the nonlocal mean filtering for images, Wang et al. [41] proposed nonlocal block for capturing long-range dependencies, which is a self-attention mechanism.

Nonlocal block ignores the Euclidean distance and calculates the relationship between two positions directly. Actually, it calculates the generalized autocorrelation matrix of features. However, the calculation efficiency is relatively high. Because after adding nonlocal operators, it is not necessary to stack too deep convolution operations for achieving the network's fitting ability. Furthermore, it does not change the size of input data and can be easily embedded in the network, so the spatial self-attention model is added in front of the first sparse block. We apply the self-attention mechanism, proposed by Zhang et al. [42], in this paper. The nonlocal block is embedded in the 3D network, which is defined as follows:

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f\left(x_i, x_j\right) g\left(x_j\right), \tag{2}$$

where i is a 3D coordinate meaning the position index of input data, j is the index of all possible positions, x is the input data, f is an autocorrelation calculation function, which can calculate the correlation between i-th position and j-th position, and g is a unary mapping function. The $1 \times 1 \times 1$ convolution is used for ascending dimension and fusing the multichannel feature in the experiments, and finally $C(x)$ is used for normalization. Using multiple $1 \times 1 \times 1$ convolution kernels in the attention model can not only achieve cross-channel interaction and information integration, but also reduce or increase the number of channels. People begin to pay attention to the $1 \times 1 \times 1$ convolution because of the network structure proposed by Lin [43]; this convolution connects two full connection layers for fusing the features linearly. After that, in Google's Inception-v4 [44] network structure, $1 \times 1 \times 1$ convolution is used in the inception module for dimensionality reduction or ascending dimension. Inspired by this advantage, in this paper, the $1 \times 1 \times 1$ convolution kernel is used to reduce the original input data dimensionality, calculate the spatial autocorrelation relationship, and then ascend the dimension of data. The different weights calculated are added back to the original data and then regularized to describe the influence on features of voxel points in different spatial positions.

4. Experiments and Results

4.1. Dataset. Radiobiological images mainly have six data formats. The NIFTI (Neuroimaging Informatics Technology Initiative) is one of them. The data format used in this paper is NIFTI. This format contains two affine coordinates, so that it can associate the physical index of voxels with its actual spatial location. The HVSMR 2016 dataset is used to evaluate the algorithm and network structure. HVSMR 2016 has a total of 10 cardiac magnetic resonance 3D scans for training and 10 scans for testing. All training sets of cardiac MRIs are from patients with CHD, including annotations of myocardium and large blood vessels.

Due to the large difference in intensity between different images, the cardiac MR images are all normalized. After normalization, the mean and unit variance are 0. To leverage the limited training data, simple data augmentation was employed to enlarge the training data. The augmentation operations include the rotation and cropping. The original training set is divided into three parts, namely, the training set, the validation set, and the testing set. The cross-validation method is used for parameter training. We use 70% of the images for training and 30% for testing. Then, we compare and briefly discuss the experimental results.

4.2. Evaluation Metrics. Medical image segmentation is an important step of medical image processing. However, it is difficult to select accurate evaluation index to evaluate the quality of segmentation by comparing segmented medical images. The following three metrics are used in this paper for measuring the results of segmentation.

4.2.1. Dice Coefficient. Dice coefficient is widely used for verifying the effect of 3D medical image segmentation. The core idea is to ensure a high recall and precision. Compared with the evaluation method of directly computing the difference between the automatic segmentation results and the original data labels, using Dice coefficient can better characterize the segmentation effect. Dice coefficient is defined as follows:

$$\text{Dice} = \frac{2|G \cap R|}{|R| + |G|} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}, \tag{3}$$

where G is the segmentation result of ground truth, which is the labeled testing data. R is the automatic segmentation result of testing data. TP, FP, and FN represent true positives, false positives, and false negatives, respectively, for each class. Ideally, the template of segmentation result and the template of label data completely overlap, which means $R = G$, and the absolute value of the Dice coefficient is 1.

4.2.2. Average Symmetric Surface Distance. Average symmetric surface distance (ADB) is defined as follows: for a single voxel point, if one or more voxel points within its 18-neighborhood are not elements of the object, they are regarded as surface voxel points. For each surface voxel point in the R, we calculate the Euclidean distance between it and the nearest surface voxel point of the real label G. Similarly, perform the same calculation for each surface voxel point in the G. $S(R)$ represents the surface voxel point set of R. The distance from any voxel point v to $S(R)$ is defined as $d(v, S(R)) = \min_{s_R \in S(R)} ||v - s_R||$. Based on this formula, the average symmetrical surface distance is defined as follows:

$$\text{ADB} = \frac{1}{|S(G)| + |S(R)|} \times \left[\sum_{s_R \in S(R)} d\left(s_R, S(G)\right) + \sum_{s_G \in S(G)} d\left(s_G, S(R)\right) \right]. \tag{4}$$

Many segmentation boundary evaluation metrics are constructed based on this distance formula, which measures the boundary difference between the segmentation result and the ground truth by calculating the voxel surface distance. The larger the value of ADB, the more dissimilar the segmentation boundary is. When the boundary of the segmentation result matches the ground truth exactly, the value of ADB is 0.

4.2.3. Hausdorff Distance. Based on the ADB, when using the maximum symmetric distance, the metric is known as Hausdorff distance, which is defined as follows:

$$\text{hausdorff}(R, G) = \max_{r \in R} \left\{ \min_{g \in G} \{d(r, g)\} \right\}, \quad (5)$$

where $d(r, g)$ represents the distance between points r and g; that is, the set consists of the shortest distance (usually expressed in Euclidean distance) from all points in the predicted segmentation set R to any point in the real label set G, and the maximum distance is selected from this set as the Hausdorff distance between the two sets R and G. This distance and the symmetrical surface distance both describe the similarity of the contour. The larger the absolute value, the less similar the segmentation.

4.3. Training. In the experiments, all weights are randomly initialized by the Gaussian distribution with $\mu = 0$, $\theta = 0.01$, and the stochastic gradient descent optimization algorithm is used. Batch size is set to 8. In order to reduce the model overfitting and speed up the convergence rate, the weight attenuation is 0.0005, and the momentum is set to 0.9, which is often used to speed up training, while making it easier to jump out of extreme points and avoid getting stuck in local optimal solutions. The drop rate is 0.2, and the initial learning rate is set to 0.01. If the learning rate is too low, the training period is too long, and the high learning rate will cause the model to be unstable and never converge. Our algorithms were trained and tested on the Dual RTX 2080 Ti GPU.

The polynomial decaying learning rate is used for ensuring the rapid convergence of the model during the initial training period and the stability of the model parameters in the later period. The initial learning rate is set relatively large, and the learning rate is reinitialized and decayed every 5000 steps. The attenuation coefficient of the learning rate is $\delta = (1 - \text{iter}/\text{max\_iter})^{\text{power}}$. After testing, the model stabilizes after 8000 iterations. The input data of SparseVoxNet consists of 8 groups of $64 \times 64 \times 64$ heart MRIs, which are cropped randomly in the same axis direction.

Multiple sets of comparative experiments and ablation experiments are designed to verify the effect of the improved method on segmentation. In the experiments, we compared our method with the traditional methods and other deep learning methods, and also compared the network only with the mixed dilated convolution and the network only with the attention mechanism and DenseVoxNet.

4.3.1. Ablation Study. We conduct ablation experiments to verify the importance of 3D dilated convolution and spatial self-attention mechanism in exploiting multiscale features. The results are presented in Table 2.

When we just add the mixed 3D dilated convolution to the model, we define this model as SparseVoxNet-D. The Dice coefficient of myocardium and blood pool gets the best results, 82.4% and 91.6%, respectively. It verifies our conjecture: dilated convolutions can exponentially expand receptive fields to obtain multiscale information without losing resolution or coverage, especially for structures with a small size or irregular boundary, such as the cardiac myocardium structures. Since the receptive field expansion speed of the dilated convolution depends on the number of holes in the dilated convolution, although the more holes will contribute a larger receptive field, the pixels in the large receptive field are not necessarily related to the current convolution. In other words, the larger receptive field is not the better. Local perception can better capture local features. Global perception can better capture the relationship characteristics of pixels at different locations. Hence, we mix the 4 layers of traditional convolution and 3 layers of dilated convolution and define the different dilation rates of dilated convolutions to better capture features.

When we just add the spatial self-attention mechanism to the model, we define this model as SparseVoxNet-S. It can be seen that ADB and Hausdorff distance of blood pool and myocardium achieve better performance than DenseVoxNet, the Hausdorff distance of myocardium outperforms DenseVoxNet by around 3.0%, and the Hausdorff distance of blood pool outperforms DenseVoxNet by around 4.8%. This indicates that with the spatial self-attention mechanism, the segmented images have been brought closer to the target domain successfully, because the self-attention in our model is complementary to the convolution for capturing long-range, global-level dependencies occurring in cardiac structure. The advantages of the attention mechanism are as follows: (a) few parameters; (b) fast calculation; (c) capturing long-range features. The problem applied in this paper is a small sample training process, so when the spatial self-attention mechanism is removed, the segmentation result is not ideal, which means the long-range features cannot be extracted efficiently. We use both dilated convolution and spatial self-attention mechanism to capture long-range features, because the method based on dilated convolution obtains information from a small number of surrounding points and cannot form dense context information. The spatial self-attention mechanism makes a single feature in any location perceive the features of all other locations, and can produce more powerful pixel-level representation capabilities. These observations demonstrate that the 3D dilated convolution and the spatial self-attention indeed play a meaningful role in exploiting multiscale features.

4.4. Results. There are segmentation results on three training images shown in Figure 2. These three slices come from different patients. The data whose indexes are 60 in the

Table 2: Results of ablation study. Bold results are the best ones.

	Myocardium			Blood pool		
Method	Dice (%)	ADB	Haus.	Dice (%)	ADB	Haus.
SparseVoxNet-D	**82.4**	0.922	5.385	**91.6**	1.073	7.736
SparseVoxNet-S	80.7	**0.853**	**5.075**	91.4	**0.951**	**5.004**
DenseVoxNet	79.2	0.943	7.175	89.48	0.955	9.608

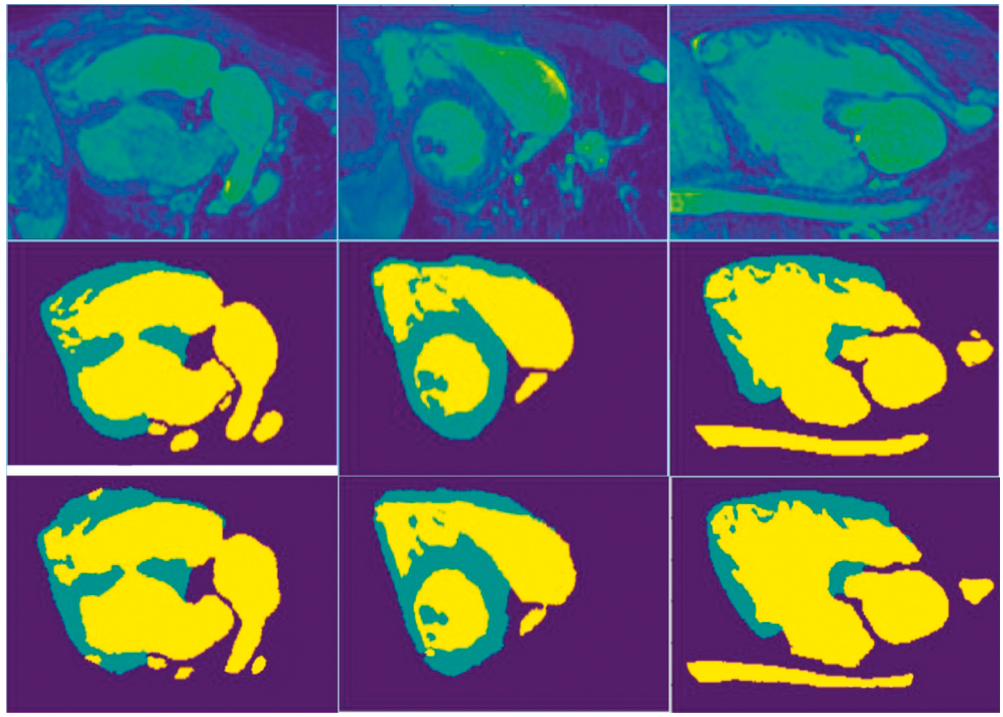

Figure 2: Segmentation results on three training images.

sample dataset have the same coronal plane view in the same dimension. The light blue and dark blue areas of the image in the first line represent the blood pool and myocardium; the dark blue and black areas belong to the background. The images in the second line are labeled, corresponding to the myocardium and blood pool in the first line of images. The third line is the results of automatic segmentation by the method proposed in this paper, where blue, yellow, and dark purple represent the myocardium, blood pool, and background, respectively. It can be seen from Figure 2 that although the cardiac structure of different patients in the training set is quite different, the method we proposed can still successfully calibrate the myocardium and blood pool from low contrast cardiac MRI, which proves that this method has a good enough fitting ability to the original data. However, there are still some disadvantages. In the first auto-segmentation result, the myocardium in the lower left corner is partially divided. In the second result, the background appeared in the myocardium. In the third result, there is extra myocardium in the upper right corner, which shows that deep learning has the ability to perceive most data features, but it does not have reasonable logical reasoning capabilities. Human segmentation will not produce these subtle logical errors.

Figure 3 shows segmentation results on three testing images. The data extraction method is the same as above. By observing the results, we can see that the method proposed in this paper also has good generalization effect on unlabeled data. However, when using the gradient descent algorithm, it is easy to fall into the local optimum and cause overfitting, because of the huge number of parameters.

4.5. Discussion. The comparison of the results between the method we proposed and other six methods is shown in Figure 4. They are mainly ranked according to the Dice coefficient. The figure also shows the auxiliary reference indexes, such as ADB and symmetric Hausdorff distance. The first three are traditional methods, such as manually extracting features and using hidden Markov random fields, and the other deep learning methods are on the HVSMR 2016 Challenge dataset. According to Figure 4, the Dice coefficient of blood pool in all methods is higher than that of myocardium, suggesting that the segmentation of blood pool is

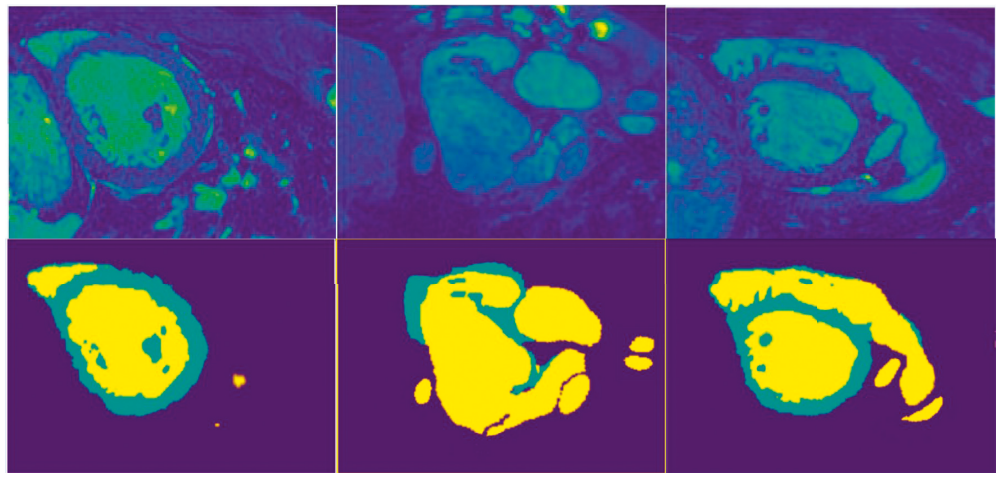

Figure 3: Segmentation results on three testing images: The three slices in the first line come from different patients. Images in the second line are the results of automatic segmentation by the method proposed in this paper.

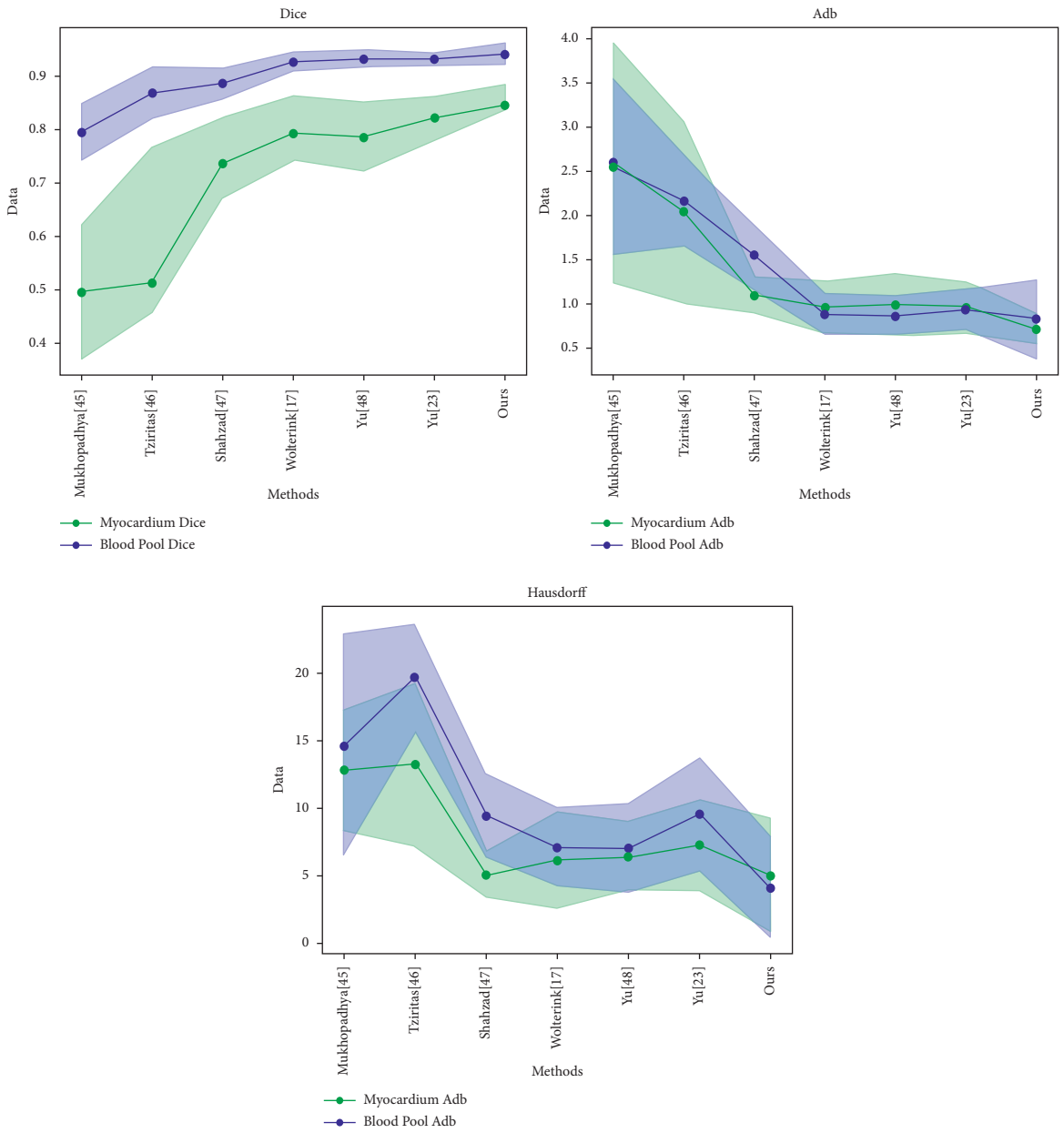

Figure 4: Comparison of experimental results between the improved method and other methods.

Table 3: Comparison of experimental results between the improved method and the 3D methods. Bold results are the best ones.

	Myocardium			Blood pool		
Method	Dice (%)	ADB	Haus.	Dice (%)	ADB	Haus.
3D U-Net [11]	69.4	2.596	12.796	79.4	2.550	14.634
V-Net [26]	70.3	2.367	10.624	81.9	2.435	12.539
VoxResNet [19]	77.4	2.041	13.199	86.7	2.157	19.723
DenseVoxNet [23]	79.2	0.943	7.175	89.48	0.955	9.608
Ours	**84.5**	**0.721**	**5.027**	**94.0**	**0.831**	**4.102**

relatively easier due to the ambiguous borders of the myocardium in the low-resolution MRIs. Regarding the segmentation of myocardium, the method we proposed achieves the best performance with the Dice; i.e., the ranking metric in the challenge, 0.861 ± 0.024, outperforms the second one by around 4%. The best result also has been achieved in blood pool segmentation with Dice; the ranking metric in the challenge, 0.94 ± 0.016, demonstrates that our sparse connected network has the capability to tackle hard cardiovascular segmentation problem. The ADB and Hausdorff distance of our method also achieved the best performance.

The results of other 3D MRI segmentation methods are mainly shown in Table 3. Firstly, the experimental parameters are compared, and the method proposed in this paper needs the least parameters. The sparse block and dilated convolution can achieve a good fitting effect with the participation of such a small number of parameters, thanks to the introduction of the attention model. The feature expression ability of sparse block will not be better than dense block in many cases, but the problem applied in this paper is medical segmentation and a small sample training process, so the sparse block can fit and generalize the data well with a small number of parameters, and the exponentially increasing receptive field provided by the dilated convolution reduces the convolution operations. The attention mechanism can well capture the features to strengthen the generalization ability of the network. Because of the small number of parameters, the amount of calculation is reduced, and the model's convergence rate is also fast.

Comparing the cross-entropy loss of DenseVoxNet and SparseVoxNet with sparse block and dilated convolution, we can find that the network only using the mixed dilated convolution can converge faster and reach lower loss values, which proves that the improved sparsely connected network structure can reduce the calculation amount and improve the efficiency and that the method of extracting long-range features by hybrid 3D dilated convolution is suitable for medical images. It has better ability to represent features and fit data .

The comparison shows that the time of one iteration of DenseVoxNet (forward and backward propagation of the network) is 0.113 s, the time of SparseVoxNet-D is 0.045 s, and the time of SparseVoxNet is 0.049 s. The proposed method has a great improvement in efficiency.

5. Conclusion

In this paper, we propose a novel and efficient 3D sparse convolutional network to segment blood pool and myocardium from 3D cardiac magnetic resonance images. This method can eliminate redundant calculations and reduce model parameters and the risk of overfitting training data on small samples. The spatial self-attention mechanism can optimize the expression ability of feature maps, and the sparse blocks can reduce the convolutional network depth. The work in this paper is an accurate pixel-level classification. Moreover, we achieve competitive results in comparison with existing methods. The proposed method can provide comprehensive information for doctors to make diagnoses of CHD.

Data Availability

The data used in this study could be accessed upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No . 61872162 and 82071995) and Natural Science Foundation of Jilin Province under Grant No. 20200201292JC.

References

[1] M. Bakator and D. Radosav, "Deep learning and medical diagnosis: a review of literature," *Multimodal Technologies and Interaction*, vol. 2, no. 2, p. 47, 2018.

[2] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, 2018.

[3] X. Zhuang, "Challenges and methodologies of fully automatic whole heart segmentation: a review," *Journal of Healthcare Engineering*, vol. 4, no. 3, pp. 371–407, 2015.

[4] W. G. Hatcher and W. Yu, *A Survey of Deep Learning: Platforms, Applications and Emerging Research Trends*, p. 1, IEEE Access, USA, 2018.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, Germany, 2015.

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2014.

[7] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern*

Recognition 14 (CVPR), IEEE Computer Society, Hawaii, USA, 2017.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Nevada, 2016.

[9] T. X. Pham, P. Siarry, and H. Oulhadj, "A multi-objective optimization approach for brain MRI segmentation using fuzzy entropy clustering and region-based active contour methods," *Magnetic Resonance Imaging*, vol. 61, pp. 41–65, 2019.

[10] D. U. Hongwei, K. Shao,, F. Bao et al., "Automated coronary artery tree segmentation in coronary CTA using a multi-objective clustering and toroidal model-guided tracking method," *Computer Methods and Programs in Biomedicine*, vol. 9963, Article ID 105908, 2020.

[11] Ö. Cicek, A. Abdulkadir, S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-net: learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016. MICCAI 2016. Lecture Notes in Computer Science*, S. Ourselin, L. Joskowicz, M. Sabuncu, G. Unal, and W. Wells, Eds., vol. 9901, Berlin, Germany, Springer, 2016.

[12] M. Habijan, H. Leventić, I. Galić, and D. Babin, "Whole heart segmentation from CT images using 3D U-net architecture," in *Proceedings of the International Conference System, Signals Image Process (IWSSIP)*, pp. 121–126, Osijek, Croatia, June 2019.

[13] X. Ding, Y. Peng, C. Shen, and T. Zeng, "Cab U-net: an end-to-end category attention boosting algorithm for segmentation," *Computerized Medical Imaging and Graphics*, vol. 84, Article ID 101764, 2020.

[14] S. Jeevakala, C. Sreelakshmi, K. Ram, R. Rangasami, and M. Sivaprakasam, "Artificial intelligence in detection and segmentation of internal auditory canal and its nerves using deep learning techniques," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, no. 11, pp. 1859–1867, 2020.

[15] G. Litjens, T. Kooi, B. E. Bejnordi et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[16] Y. Fisher and V. Koltun, *Multi-Scale Context Aggregation by Dilated Convolutions*, 2016.

[17] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "Dilated convolutional neural networks for cardiovascular MR segmentation in congenital heart disease," in *Proceedings of the International Workshop on Reconstruction and Analysis of Moving Body Organs International Workshop on Whole-Heart and Great Vessel Segmentation from 3D Cardiovascular MRI in Congenital Heart Disease*, pp. 95–102, Athens, Greece, 2017.

[18] Y. Fisher, V. Koltun, and T. Funkhouser, "Dilated residual network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 472–480, Honolulu, HI, USA, 2017.

[19] H. Chen, Q. Dou, L. Yu, Q. Jing, and H. Pheng-Ann, "VoxResNet: deep voxelwise residual networks for brain segmentation from 3D MR images," *NeuroImage*, vol. 170, pp. 446–455, 2017.

[20] L. Yu, X. Yang, H. Chen, J. Qin, and P. A. Heng, "Volumetric ConvNets with mixed residual connections for automated prostate segmentation from 3D MR images," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 66–72, AAAI Press, 2017.

[21] A. Fakhry, T. Zeng, and S. Ji, "Residual deconvolutional networks for brain electron microscopy image segmentation," *IEEE Transactions on Medical Imaging*, vol. 36, no. 2, pp. 447–456, 2017.

[22] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, USA, 2017.

[23] L. Yu, J.-Z. Cheng, Q. Dou et al., "Automatic 3D cardiovascular MR segmentation with densely-connected volumetric convNets," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 287–295, Canada, 2017.

[24] M. F. Stollenga, W. Byeon, M. Liwicki, and J. Schmidhuber, "Parallel multi-dimensional LSTM, with application to fast biomedical volumetric image segmentation," *Advances in Neural Information Processing Systems*, vol. 2, pp. 2998–3006, 2015.

[25] Q. Dou, H. Chen, L. Yu et al., "Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1182–1195, 2016.

[26] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: fully convolutional neural networks for volumetric medical image segmentation," in *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, IEEE, Stanford, California, 2016.

[27] S. Andermatt, S. Pezold, and P. Cattin, "Multi-dimensional gated recurrent units for the segmentation of biomedical 3D-data," in *Proceedings of the International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis International Workshop on Deep Learning in Medical Image Analysis*, pp. 142–151, Athens, Greece, 2016.

[28] K. Kamnitsas, C. Ledig, V. F. J. Newcombe et al., "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Medical Image Analysis*, vol. 36, pp. 61–78, 2017.

[29] J. Dolz, C. Desrosiers, and I. B. Ayed, "3D fully convolutional networks for subcortical segmentation in MRI: a large-scale study," *NeuroImage*, vol. 170, pp. 456–470, 2017.

[30] E. Hosseini-Asl, R. Keynto, and A. El-Baz, "Alzheimer's disease diagnostics by a deeply supervised adaptable 3D convolutional network," in *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, USA, 2016.

[31] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng, "3D deeply supervised network for automatic liver segmentation from CT volumes," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, Athens, Greece, 2016.

[32] J. Cheng, L. Dong, and M. Lapata, *Long Short-Term Memory-Networks for Machine Reading*, EMNLP, 2016.

[33] A. P. Parikh, O. Tackstr ¨ om, D. Das, and J. Uszkoreit, *A Decomposable Attention Model for Natural Language Inference*, EMNLP, 2016.

[34] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," 2017, https://arxiv.org/abs/1706.03762.

[35] Q. Dou, C. Quyang, C. Chen et al., "PnP-AdaNet: plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation," 2018, https://arxiv.org/abs/1812.07907.

[36] Q. Dou, C. Quyang, C. Chen, H. Chen, and P.-A. Heng, "Unsupervised cross-modality domain adaptation of convnets

for biomedical image segmentations with adversarial loss," 2018, https://arxiv.org/abs/1804.10916.

[37] J. Schlemper, O. Oktay, M. Schaap et al., "Attention gated networks: learning to leverage salient regions in medical images," *Medical Image Analysis*, vol. 53, pp. 197–207, 2019.

[38] Z. Shi, G. Zeng, L. Zhang et al., "Bayesian VoxDRN: a probabilistic deep voxelwise dilated residual network for whole heart segmentation from 3D MR images," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 569–577, Spain, 2018.

[39] K. Liu, Y. Ma, and H. Xiong, *Medical Image Super-Resolution Method Based on Dense Blended Attention Network*, 2019.

[40] C. Kaul, S. Manandhar, and N. Pears, "FocusNet: an attention-based fully convolutional network for medical image segmentation," in *Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI)*, Italy, 2019.

[41] X. Wang, R. Girshick, A. Gupta, and H. Kaiming, "Non-local Neural Networks," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7794–7803, Salt Lake City, UT, USA, 2018.

[42] H. Zhang, I. Goodfellow, and D. Metaxas, "Self-attention generative adversarial networks" augustus odena proceedings of the 36th international conference on machine learning," *PMLR*, vol. 97, pp. 7354–7363, 2019.

[43] M. Lin, Q. Chen, and S. Yan, "Network In Network," 2013, https://arxiv.org/abs/1312.4400.

[44] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, California, USA, 2017.

[45] A. Mukhopadhyay, "Total variation random forest: fully automatic MRI segmentation in congenital heart disease," *Reconstruction, Segmentation, and Analysis of Medical Images*, vol. 10129, 2016.

[46] G. Tziritas, "Fully-automatic segmentation of Cardiac images using 3-D MRF model optimization and substructures tracking," in *Proceedings of the International Workshop on Reconstruction & Analysis of Moving Body Organs*, Springer International Publishing, Athens, Greece, 2016.

[47] R. Shahzad, S. Gao, and Q. Tao, "Automated cardiovascular segmentation in patients with congenital heart disease from 3D CMR scans: combining multi-atlases and level-sets," in *Proceedings of the International Workshop on Reconstruction & Analysis of Moving Body Organs*, Springer International Publishing, Athens, Greece, 2016.

[48] L. Yu, X. Yang, J. Qin, and P. Heng, "3D FractalNet: dense volumetric segmentation for cardiovascular MRI volumes," *Reconstruction, Segmentation, and Analysis of Medical Images*, vol. 10129, pp. 103–110, 2017.