Temporal-Spatial Graph Attention Networks for DCE-MRI Breast Tumor Segmentation

Tianxu Lv
lvtianxu1@icloud.com
Xiang Pan*
xiangpan@jiangnan.edu.cn
The School of Artificial Intelligence and Computer Science
Jiangnan University
Wuxi, China

Abstract

Recent researches on medical image segmentation resort to the combination of natural image segmentation models and medical domain knowledge. However, prior methods only focus on single image segmentation or 3D convolutional operation based volume segmentation, and overlook the spatial correlations of inter-slice and temporal correlations of the inter-sequence in DCE-MRI images. In this paper, we propose a novel end-to-end temporal-spatial graph attention network (*TSGAN*), which precisely segments tumor of 4D (volume space, time) DCE-MRI images by conjointly exploiting the spatial contextual dependency of inter-slice and temporal contextual dependency of inter-sequence. Specially, we design a graph temporal attention module to integrate the temporal-spatial representations hidden in 4D data into deep segmentation. The spatial dependency is learnt by graph attention operation, which attends over its neighbourhoods' features for each vertex. Meanwhile, the spatial representations learnt by the graph attention layer are combined with the temporal representations by a temporal attention operator. Then the temporal dependency is exploited by spreading on the graph. We also design a tumour structural similarity (TSS) loss used to exploit the tumour structural dependency and enhance inter-voxel similarity within the same tissue for segmentation. We demonstrate that the proposed model outperforms recent state-of-the-art methods through comprehensive experiments.

1 Introduction

Breast cancer [8] is the primary cause of death from cancer among women. Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) can non-invasively uncover both temporal and spatial properties of the physiological tissue, which plays an important role in diagnosis and staging of breast tumors [23] [43]. For treatment selection and therapy evaluation, it is a vital process to precisely segment breast tumor in DCE-MRI images. However, it is not only challenging but also time-consuming to segment breast tumor manually. So many automatic breast tumor segmentation methods have been proposed.

The classical methods of breast tumor segmentation consist of three types: threshold based methods [15, 47], graph partitioning methods [13, 25] and cluster based methods [2, 3, 33]. Threshold based methods tried to automatically select an optimal threshold for

*Xiang Pan is the corresponding author.

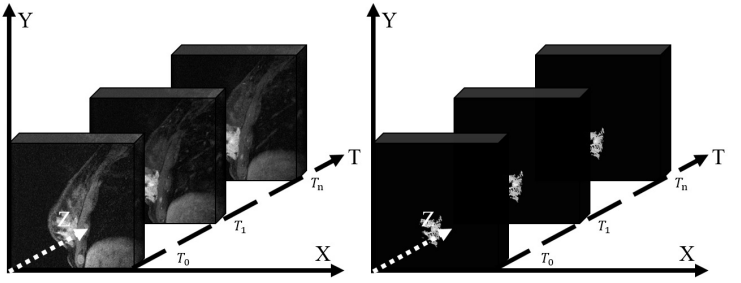

Figure 1: DCE-MRI images and tumor segmentation of breast cancer of one patient. The left displays the breast tissue and the right is the corresponding tumor segmentation.

identifying differences between background and tumors. Graph partitioning methods were proposed to take advantage of the similarity between the neighborhood pixels. The image can be seen as a graph, in which each pixel was regarded as a node and the similarity between pixels were seen as the edges. Breast tumor was segmented by partitioning the graph according to a specific criterion. Cluster based methods was utilized to learn statistical information between background and tumors. However, these methods are difficult to obtain accurate breast tumor segmentation due to the complexity of tumor texture characteristics [1, 34] and intensity similarity between background and tumors.

To overcome these difficulties, deep learning based segmentation approaches have been proposed. Many segmentation models based on Convolutional neural networks have achieved remarkable performance. U-Net [35] with a contracting path and a symmetric expanding path was proposed for accurate semantic segmentation. Inspired by U-Net, V-Net [30] was proposed for using volumetric convolutions to process 3D images. Pixel deconvolutional layer (PixelDCL) [11] tried to focus on the intrinsic relations among neighboring pixels during the up-sample process. However, these models did not take into account non-local information for a limited receptive field. Several methods have been presented to deal with this issue. Yu *et al.* [45] integrated multi-scale contextual information and enlarge the receptive field within dilated convolutions. In addition, multi-scale self-guided attention [38] and criss-cross attention [17] modules were proposed to capture rich contextual dependencies and yield discriminative features. More recently, several spatio-temporal architectures [7, 21] were presented to capture spatio-temporal information for segmentation. Volumetric Spatio-temporal memory networks [21] were deployed to exploit spatio-temporal information from CT scans to improve the performance. Although the aforementioned methods enforce global contextual dependencies, they are generic and are not optimal for specific applications. Particularly, a core challenge for DCE-MRI image segmentation is how to effectively model spatial and temporal relations.

In this paper, we focus on addressing the following segmentation challenges raised by DCE-MRI images. First, inter-slices of medical images contain spatial contextual dependency. Although 3D convolutional operation was used to exploit it, the computational cost is high. Second, the imaging of the same position in different time sequences contain temporal contextual dependency, which is not yet used for tumor segmentation.

To tackle the above challenges, we propose a novel end-to-end DCE-MRI image segmentation framework called Temporal-Spatial Graph Attention Network (*TSGAN*) that exploits spatial and temporal contextual dependencies. First, we encode the input volume data as a

spatial graph, where vertexes represent slices and edges measure spatial relations between them. In order to effectively exploit the spatial contextual dependency, we introduce graph attention mechanism to exploit the correlations between each pair of vertexes. Second, we design a temporal attention operator, which is applied to representations of different time points, to implicitly incorporate temporal relations into the graph. In addition, we deploy a tumour structural similarity (TSS) loss to exploit the tumour structural information for segmentation.

In summary, we make the following contributions: (1) We design a novel end-to-end temporal-spatial graph attention network to perform DCE-MRI image segmentation by exploiting spatial and temporal contextual dependencies. To the best of our current knowledge, this is the first time that spatial and temporal contextual dependencies are simultaneously leveraged for 4D breast tumour segmentation. (2) We propose a graph temporal attention module, which integrates temporal attention with graph attention, to jointly consider the spatial and temporal relations. TSS loss is introduced to exploit tumour structural information. (3) Comprehensive experiments show that the proposed model outperforms the state-of-the-art methods. Further experiments verify the contribution of different temporal volumes for segmentation.

2 Related Work

2.1 Tumor Segmentation

Recently, convolutional neural networks (CNNs)-based architecture has achieved great results in the domain of tumor segmentation. Fully convolutional networks (FCNs) [27] and U-Net [35] have been widely used for semantic segmentation. Both networks adopt the encoder-decoder architectures to promote the accuracy and speed of segmentation. Havaei *et al.* [14] proposed a two-pathway CNNs model for exploiting local details and contextual information. However, the above methods focus on 2D slices obtained from 3D volume data, leading to the spatial contextual information missing. To handle this issue, 3D convolutional kernels are performed on 3D origin volume data. For instance, Myronenko [31] introduced a 3D encoder-decoder structure with a variational auto-encoder branch for tumor segmentation. V-Net [30] utilized volumetric convolutions and a dice coefficient based objective function to segment the volume data at once. In addition, other methods that exploit spatial context consist of formulating anisotropic and dilated convolution operations [4], combining conditional random field (CRF) [48], and employing attention mechanisms [26, 45]. While the aforementioned methods show promising consequences, it requires further efforts to explore how to jointly exploit spatial and temporal contextual dependencies for tumor segmentation in DCE-MRI images.

2.2 Graph Neural Networks (GNNs)

GNNs were introduced in Gori *et al.* [12] and Scarselli *et al.* [36] to effectively handle general graphs like undirected and directed graphs. Graph convolutional networks (GCNs) [20] were proposed for learning hidden layer representations using a first-order approximation of graph spectral convolutions. GCNs and its variants [16, 22, 44] have attracted a surge of interest. However, the graph convolution operation requires specific graph structure. Veličković *et al.* [39] proposed graph attention networks that deal with graphs using self-

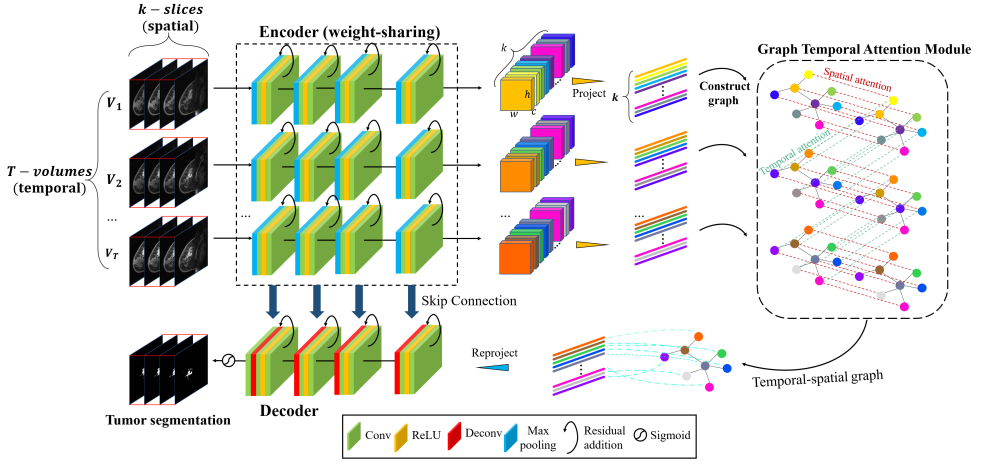

Figure 2: Overview of the proposed model. The input DCE-MRI images have T sequences and every sequence has k slices. We use a weight-sharing encoder to extract the feature map for each slice. Then, feature maps are projected to obtain the useful representations, which are corresponding to the slices. The representations at the first time are employed to construct a original graph, whose vertexes represent slices and edges indicate the correlation between slices. Following, the temporal-spatial contextual dependencies are extracted by graph temporal attention module. Ultimately, The temporal-spatial graph is reprojected into Decoder to obtain the tumor segmentation.

attentional layers, in which nodes can attend over the features of their neighbourhoods. In recent years, GNNs have been applied to tackle computer vision tasks, such as hand-object pose estimation [9], 3D object detection [5, 37], video-based person re-identification [42] and visual relationship detection [29].

Graph-based methods have been paid more attention in segmentation. Specifically, Li *et al.* [24] proposed spatial pyramid based graph reasoning networks to exploit multiple long-range contextual patterns through graph reasoning in the feature space. A sparse layered graph [19] was proposed for graph cut segmentation by adding explicit object interactions. Wu *et al.* [41] explored intra-modular and inter-modular relationships between background and foreground things for panoptic segmentation. Although these methods explore spatial information, the spatial and temporal contextual dependencies are not fully exploited.

3 Proposed Methods

3.1 Overview

As shown in Figure 2, the proposed model consists of three key components: (a) Encoder, (b) Graph temporal attention module (GTAM), and (c) Decoder. Particularly, given DCE-MRI images, we represent it as $U = \{V_1, V_2, ..., V_T\}$, where T is the number of samples after the contrast injection. Meanwhile, for each volume of DCE-MRI images, we denote it as $V_i = \{S_1, S_2, ..., S_K\}$, where K is the number of slices in each volume. Firstly, we apply a weight-sharing encoder to extract their feature maps $\chi = \{\mathbf{X}^l\}_{l=1}^{T}$ as the multi-temporal fea-

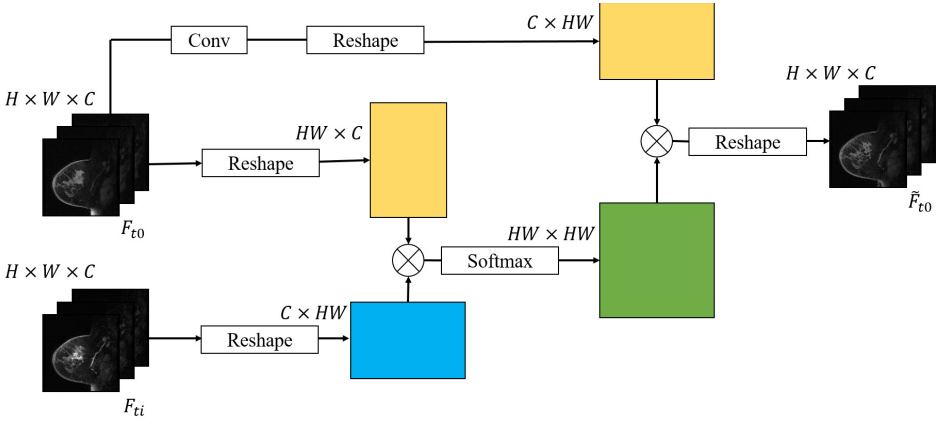

Figure 3: The computation procedures of the temporal attention mechanism. F_{t0} means the origin features of each vertex, F_{ti} represents the temporal features of the corresponding vertex and $\widetilde{F}_{t0}$ represents the output features.

ture representations of U. And each $\mathbf{X}^l \in \chi$ consists of K feature maps: $\mathbf{X}^l = F_1, F_2, ...F_K$, where F_j is the feature map of j-th slice in the volume, $F_j \in \mathbb{R}^{h \times w \times c}$, where h, w, c represents the height, width and channel number. The whole encoder is based on 2D convolution operation to reduce the computation complexity. Secondly, for each $\mathbf{X}^l \in \chi$, we design a spatial projection by 1×1 convolution kernel operation, which is used to transform the feature map into vertex-based feature map while preserving the spatial consistency. Afterwards, to effectively capture spatial contextual dependency, the vertex-based features of V_i are applied to the construction of a origin graph G. Then, for the graph G, we design a temporal attention operation that is adaptive to our DCE-MRI image segmentation task, which can well exploit inter-sequence correspondence while preserving the temporal contextual dependency. Therefore, the $\widetilde{G}$ is fed into temporal attention operation, producing a temporal graph $\widetilde{G}$, which preserves temporal contextual dependency. Thirdly, $\widetilde{G}$ is integrated into a graph attention layer, updating vertexes by attending over their neighbourhoods' features and generating the temporal-spatial graph $\widetilde{\widetilde{G}}$. Finally, the concatenated features $\chi \copyright \widetilde{\widetilde{G}}$ are fed into a decoder, generating the predicted DCE-MRI tumor segmentation.

3.2 Spatial Graph Construction

As aforementioned, to exploit and utilize the spatial contextual dependency of inter-slices, we employ graph attention layer (GAT) to model the spatial relations between slices. We define the constructed spatial graph as $G(V,E)$ of K vertexes with vertexes $v_i \in V$ and edges $e_{ij} = (v_i, v_j) \in E$. Each slice is considered as a vertex and the edges represent the spatial relations between slices. To reduce the computation complexity, we select q slices as its neighbours to construct an undirected graph.

The spatial relations between every two slices in the graph are initialized as follows:

$$e(v_i, v_j) = \frac{exp(ReLU(v_i - v_j))}{\sum_{q \in K_i} exp(ReLU(v_i - v_j))} \tag{1}$$

3.3 Graph temporal attention Module

To effectively explore the temporal contextual dependency of inter-sequences, we introduce a temporal attention operation to model the temporal relations between sequences. We use F_0 to represent the origin feature of each vertex and F_i to represent the temporal feature of the corresponding vertex. After temporal attention operation, we can obtain the output features $\widetilde{F_0}$, which learns the temporal contextual dependency. First, we calculate the attention map $\widetilde{F_{ti}}$ by F_0 and F_i. The computational process of $\widetilde{F_{ti}}$ is shown as Fig.3, and the formula is as follows:

$$\widetilde{F_{ti}} = \phi(F_{t0}; W_\phi) Softmax(\phi^T(F_{t0})\phi(F_{ti})) \tag{2}$$

where W_ϕ is a learnable parameter.Then, the final temporal attention is defined as :

$$\widetilde{F_{t0}} = \alpha \widetilde{F_{ti}} + F_{t0} \tag{3}$$

where α is a trainable parameter and $\widetilde{F_{ti}}$ focuses on temporal relations of inter-sequence.

Therefore, after temporal attention operation, we can obtain a temporal graph $\widetilde{G}(\widetilde{V}, E)$ with vertexes $\widetilde{v_i} \in V$. The next step is to excavate the spatial contextual dependency. We employ a learnable parameters $W \in \mathbb{R}^{M' \times M}$ for linear transformation, in which M represents the dimension of $\widetilde{v_i}$. The spatial graph attention coefficients are computed as follows:

$$e_{ij} = \alpha(W\widetilde{v_i}, W\widetilde{v_j}) \tag{4}$$

where $\widetilde{v_j}$ is a neighbor of $\widetilde{v_i}$, e_{ij} indicates the significance of slice j-th features to slice i and α is a trainable parameter. For reducing the computation complexity, we select q neighbours of each vertex i, which are defined as: $\{\widetilde{v}_{i-q/2}, ..., \widetilde{v}_{i-1}, \widetilde{v}_{i+1}, ..., \widetilde{v}_{i+q/2}\}$. In our experiments, the q is usually set as 4. Following, to make the attention coefficients simply comparable with every vertex, we normalize the acquired spatial graph attention by softmax operation. The final spatial graph attention coefficients are calculated as follows:

$$e_{ij} = \frac{exp(f(a^T[W\widetilde{v_i} \parallel W\widetilde{v_j}]))}{\sum_{q \in K_i} exp(f(a^T[W\widetilde{v_i} \parallel W\widetilde{v_q}]))} \tag{5}$$

where f represents LeakyRelu function, $\parallel$ represents concatenation operation and a^T is a learnable weight. And we can get the final temporal-spatial representation of each vertex by a linear combination:

$$\overline{\widetilde{v_i}} = f(\sum_{j \in K_i} e_{ij} W\widetilde{v_j}) \tag{6}$$

in which f signifies LeakyRelu function. Then the learnt temporal-spatial representation $\overline{\widetilde{v_i}}$ is passed through a decoder to produce the final tumour segmentation results.

3.4 Loss Function

Breast tumour contains abundant structural information, which is useful to calculate for segmentation. However, most traditional segmentation loss functions overlook this point. Structural Similarity is an indicator to measure image similarity, which is often used in image restoration and enhancement domain. Hence, we intuitively employ it to exploit the tumour structural information. Formally, we denote P_i as the prediction map of i-th slice and T_i as

the target map of i-th slice. Then we compute the tumour structural similarity loss L_{TSS} as the following equation:

$$L_{TSS} = -\frac{(2\mu_{P_i}\mu_{T_i} + c_1)(2\sigma_{P_iT_i} + c_2)}{(\mu_{P_i}^2 + \mu_{T_i}^2 + c_1)(\sigma_{P_i}^2 + \sigma_{T_i}^2 + c_2)} \tag{7}$$

where μ_{P_i} is the mean of P_i, μ_{T_i} is the mean of T_i, $\sigma_{P_i}^2$ is the variance of P_i, $\sigma_{T_i}^2$ is the variance of T_i, c_1 and c_2 is constant to maintain stability, and $\sigma_{P_iT_i}$ is the covariance of P_i and T_i.

By integrating the tumour structural similarity minimization principle with *TSGAN* model, we can obtain the objective function:

$$L_{TSGAN} = L_S + \lambda_{tss} L_{TSS} \tag{8}$$

in which L_S is the dice loss to evaluate the segmentation performance, L_{TSS} measures the tumour structural similarity and λ_{tss} is the regularization parameter to balance the trade-off.

4 Experiments

4.1 Datasets and Implementation Details

We conduct experiments on the publicly available Breast-MRI-NACT-Pilot dataset [32] to evaluate our proposed *TSGAN*. The Breast-MRI-NACT-Pilot dataset includes DCE-MRI data of 64 patients with breast cancer, which are obtained on a 1.5-T scanner (Signa, GE Healthcare, Milwaukee, WI) using a bilateral phased array breast coil, high spatial resolution and low temporal resolution. The section thickness is 2 mm and the size of MR matrix is 256×256×60. Each DCE-MRI data contains three time points, including initial moment, 2.5 minutes and 7.5 minutes from the start of the contrast injection. We randomly select 80% patient data for training and the other 20% patient data for testing to maintain the consistency of data distribution.

The proposed model was implemented with Pytorch and the training was executed on double NVIDIA RTX 2080Ti GPUs. We utilize the Adam optimizer to optimize the model, whose hyper-parameter is set as $\beta_1 = 0.5$, $\beta_2 = 0.999$ empirically. The initial learning rate is set as $\alpha = 2\times10^{-5}$ and is decreased with a weight decay of 2×10^{-6}. For the hyper-parameters in the proposed method, λ_{tss} is set as 0.2 empirically. In all the experiments, we use 256×256 original size with batch size of 9 and training is stopped when the learning rate drops below 10^{-9} or 2000 epochs are exceeded. We do not use any data augmentation.

The *PA* (Pixel Accuracy), *F1-score*, *mIoU* (mean intersection over union), *RVD* (Relative volume Difference) and *DSC* (Dice Similarity Coefficient) metrics are adopted for evaluation. *PA* measures the match between ground truth and predicted segmentation in simple pixel level. *F1-score* measures the harmonic average of the precision and recall to balance the difference between precision and recall. *mean-IoU* measures the match between ground truth and predicted segmentation by calculating the ratio of intersection and union. *RVD* measures the volume difference between predicted segmentation and ground truth. *DSC* is used to evaluate the similarity between predicted results and ground truth by computing intersection area and total area. A better segmentation method has a larger value of *PA*, *F1-score*, *mean-IoU* and *DSC* while a smaller value of *RVD*.

Table 1: Quantitative comparison with other methods (Mean±std).

Method	*PA*(%)	*F1*(%)	*mIoU*(%)	*RVD*(voxel)	*DSC*(%)
3D-Unet [6]	94.2±1.6	38.5±3.3	52.8±3.8	8.7±1.4	48.5±3.5
Vnet [30]	97.4±1.1	45.9±2.9	65.1±3.4	3.4±1.0	52.9±3.2
NlNet [40]	96.8±0.9	45.6±2.6	64.6±3.7	3.9±1.1	52.2±3.1
DuANet [10]	96.5±1.3	45.1±2.9	63.4±2.9	6.3±1.6	52.3±3.1
MultiResUNet [18]	98.0±0.9	55.4±2.8	70.2±2.7	2.4±1.0	55.6±2.8
LNet [7]	97.6±1.2	48.1±2.4	65.1±3.1	2.7±1.3	52.3±2.9
DCUNet [28]	98.1±0.9	49.1±2.5	67.7±2.7	2.1±0.9	53.1±2.6
TSGAN	**98.5±0.7**	**63.2±2.0**	**74.7±2.5**	2.2±0.8	**63.5±2.5**

4.2 Experimental Results

We compare our *TSGAN* with other methods for tumor segmentation on Breast-MRI-NACT-Pilot dataset, including 3D-Unet [6], Vnet [30], Non-local Network (NlNet) [40], Dual Attention Network (DuANet) [10], MultiResUNet (MRUNet) [18], Longitudinal Network (LNet) [7] and DCUNet [28]. 3D-Unet and Vnet are designed to solve the 3D volume segmentation, and has been applied to many medical image segmentation works [46] [6]. NlNet and DuANet are presented to exploit global information by non-local block or attention mechanism. MultiResUNet and DCUNet are the work of improvement of U-Net. LNet is proposed to learn from spatio-temporal changes to guide the network for Multiple Sclerosis Lesion Segmentation. To make a fair comparison, we not only use the same data preprocessing for all methods, but retrain all models using unified implementation.

Table 1 shows the quantitative comparison with other methods. We report the mean value and standard deviation of repeat experiments for each method. Table 1 reveals that the proposed model can achieve the improved performance on most evaluation metrics compared with other methods. These deep learning methods including 3D-Unet, Vnet, Non-local Network, Dual Attention Network, MultiResUNet and DCUNet only extract the spatial features based on 3D convolutional operation or attention mechanism. Due to DCE-MRI images are four-dimensional data containing spatial and temporary information, it is the most appreciate to take full advantage of spatial and temporal features simultaneously. The proposed *TSGAN* can exploit both spatial and temporal features based on graph temporal attention module, which can guide the tumour segmentation. We note that DCUNet can achieve better results on *RVD* metric. However, the proposed model can outperform the DCUNet by 0.4%, 14.1%, 7.0% and 10.4% in *PA*, *F1*, *mIoU* and *DSC*. In comparison with the results of LNet learning spatio-temporal changes, the mean *PA*, *F1*, *mIoU*, *RVD* and *DSC* increase by approximately by 0.9%, 15.1%, 9.6%, 0.5 and 11.2%, respectively. In general, the proposed method can achieve improvements among most metrics (*PA*: 0.4%, *F1-score*: 7.8%, *mean-IoU*: 4.5%

Table 2: Time of each epoch in training.

Method	3D-Unet	Vnet	NlNet	DuANet	MRUNet	LNet	DCUNet	**Ours**
Time(s)	70.3	32.5	29.3	52.0	42.0	39.8	51.8	**20.4**

Table 3: Quantitative comparison of different temporal volumes. (Mean±std)

	PA(%)	*F1*(%)	*mIoU*(%)	*RVD*(voxel)	*DSC*(%)
T0+T1	95.6±0.8	33.6±3.8	59.1±3.6	9.8±1.3	50.1±3.1
T0+T2	96.1±0.8	32.1±3.4	58.4±3.7	5.9±0.9	49.2±2.8
T1+T2	97.7±0.6	55.6±2.1	70.8±2.2	4.1±1.1	58.9±2.8
T0+T1+T2	**98.5±0.7**	**63.2±2.0**	**74.7±2.5**	**2.1±0.8**	**63.5±2.5**

Table 4: Mean pixel intensity of different temporal volume subtraction.

	T1-T0	T2-T0	T2-T1
Mean pixel intensity	164.89884	159.57919	25.49661

and *DSC*: 7.9%).

We also compare running speed comparison with other methods. Table 2 shows the running time of each epoch. All the experiments are implemented on a device with a 3.60GHz Intel(R) Core(TM) i9-9900K CPU, 16GB RAM, and two NVIDIA RTX 2080Ti GPUs. It is distinct that the proposed model is faster than other methods. We attribute this to that the proposed model is based on 2D operation, which can avoid complex computations.

4.3 Ablation Study

To further verify the contribution of different temporal volumes, we design ablation experiments to figure out the effect. The ablation experiments include three kinds of combination of temporal volumes (T0+T1, T0+T2 and T1+T2) for that the dataset have three time points. Table 3 shows the quantitative results of different temporal volumes. From Table 3, we can see that the segmentation performance of T0+T1 and T0+T2 are inferior to the T0+T1+T2 on all evaluation metrics. We can also find that the segmentation results of T1+T2 outperform T0+T1 and T0+T2. We attribute this to that different temporal volumes contain diverse temporal-spatial contextual information and play different roles for tumour segmentation. In addition, the segmentation results of T0+T1+T2 excels other three combinations by 0.8%, 7.6%, 3.9%, 2.0 voxels and 4.6% in *PA*, *F1*, *mIoU*, *RVD* and *DSC*. We attribute it to that three temporal volumes own more abundant temporal-spatial contextual information. From Table 3, we can deduce that T1+T2 temporal volumes account for the major contribution for tumour segmentation. To confirm the above assumption, for two temporal volumes, we utilize the subtraction and then compute the mean pixel value to represent the temporal characteristic value. We calculate mean pixel intensity of different temporal volume subtraction to seek the potential relationship. Table 4 shows the value of mean pixel intensity. From Table 4, we can see that the mean pixel intensity of T1-T0 and T2-T0 is quite similar, and both are different from T2-T1. Thus, this is a element influencing the segmentation results for it has different temporal contextual information.

In addition, we conduct an ablation study to identify the impact of the proposed loss function. The qualitative comparison of different components is shown in Figure 4. We observe that the segmentation details achieve significant improvement when compared to the baseline model. Table 5 reports the quantitative comparison of different components. It is clearly that the tumour structural similarity loss brings overall performance improvement

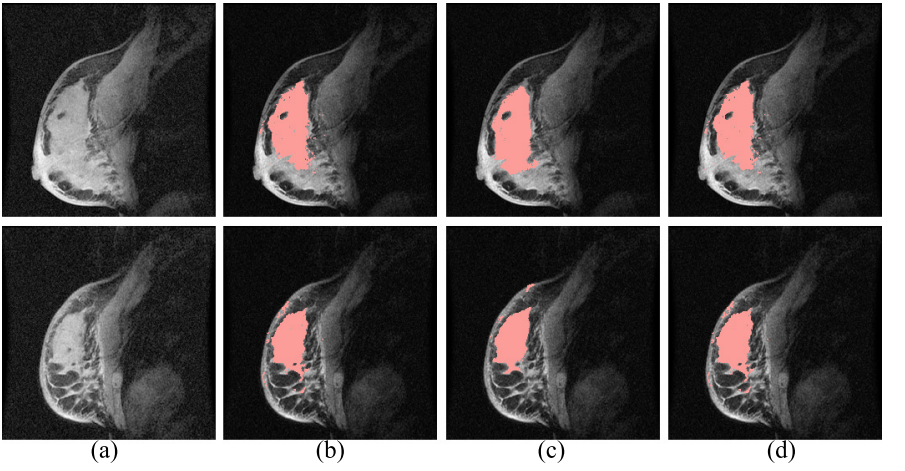

Figure 4: Visualization comparison of different components. (a) MR image. (b) ground truth. (c) Baseline. (d) Baseline+L_{TSS}

Table 5: Ablation analysis of different components in the proposed method. (Mean±std)

	PA(%)	*F1*(%)	*mIoU*(%)	*RVD*(voxel)	*DSC*(%)
Baseline	98.0±0.8	49.2±2.5	68.1±2.7	2.4±0.9	58.5±2.6
Baseline+L_{TSS}	**98.5±0.7**	**63.2±2.0**	**74.7±2.5**	**2.2±0.8**	**63.5±2.5**

(*DSC*: 0.5%, *F1*: 14.0%, *mIoU*: 6.6%, *RVD*: 0.2 voxel and *DSC*: 5%) compared with baseline, which is implemented with only dice loss. This indicates that the proposed L_{TSS} can contribute to the learning of breast tumor structural information for better segmentation.

5 Conclusion

In this paper, we propose a novel temporal-spatial graph attention network (*TSGAN*) to perform 4D DCE-MRI tumor segmentation. We explore the spatial contextual dependency of inter-slice and temporal contextual dependency of inter-sequence. We design a graph temporal attention module to integrate the temporal-spatial information hidden in DCE-MRI images into tumor segmentation. The whole network is based on 2D operations, so it can avoid complex and expensive computations. The experimental results demonstrate the superiority of our *TSGAN*, which achieves competitive performance both in accuracy and efficiency.

Acknowledgement

This work is supported in part by the National Key R&D Program of China under Grants 2018YFA0701700 and 2017YFC0109402, and is supported by National Natural Science Foundation of China grants 61602007 and 61731008, and Zhejiang Provincial Natural Science Foundation of China (LZ15F010001).

References

[1] André Victor Alvarenga, Wagner CA Pereira, Antonio Fernando C Infantosi, and Carolina M Azevedo. Complexity curve and grey level co-occurrence matrix in the texture evaluation of breast tumor on ultrasound images. *Medical physics*, 34(2):379–387, 2007.

[2] Agus Zainal Arifin and Akira Asano. Image segmentation by histogram thresholding using hierarchical cluster analysis. *Pattern recognition letters*, 27(13):1515–1521, 2006.

[3] Sukbin Cha, Ken W McCleary, and Muzaffer Uysal. Travel motivations of japanese overseas travelers: A factor-cluster segmentation approach. *Journal of travel research*, 34(1):33–39, 1995.

[4] Chen Chen, Xiaopeng Liu, Meng Ding, Junfeng Zheng, and Jiangyun Li. 3d dilated multi-fiber network for real-time brain tumor segmentation in mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 184–192. Springer, 2019.

[5] Jintai Chen, Biwen Lei, Qingyu Song, Haochao Ying, Danny Z Chen, and Jian Wu. A hierarchical graph network for 3d object detection on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 392–401, 2020.

[6] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. *CoRR*, abs/1606.06650, 2016. URL http://arxiv.org/abs/1606.06650.

[7] Stefan Denner, Ashkan Khakzar, Moiz Sajid, Mahdi Saleh, Ziga Spiclin, Seong Tae Kim, and Nassir Navab. Spatio-temporal learning from longitudinal data for multiple sclerosis lesion segmentation, 2020. URL https://arxiv.org/abs/2004.03675.

[8] Carol DeSantis, Jiemin Ma, Leah Bryan, and Ahmedin Jemal. Breast cancer statistics, 2013. *CA: a cancer journal for clinicians*, 64(1):52–62, 2014.

[9] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6608–6617, 2020.

[10] Jun Fu, Jing Liu, Haijie Tian, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. *CoRR*, abs/1809.02983, 2018. URL http://arxiv.org/abs/1809.02983.

[11] Hongyang Gao, Hao Yuan, Zhengyang Wang, and Shuiwang Ji. Pixel deconvolutional networks. *arXiv preprint arXiv:1705.06820*, 2017.

[12] Marco Gori, Gabriele Monfardini, and Franco Scarselli. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 729–734. IEEE, 2005.

[13] Leo Grady and Eric L Schwartz. Isoperimetric graph partitioning for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 28(3):469–475, 2006.

[14] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.

[15] Karla Horsch, Maryellen L Giger, Luz A Venta, and Carl J Vyborny. Computerized diagnosis of breast lesions on ultrasound. *Medical physics*, 29(2):157–164, 2002.

[16] Z. Huang, X. Li, Y. Ye, and M. K. Ng. Mr-gcn: Multi-relational graph convolutional networks based on generalized tensor product. In *Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence IJCAI-PRICAI-20*, 2020.

[17] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 603–612, 2019.

[18] Nabil Ibtehaz and M.Sohel Rahman. Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural Networks*, 121:74–87, Jan 2020. ISSN 0893-6080. doi: 10.1016/j.neunet.2019.08.025. URL http://dx.doi.org/10.1016/j.neunet.2019.08.025.

[19] Niels Jeppesen, Anders N Christensen, Vedrana A Dahl, and Anders B Dahl. Sparse layered graphs for multi-object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12777–12785, 2020.

[20] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[21] Tetiana Klymenko, Seong Tae Kim, Kirsten Lauber, Christopher Kurz, Guillaume Landry, Nassir Navab, and Shadi Albarqouni. Butterfly-net: Spatial-temporal architecture for medical image segmentation. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 616–620, 2021. doi: 10.1109/ISBI48211.2021.9433939.

[22] Ruoyu Li, Sheng Wang, Feiyun Zhu, and Junzhou Huang. Adaptive graph convolutional neural networks. *arXiv preprint arXiv:1801.03226*, 2018.

[23] Xia Li, Lori R Arlinghaus, Gregory D Ayers, A Bapsi Chakravarthy, Richard G Abramson, Vandana G Abramson, Nkiruka Atuegwu, Jaime Farley, Ingrid A Mayer, Mark C Kelley, et al. Dce-mri analysis methods for predicting the response of breast cancer to neoadjuvant chemotherapy: Pilot study findings. *Magnetic resonance in medicine*, 71 (4):1592–1602, 2014.

[24] Xia Li, Yibo Yang, Qijie Zhao, Tiancheng Shen, Zhouchen Lin, and Hong Liu. Spatial pyramid based graph reasoning for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8950–8959, 2020.

[25] Zhenguo Li, Xiao-Ming Wu, and Shih-Fu Chang. Segmentation using superpixels: A bipartite graph partitioning approach. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 789–796. IEEE, 2012.

[26] Zhihua Liu, Lei Tong, Long Chen, Feixiang Zhou, Zheheng Jiang, Qianni Zhang, Yinhai Wang, Caifeng Shan, Ling Li, and Huiyu Zhou. Canet: Context aware network for 3d brain tumor segmentation. *arXiv preprint arXiv:2007.07788*, 2020.

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[28] Ange Lou, Shuyue Guan, and Murray H Loew. Dc-unet: rethinking the u-net architecture with dual channel efficient cnn for medical image segmentation. In *Medical Imaging 2021: Image Processing*, volume 11596, page 115962T. International Society for Optics and Photonics, 2021.

[29] Li Mi and Zhenzhong Chen. Hierarchical graph attention network for visual relationship detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13886–13895, 2020.

[30] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

[31] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI Brainlesion Workshop*, pages 311–320. Springer, 2018.

[32] David Newitt and Nola Hylton. Single site breast dce-mri data and segmentations from patients undergoing neoadjuvant chemotherapy. *The Cancer Imaging Archive*, 2016.

[33] J-R Ohm and Phuong Ma. Feature-based cluster segmentation of image sequences. In *Proceedings of International Conference on Image Processing*, volume 3, pages 178–181. IEEE, 1997.

[34] Fanny Orlhac, Michaël Soussan, Jacques-Antoine Maisonobe, Camilo A Garcia, Bruno Vanderlinden, and Irène Buvat. Tumor texture analysis in 18f-fdg pet: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis. *Journal of Nuclear Medicine*, 55(3):414–422, 2014.

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[36] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2008.

[37] Weijing Shi and Raj Rajkumar. Point-gnn: Graph neural network for 3d object detection in a point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1711–1719, 2020.

[38] Ashish Sinha and Jose Dolz. Multi-scale self-guided attention for medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2020.

[39] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[40] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *CoRR*, abs/1711.07971, 2017. URL http://arxiv.org/abs/1711.07971.

[41] Yangxin Wu, Gengwei Zhang, Yiming Gao, Xiajun Deng, Ke Gong, Xiaodan Liang, and Liang Lin. Bidirectional graph reasoning network for panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9080–9089, 2020.

[42] Jinrui Yang, Wei-Shi Zheng, Qize Yang, Ying-Cong Chen, and Qi Tian. Spatial-temporal graph convolutional network for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3289–3299, 2020.

[43] Thomas E Yankeelov, Martin Lepage, Anuradha Chakravarthy, Elizabeth E Broome, Kenneth J Niermann, Mark C Kelley, Ingrid Meszoely, Ingrid A Mayer, Cheryl R Herman, Kevin McManus, et al. Integration of quantitative dce-mri and adc mapping to monitor treatment response in human breast cancer: initial results. *Magnetic resonance imaging*, 25(1):1–13, 2007.

[44] Fanfan Ye, Shiliang Pu, Qiaoyong Zhong, Chao Li, Di Xie, and Huiming Tang. Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition. *arXiv preprint arXiv:2007.14690*, 2020.

[45] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[46] L. Zhang, J. Zhang, P. Shen, G. Zhu, and M. Bennamoun. Block level skip connections across cascaded v-net for multi-organ segmentation. *IEEE Transactions on Medical Imaging*, PP(99):1–1, 2020.

[47] Ling Zhang. A novel segmentation method for breast cancer ultrasound cad system. In *Proceedings of the 2011 International Conference on Informatics, Cybernetics, and Computer Engineering (ICCE2011) November 19–20, 2011, Melbourne, Australia*, pages 307–313. Springer, 2011.

[48] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1529–1537, 2015.

[49] Chenhong Zhou, Changxing Ding, Xinchao Wang, Zhentai Lu, and Dacheng Tao. One-pass multi-task networks with cross-task guided attention for brain tumor segmentation. *IEEE Transactions on Image Processing*, 29:4516–4529, 2020.