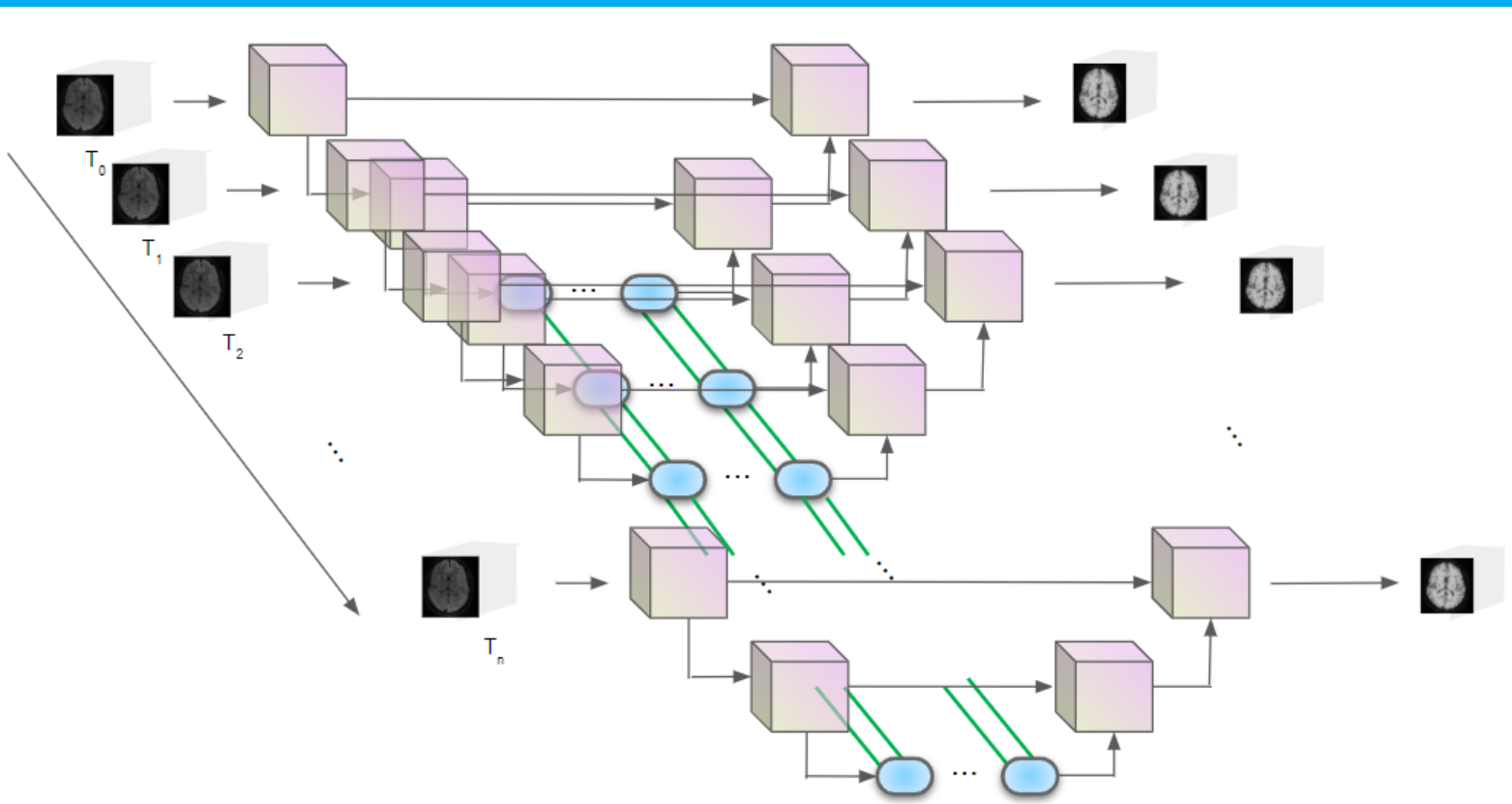
Deep Learning for 4D Longitudinal Segmentation of MRI Brain Tissues and Glioma

HAO NI

Master Thesis
Program: Msc Mechanical Engineering
Track: Biomechanical Design
Specialisation: Biorobotics



TU Delft

Deep Learning for 4D Longitudinal Segmentation of MRI Brain Tissues and Glioma

by

HAO NI

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday September 8, 2020 at 10:00 AM.

Student number:	4777166	
Project duration:	February 4, 2020 – July 31, 2020	
Thesis committee:	Prof. dr. ir. M. Wisse,	TU Delft, chairman
	Dr. W. Pan,	TU Delft, supervisor
	Dr. ir. S. Klein,	Erasmus MC, external supervisor
	K.A. van Garderen,	Erasmus MC, daily supervisor

An electronic version of this thesis is available at
http://repository.tudelft.nl/.

TU Delft BIGR Biomedical Imaging Group Rotterdam

Abstract

Glioma is a kind of slow-growing brain tumor which may result in severe seizures. Currently a major tool used to detect and diagnose the glioma is MRI scan. To better analyze the medical image, segmentation is usually conducted as a basic step for further processing, which partitions an integrate image into multiple physically meaningful regions by annotating objects and boundaries. Deep learning based segmentation methods have attracted significant interest due to their high efficiency and strong generalization ability. With the increasing demands of high-quality segmentation of bio-tissues in medical region, plenty of innovative approaches were proposed to expand the boundary of segmentation capability of deep learning models by taking the spatial or temporal constraints of bio-structure into consideration. Although, longitudinal segmentation in 2D natural image sequences has made a lot of success, the potential of deep learning network in segmenting a series of chronological 3D MRI images in terms of improving consistency remains unclear.

This thesis aims to investigate whether deep learning models are able to increase segmentation accuracy as well as consistency in longitudinal 3D images, specifically focusing on introducing Recurrent Neural Network(RNN) to 3D Convolutional Neural Network(CNN) for 4D segmentation. In addition to the implementation of several U-Net variants as CNN backbone, three types of longitudinal connection strategies are proposed. A hierarchical workflow is followed to create the optimal version of longitudinal network based on combining multiple CNN variants and connection strategies. The evaluation of the 4D network shows that segmentation accuracy of the longitudinal model is limited by its CNN backbone and temporal information can partially improve the segmentation consistency with regard to maintaining the highest proportion of normal tissue unchangeable over time.

Key words: MRI, CNN, U-Net, longitudinal network, 4D segmentation, segmentation accuracy, segmentation consistency, brain tumor segmentation

Acknowledgements

I would like to express my appreciate to the following people who provide me with assistant and support in accomplishing this thesis. First I want to thank Karin van Garderen, PhD at Erasmus MC. Without your tutor and patient, I could not finish this project in six months starting from a freshman in the medical image segmentation field. Whenever I have questions, you are always there and willing to help me address trouble. I would also thank Stefan Klein, my supervisor at Erasmus MC. Your expertise in image registration filed left me a deep impression. Every time when I show the progress, you can point out the problems existing in my methods and give me fresh ideas and useful instructions. The encouragement from Karin and Stefan supports me to go through this difficult time period. And Wei Pan, my supervisor from TU Delft, your valuable advice not only on arranging Master thesis but also on preparing my future career helps me stay away from anxiety. In addition I cannot ignore my parents, who stand by me when I decided to study abroad and provide the financial support. Without you, I will not have all these unforgettable experiences. Finally, thank you to all my friends and your accompany to help me survive in this tough time.

Contents

List of Figures

1 Introduction

1.1. Low grade glioma and Magnetic Resonance Imaging (MRI)

Primary brain tumors are severe cancers which take many lives around the world every year. Glioma is an important class of primary brain tumor that can be classified into four grades: I, II, III and IV, among which, grade I and II are termed as low grade glioma[8], a type of slow-growing tumor. Seizures are the most typical symptoms caused by low grade glioma. Based on the time and location it progresses inside brain, the symptoms can be mild or extremely serious. The exact causes of low grade glioma remain unknown. Currently the optimal treatment of this disease is controversial, with a combination of surgery, observation, and radiation. If the location of glioma within brain is safe enough, then removing it as much as possible by operation would be an idea option. Even though, there is no guarantee that the tumor will not come back again after resection. If so, following surgery have to be considered. To monitor the growth of glioma, patients are required to take the brain examination regularly[7]. Magnetic Resonance Imaging (MRI) scan is an important tool used for the non-invasive diagnosis and and quantitative analysis during the treatment of low grade glioma.

With the rapid progress of technology and massive research on medical instruments in past years, various imaging technologies are applied to healthcare sector, including computed tomography scan (CT), MRI, ultrasound, positron emission tomography (PET) etc. Among them, MRI is featured with little health hazards because of no radiation to human body and no limits to the number of scans on the subject, which leads it to the prominent position in current medical diagnose and clinical research. Frequently used MRI in diagnose is multi-parametric MRI, also known as multi-modality MRI, composed of several different pulse sequences. The most common MRI sequences include T1-weighted, T2-weighted, Fluid Attenuated Inversion Recovery (Flair) and Diffusion weighted imaging (DWI). Some frequently used modalities are shown in Figure 1.1.

MRI is powerful to detect anomalies of the brain and spinal cord, injuries or abnormalities of the joints. Especially in brain, MRI is capable to differentiate between white matter, grey matter and other tissues by combining several modalities, which is in favor of diagnosing lesions and tumors. The application of MRI in diagnosis of glioma is a typical example. However, efficiently reading raw MRI images by human eyes to obtain useful information demands specialized knowledge and rich experience. In recent years, researchers have been trying to facilitate the extraction of information from MRI by utilizing many image processing technologies. A basic step for such processing is semantic segmentation.

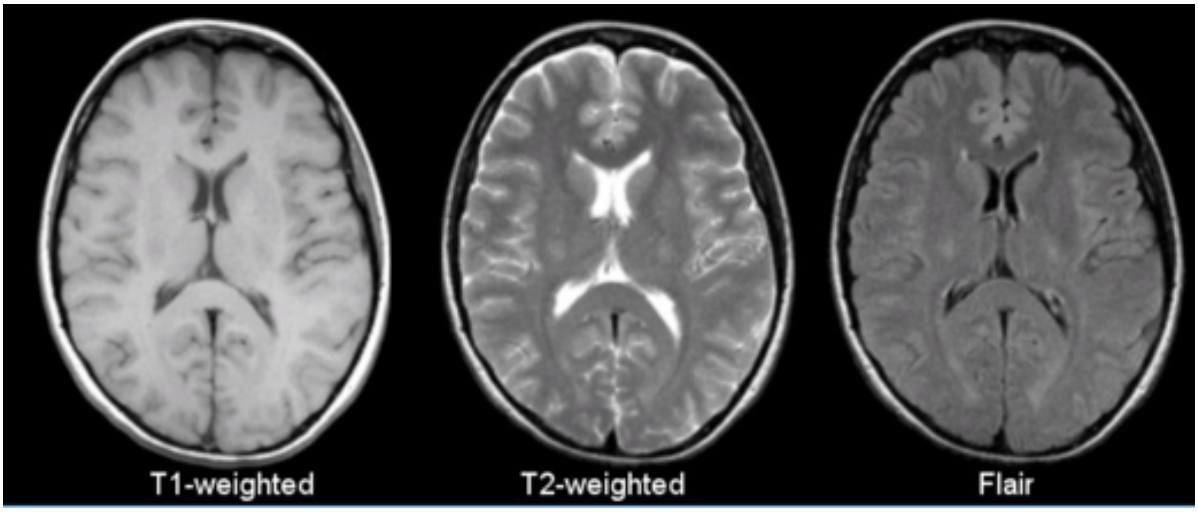


Figure 1.1: Three types of modalities of a normal brain MRI image[5].

1.2. Image Segmentation

Image segmentation is the process to assign labels to each pixel where the regions sharing the same characteristic or having identical semantic meaning for the given problem will be given the same label. Two subdivisions of image segmentation task are semantic segmentation and instance segmentation, respectively. The former one classifies the pixels to the same categories while the later one provides more sophisticated results in terms of classifying multiple objects of the same class. In medical image segmentation tasks, we mainly focus on semantic segmentation. Clear segmentation and distinguishable regions are essential for further analysis as they involve the determination of homogeneity levels of texture or layer thickness, as shown in Figure 1.2. Conventionally, the gold standard of medical image segmentation is the manual delineation created by clinical experts. However, manual segmentation is quite time-consuming and requires specialized knowledge on certain domains, leading to high human labour cost. Moreover, labelling process is subjective and error-prone. Different experts may give different results based on their level of knowledge and experience, resulting in a high variance segmentation results.

An alternative method is automatic image segmentation. Classical approaches are dominated by graph theory based(normalized cut[32], graph cut[12], etc.) and clustering based methods(K-means[24], watershed segmentations[34], etc.). However, the results of such classical machine learning based methods depends on the quality of artificial feature engineering and introduce many additional constraints, which limits the efficiency. Recently, with the soaring computation capability of hardware like CPU and GPU, one branch of machine learning study, Deep Neural Network (DNN), has shown its incredible generalization ability and outstanding performance in various fields. Because of its supremacy of modeling complex nonlinear relationships between variables compared to conventional algorithms, DNN dominates the applications in computer vision field. Medical image segmentation is also one sub-field which benefits a lot from this advanced technique.

1.3. Problem Statement

The diagnosis of low grade glioma usually requires longitudinal analysis for the evaluation of development of normal tissues and tumors over time. Since the morphological changes of tumor are interacting with other normal tissues, separately segmenting each target and then combining them together to build up a whole brain segmentation would be inefficient and bias-prone. On the other hand, it is assumed that the status of

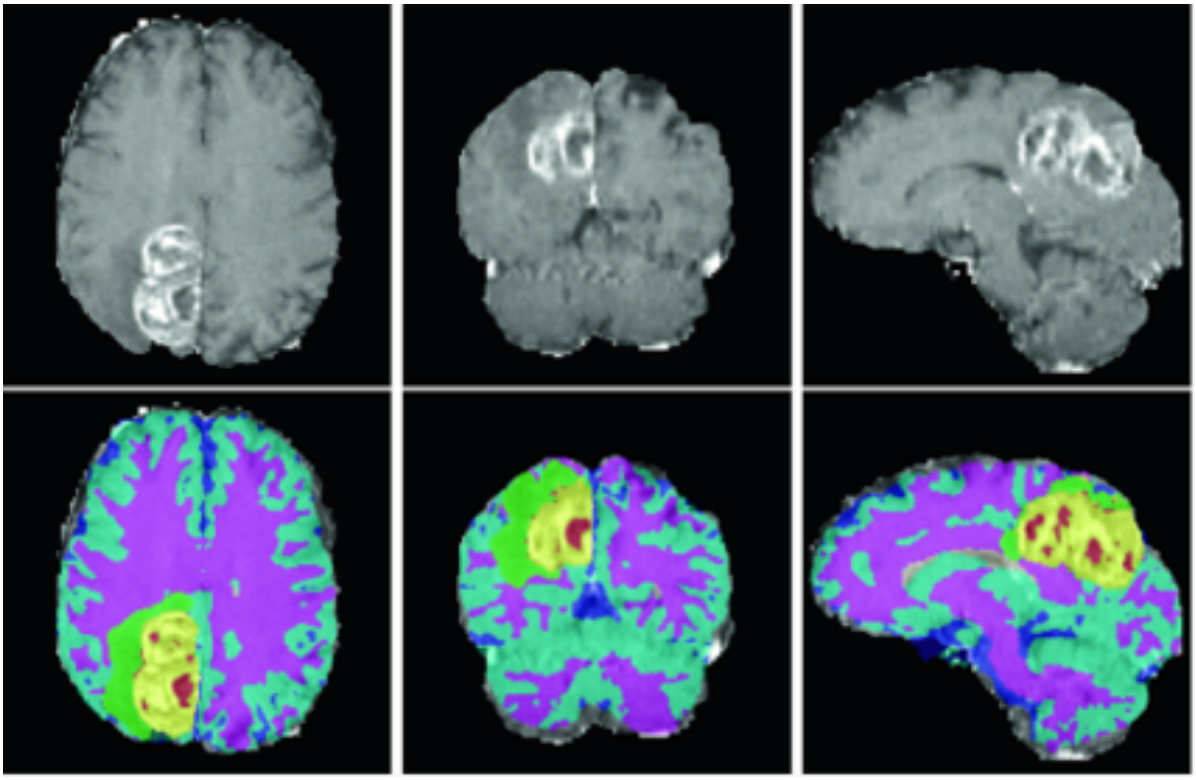


Figure 1.2: A 3D MRI image of brain from three orthogonal directions. The upper row is raw image, and the bottom row is semantic segmented results. Different colors represent distinct tissues[19].

brain tissues should keep stable within a short time period since the growth rate of low grade glioma is very slow. Concerning these two reasons, a longitudinal, multi-target segmentation method is desired to be developed for better brain segmentation results.

Multi-target segmentation has been widely used in image segmentation field and proven to be effective. The main goal of this research is to examine whether the cerebral longitudinal information is helpful to improve the semantic segmentation of MRI brain in terms of both accuracy and consistency. The dataset used in this project is a series of chronological MRI brain images of low grade glioma patients. Every patient has several MRI scans and experienced one or more times brain operation to remove the tumor. The segmentation masks are generated automatically by HD-GLIO[4] and FAST[40], instead of "golden standard" manual delineation from experts. Therefore, they maybe not that close to the real ground truth and probably contain artefact errors. As a result of utilizing these masks as training labels, the approximation to masks could reveal the capability of model but may not necessarily lead to the conclusion that the prediction is close to the ground truth. Compared to creating a powerful model to optimally mimic the fake mask examples, we are more interested in a result that can better approach truth. In other words, the difference between masks and model prediction is not always a bad thing since it is possible for the incorporated longitudinal information to remove the unreasonable errors existing in masks. The judgement of whether prediction is better than mask, however, can only be conducted by professional experts and will not discussed in this thesis. As stated before, the reason to take longitudinal information into consideration as a constraint of creating models is based on the fact that brain tissues should keep consistent over time. Therefore, conventional accuracy-only metric is not enough to evaluate the model's performance. We need more metrics which can reflect the temporal consistency level of segmentation, as a supplement of evaluation on the approximation to truth.

The research goal can be transferred into answering the following sub-questions:

- How to design longitudinal multi-target segmentation model with limited memory usage?
- How to evaluate the segmentation consistency?
- Can longitudinal information help improve segmentation accuracy of each target region?
- Can longitudinal information help improve segmentation consistency of each target region?

1.4. Thesis Overview

The thesis is composed of five themed chapters. Chapter 2 presents the historical work done in medical image segmentation field, which includes both CNN based and combined CNN and RNN approaches. The third chapter is concerned with the methodology and experiments setup used for this study. In particular, three different created 4D segmentation networks will be illustrated after the explanation of several 3D CNN variants. The experiment conditions are introduced in details as well. In chapter 4, results derived from multiple experiments are provided and the analysis of capabilities and performance of the proposed models is conducted. Finally, the findings, conclusions and future work will be discussed in chapter 5.

2

Related Work

As mentioned in last chapter, medical image segmentation is a fundamental but important step for further diagnosis. Numerous deep learning models have been developed in recent years in hope of obtaining accurate segmenting results. Since DNN is a kind of general method, the scope of the literature review was expanded to all types of the medical image segmentation problem, instead of being limited to MRI brain image, with a special focus on consistent segmentation solutions. In previous work, enforcing the segment consistency can be achieved by integrating segmentation with traditional algorithms, or incorporating RNN into CNN backbone. This chapter will give a gross introduction to both strategies.

2.1. Deep Learning Incorporated with Traditional Methods

2.1.1. Image Registration

Image registration is the process to align several images to an identical coordinate system by finding an optimal spatial transformation. Given a couple of images, one of them is treated as a fixed image, and the others are moving images(target images). By iteratively compare the moving images to the fixed image, optimal feature correspondence is expected to obtained under certain similarity measurement. If the target images are in a time series, some temporal feature relationship in neighbouring images would be built up. This technology can be combined with segmentation either to improve the latter's performance or optimize simultaneously. According to the role played by image registration, the integration of segmentation and registration can be classified into three types: 1. registration based segmentation; 2. registration as preprocessing of segmentation; 3. fused segmentation and registration.

Registration Based Segmentation Given an atlases image and corresponding segmentation labels, target images are aligned with atlas by registration, and labels are estimated by the label fusion of warped atlas. Consequently, the segmentation relies more on prior knowledge from atlases but less on training, with more smoothness constraints applied. Hu et al.(2018) proposed a label-driven correspondence learning framework, achieving label propagation by warping a generated dense displacement field with available labels of moving images to match their corresponding counterparts in the fixed image [22]. Liang et al.(2019) created a CompareNet consisting of a classification subnet, a features embedding layer and a label fusion subnet[27], as shown in Figure 2.1. First, a warped atlas images is generated to feed classification subnet for predicting a unary potential of segmentation. The feature embedding layer embeds deep features of the target and atlas image for the label fusion sub-net to produce a pairwise potential of segmentation, and the final segmenta-

tion is the weighted sum of unary and pairwise results.

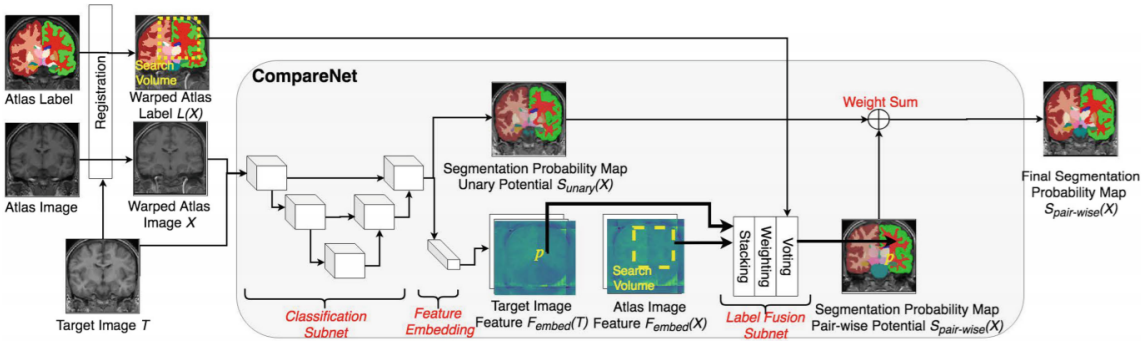


Figure 2.1: Overview of CompareNet. The atlas image and label are aligned to the target image T before CompareNet[27].

Registration as Preprocessing of Segmentation This approach takes advantage of image registration technique to generate the prior information as the input of deep network while the training pipeline is independent from atlas images in the following stages. For instance, Wu et al. (2019) [35] proposed a joint multi-atlas guided 3D+2D hybrid network to better learn the boundary features, as shown in Figure 2.2. The preprocessing comprises histogram matching, affine registration and large deformation diffeomorphic metric mapping (LDDMM). The 3D patches together with preprocessed atlas patches are introduced into 3D multi-atlas network as prior information to produce structure-specific 2D probability maps in three orthogonal views; then a 2D attention U-Net takes the 2D slice from each view respectively to generate final segmentation results. Vandewinckele et al.(2019) integrated deformable image registration(DIR) and CNN for the longitudinal CT scans segmentation[33]. The previous CT scans and corresponding segmentation are aligned to current scans and segmentation prediction to yield deformed images and deformed segmentation. These generations with current scans and segmentation are used as the input of a four-layers CNN, which predicts the final segmentation of current scans.

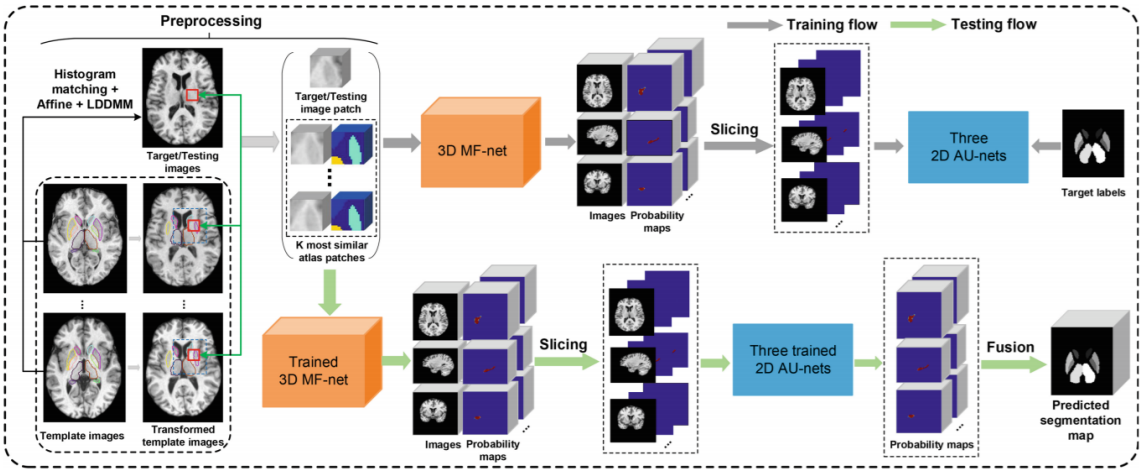


Figure 2.2: Illustration of the joint 3D+2D segmentation pipeline[35].

Fused Segmentation and Registration The final integration type tries to optimize the registration and segmentation jointly. In this way, image registration results could benefit from segmentation mutually. In often cases, the co-training of two parts is achieved by introducing joint loss function. A typical example is DeepAtlas, proposed by Xu et al.(2019), which consists of a weakly-supervised registration learning and a semi-

supervised segmentation learning[37], as shown in Figure 2.3. The registration net takes moving images and target images as input to yield displacement field, which is supervised by a regularization loss, and warped moving images, that are supervised by the intensity similarity loss. Segmentation net takes the same inputs but generates moving segmentation and target segmentation, supervised by segmentation loss. The moving segmentation is warped with the displacement field from registration net to produce a warped moving segmentation and penalized by the anatomy similarity loss. These four loss functions are jointly optimized to realize the co-learning of registration and segmentation. Li et al.(2019) also proposed a jointly training hybrid network for the longitudinal consistency analysis[26], where the improvement of registration and segmentation is achieved by training a single CNN supervised by four optimization goals simultaneously: segmentation accuracy, similarity between registered images, deformation field smoothness and segmentation consistency.

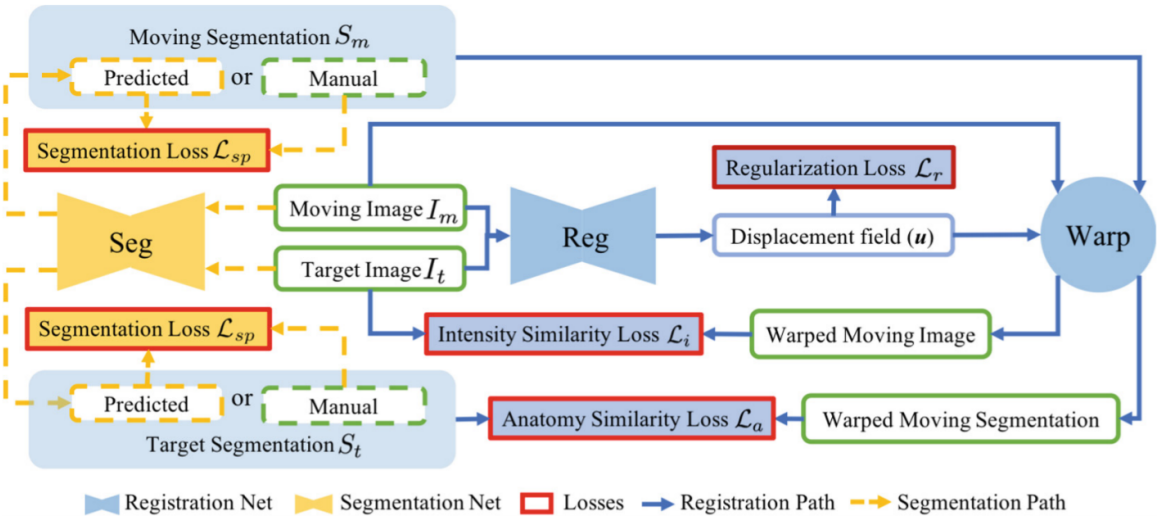


Figure 2.3: Workflow of DeepAtlas for joint learning of weakly supervised registration and semisupervised segmentation[37].

2.1.2. Conditional Random Field(CRF)

Conditional random field is a powerful probabilistic model used in image segmentation field. The key idea behind CRF is to formulate the segmentation problem as the spatially and temporally neighbouring label probability inference problem. CRF can be applied directly for segmentation like in [10], where a 4D CRF non-parametric growth model incorporating growth and inclusion constraints is proposed to obtain brain tumour segmentations and detect tumor regrowth in longitudinal sequences; or, in more cases, combined with deep network to refine the segmentation results. The framework of this work is presented in Figure 2.4. Deeplab[14] is a famous model which incorporates CRF into the final layer of a CNN as a postprocessing part to fine tune the boundary localization segmentation. Zheng et al. (2015) improved the combination strategy by implementing the CRF as recurrent neural network, which will be talked in next chapter, plugged in as a part of CNN[41]. However, both of these two models are used for semantic segmentation of natural image.

2.2. CNN Combined with RNN

Recurrent Neural Network(RNN) is known as its specialization at dealing with sequential data. Taking advantage of this property to enhance the segmentation consistency and smoothness has been investigated by many researchers. Chen et al.(2016) connect the RNN structure at the out of U-Nets to leverage 3D image anisotropism[13]. They proposed a kU-Net structure as the backbone to extract feature map from original image. kU-Net is a successive connection of submodules of FCN where each submodule take a down-scaled image after max pooling as input, i.e., the k-th submodule has an input after k-1 times max pooling while the

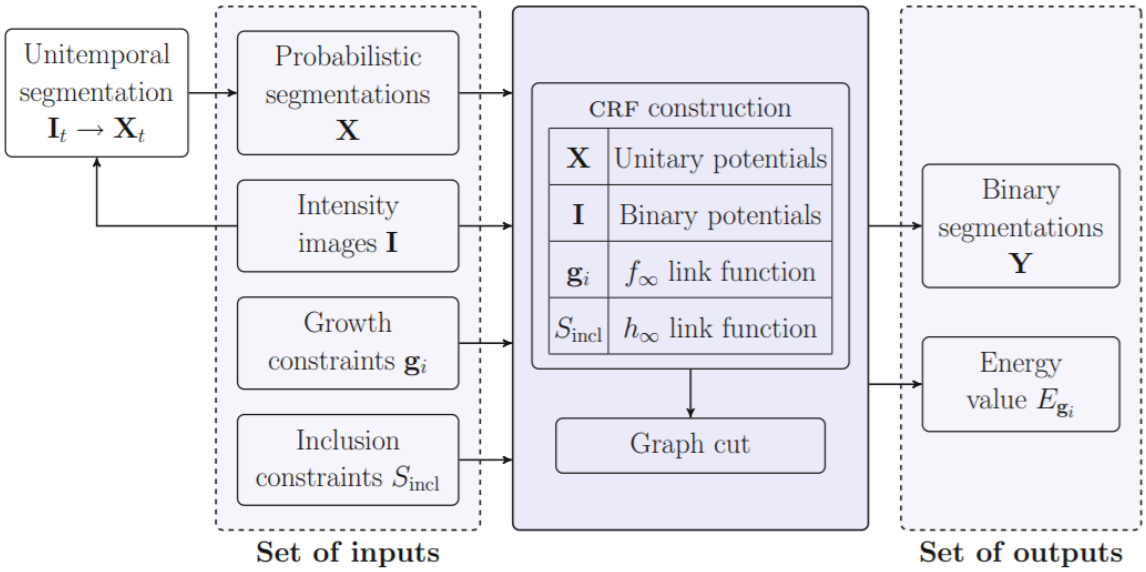


Figure 2.4: Overview of a 4D CRF as a Nonparametric Growth Model (NPGM)[10].

first one has the original resolution.The k value is chosen as 2 in the papar. The propagation of information between modules is achieved by fusing the the output from previous U-net into next one as input. The outputs of kU-Nets are extended to a stacked bidirectional ConvLSTM network. Like the FCN structure, some operations such as convolution, pooling and deconvolution are used between BiLSTM layers to configure "deep" architecture. The proposed network is shown in Figure 2.5.

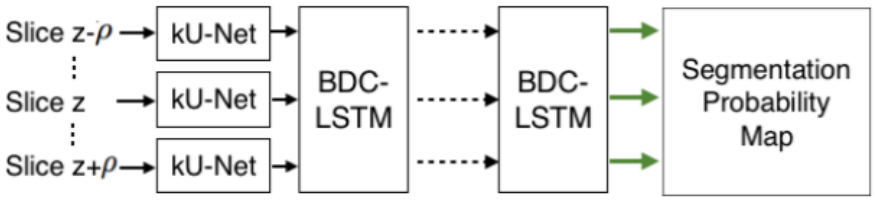


Figure 2.5: The framework of back-connected LSTM proposed by Chen et al. BDC-LSTM refers to bi-directional convolutional LSTM[13].

Apart from connecting the RNN at the end of CNN network, the combination can also take place within the body of CNN backbone. Poudel et al.(2016) [30] place the RNN component at the bottom of U-Net backbone when doing sequential cardiac segmentation since they thought the end of contraction path is equipped with the most compressed global context. The RNN unit they chosen is GRU for sake of less computation load. Azad et al.(2019) [11] used BiLSTMs at each skip connection, for each head of BiLSTM receiving the copied feature map from encoding path and up-sampled feature map from decoding path, respectively, as shown in Figure 2.6. In this way, the local and global spatial information could be better fused than a simple skip connection. Apart from than, they introduced a dense connection at the bottom of U-Net to learn a diverse of feature set more efficiently, which idea is analogous to layers in DenseNet[23]. However, this architecture can only improve the segmentation quality for single 2D slice.

Novikov et al.(2018) combined the middle-connected and back-connected strategy together for segmenting vertebrae and liver in 3D CT scans [29]. Two groups of Bi-Conv LSTMs are placed at the bottom and after the

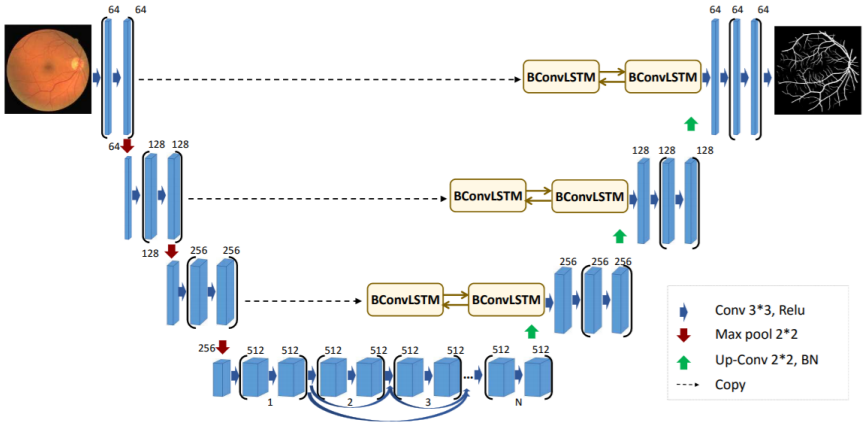


Figure 2.6: Architecture of BCDU-Net with bi-directional ConvLSTM in the skip connections and densely connected convolution[11].

output of U-Net backbone, respectively. The middle connected one is designed for adding explicit dependency of the low-dimensional high abstract features whilst the back connected one for the high-dimensional high-abstract features. With only 3 slices per sequence as input, this method can achieve comparable or even superior performance with strong generalization capacity compared to other state-of-the-art proposals. In semantic video frame segmentation field, the fusion of FCN and RNN is also a common method. The training framework is shown in Figure 2.7

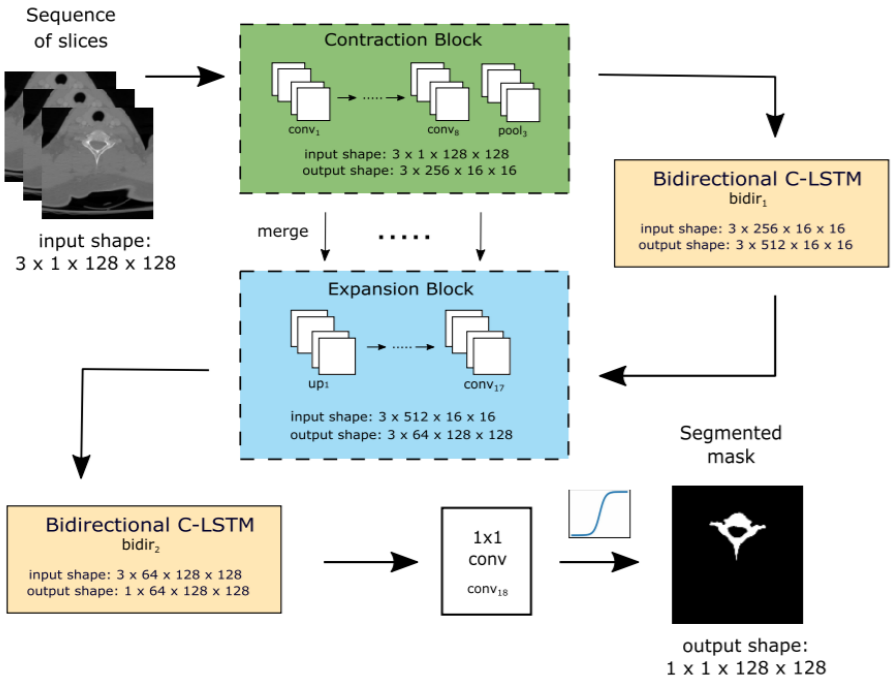


Figure 2.7: The framework of 3D sequential segmentation network proposed by Novikov et al[29].

Gao et al.(2018) proposed a fully convolutional structured LSTM Networks(FCSLSTMs) by integrating FCN layers into ConvLSTMs [18]. Starting from the ConvLSTM, the authors replaced convolution operator at each gate by more complex operators. These operators come from a lightened FCN model modified from VGG-16 structure where the number of feature map channels declines without sacrificing much accuracy. The

convolution layers in this FCN are used as the complex operator in replacement of normal dot product at each 'gate' in stacked LSTM layer by layer, i.e. conv1 is added to the first layer of LSTM network and so on. In this way, a structured network will be built up through stacking LSTM layers. The structure of FCSLSTMs is shown in Figure 2.8.

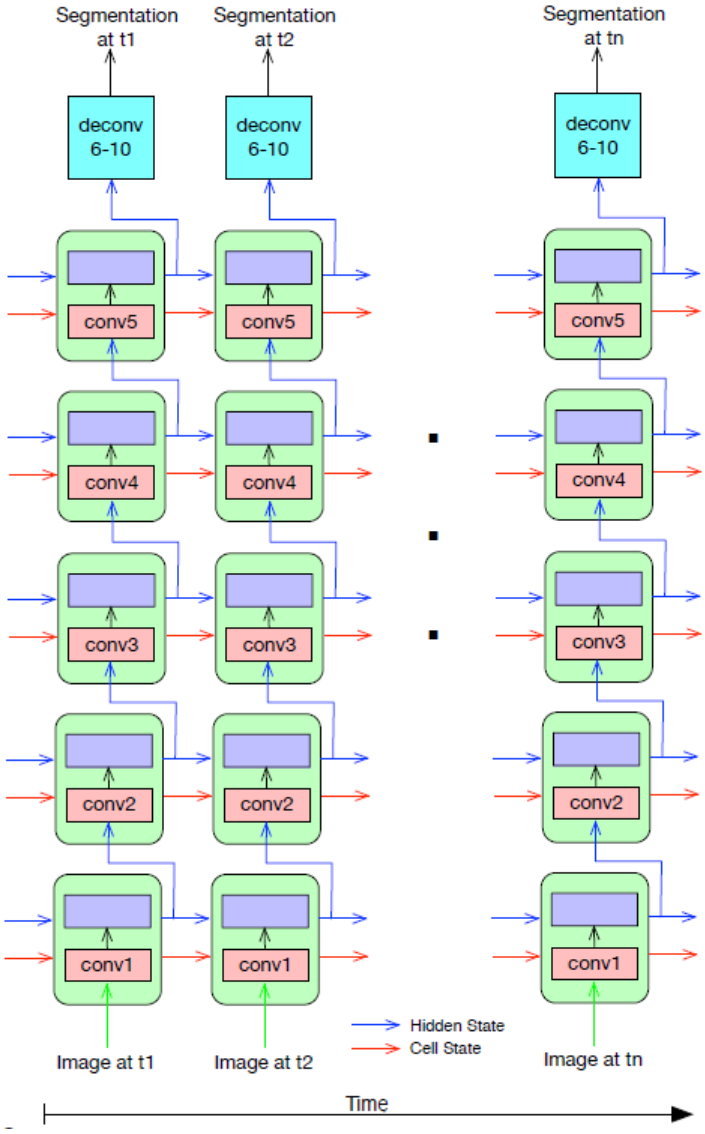


Figure 2.8: The overview of FCSLSTMs[18].

3

Methodology

Most existing work concentrates on segmenting either 2D or 3D image at each time step individually. In fact, the development of brain tissues and glioma is considered to be gradual and in consistent chronological order. Despite the spatial features relationship within an individual image, the constraints brought by this temporal property give another prior estimation on the shape of interested region from previous morphological characteristics. The main aim of this study is to investigate whether the longitudinal information will help improve the 3D segmentation quality. In this chapter, the proposed 4D longitudinal segmentation models will be illustrated, together with its building blocks. After that, the experiment setup will be introduced.

3.1. Baseline: 3D segmentation models

Before the appearance of 3D CNN, the 3D images are processed by 2D CNN slice by slice along a certain axis, and then the results are concatenated together directly to reconstruct the 3D volume. Generally speaking, this approach sacrifices the integrity of 3D information. In 2016, the proposal of 3D U-Net [16] and V-Net[28] alleviated this trouble by using 3D convolutional manipulator as basic components of network. In this section, a detailed description of original 3D U-Net and its variants is given, which will be used as the baselines for the construction of 4D longitudinal models later.

3.1.1. Original 3D U-Net Architecture

3D U-Net, proposed by Özgün Çiçek et al. in 2015, has been frequently considered as a standard baseline in medical image segmentation problems due to its simple structure and effective performance. The model comprises of three parts, an encoder path, a bridge and a decoder path. The encoder path is a series of $3 \times 3 \times 3$ convolution blocks and max pooling operators with stride of two to narrow down the feature map size, meanwhile increasing feature map channels and the scope of receptive field. A batch normalization layer(BN), which normalizes the data distribution to speed up calculation, as well as a rectified linear function(ReLU) used for introducing nonlinearity are included in convolution blocks. In the contrast, the decoder path is a series of $3 \times 3 \times 3$ convolution blocks with transposed convolution operator with stride of two to recover feature map size and decrease feature channels. The bridge at the end of encoder path consists of only convolution blocks with stride of one, which is responsible for connecting above two opposite process. At each convolution level, an identical mapping concatenates the convolution blocks from encoder to decoder to fuse the extracted features with recovered features. This manipulation is proven to be effective at increasing the segmentation accuracy. At the end of decoder, a $1 \times 1 \times 1$ convolution layer and a sigmoid activation

layer is used to project the multichannel feature maps into the desired segmentation. The architecture of the original 3D U-Net is shown in Figure 3.6.

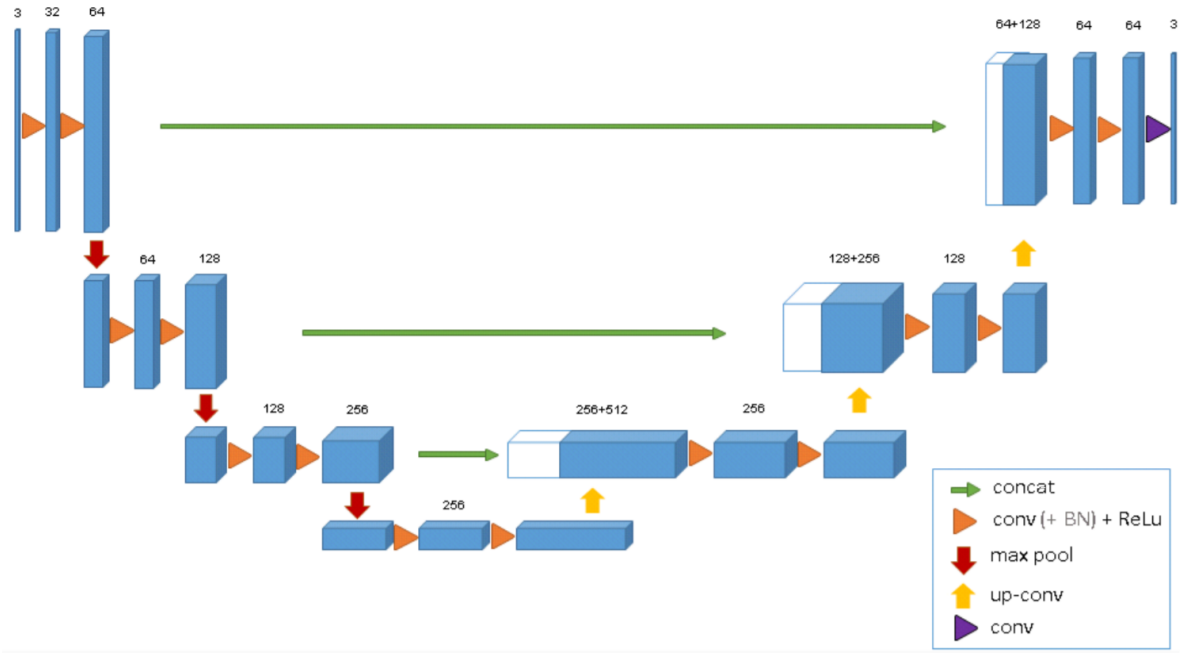


Figure 3.1: Original 3D U-Net structure. Retrieved from Özgün Çiçek et al[16].

In this project, the original 3D U-Net was adopted as the basic model for evaluation. The detailed parameters of each component are presented in Table 3.1.

Table 3.1: Parameters of 3D U-Net baseline

	Conv level	Conv layer	Filter / Channels	Stride
Encoder	Level 1	Conv 1	3 x 3 x 3 /16	1
		Conv 2	3 x 3 x 3 /16	1
		Maxpooling 1		2
	Level 2	Conv 3	3 x 3 x 3 / 32	1
		Conv 4	3 x 3 x 3 / 32	1
		Maxpooling 2		2
	Level 3	Conv 5	3 x 3 x 3 / 64	1
		Conv 6	3 x 3 x 3 / 64	1
		Maxpooling 3		2
Bridge	Level 4	Conv 7	3 x 3 x 3 / 128	1
		Conv 8	3 x 3 x 3 / 128	1
Decoder	Level 3	Trans Conv	3 x 3 x 3 / 128	2
		Conv 9	3 x 3 x 3 / 64	1
		Conv 10	3 x 3 x 3 /64	1
	Level 2	Trans Conv	3 x 3 x 3 / 64	2
		Conv 11	3 x 3 x 3 / 32	1
		Conv 12	3 x 3 x 3 / 32	1
	Level 1	Trans Conv	3 x 3 x 3 / 32	2
		Conv 13	3 x 3 x 3 / 16	1
		Conv 14	3 x 3 x 3 / 16	1
Output		Conv 15	1 x 1 x 1 / 5	1

3.1.2. 3D Res U-Net

The depth of network was considered as an important factor influencing the feature expression and extraction. However, with increasingly stacked layers, the vanishing of gradient during propagation becomes severe, resulting in degraded performance of network. To alleviate this problem, He et al. proposed the deep residual learning framework[21]. The main idea of this work is to perform identity mapping from the inputs to the outputs of stacked layers by shortcut connections skipping one or more layers, resulting in a residual block. The output of each layer is no longer the traditional mapping of input but the summation of input and the mapping. If the dimensions of mapping and input are different, a linear mapping of input will be applied to match the dimension before summation. This residual connection can be expressed mathematically in the following equation:

$$y = \mathscr{F}(W, x) + \mathscr{I}(x) \tag{3.1}$$

where $\mathscr{F}$ represents any nonlinear activation function and $\mathscr{I}$ is identical mapping. In this project, the residual connection concept was applied to basic 3D U-Net for creating a 3D Res U-Net network, as shown in Figure 3.2. There are three major different points in 3D Res U-Net compared to 3D U-Net: 1. All the convolution blocks are replaced by residual blocks. 2. Instead of using Max Pooling, 3x3x3 convolution with stride size of 2 is used to downsize the feature maps in encoder. 3. Transpose convolution is replaced by the interpolation up-sampling algorithm to reduce the parameter amount.

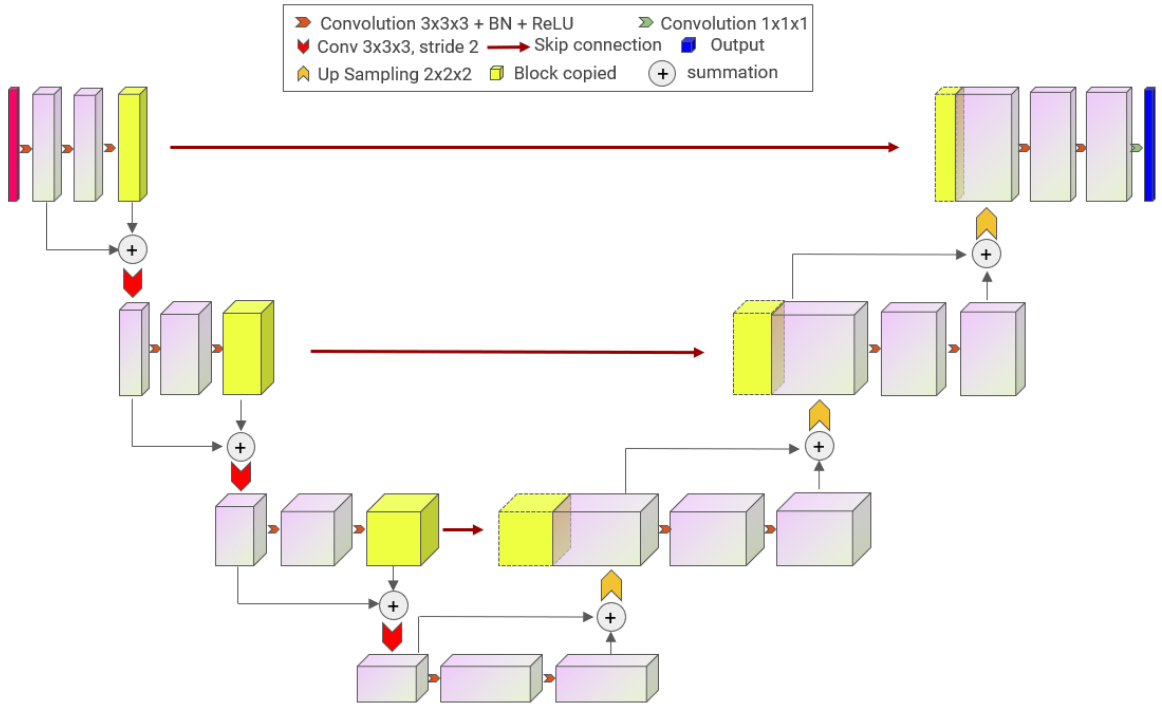


Figure 3.2: 3D Res U-Net structure.

3.1.3. 3D Dilation Res U-Net

A pooling layer, either Average Pooling or Max Pooling, is used for reducing resolution and enlarging the receptive field to integrate global contextual information. However, detailed local information will loss during successive pooling, which hinders the full-resolution dense prediction as segmentation output. Dilated convolution was proposed by Fisher et al. in 2016 to solve this problem [38]. Traditional convolution can be seen as the dilated convolution with dilation factor of 1. With exponentially increased dilation factor(1, 2, 4, ...),

the receptive field of an element in layer F_i is $(2^{i+1}-1)\times(2^{i+1}-1)$, for $i=0,1,...,$ if the filter size is 3×3 [38], as shown in 3.3. This dilation convolution can be stacked in parallel to aggregate multi-scale contextual information. Inspired by the work of Feng et al., the 3D dilation convolution group (shown in Figure 3.4(a)) was implemented at the bridge of 3D Res U-Net body to build up the 3D Dilation Res U-Net, as shown in Figure 3.4(b). The dilation part is split into four sub-paths, with increasing dilation factors.

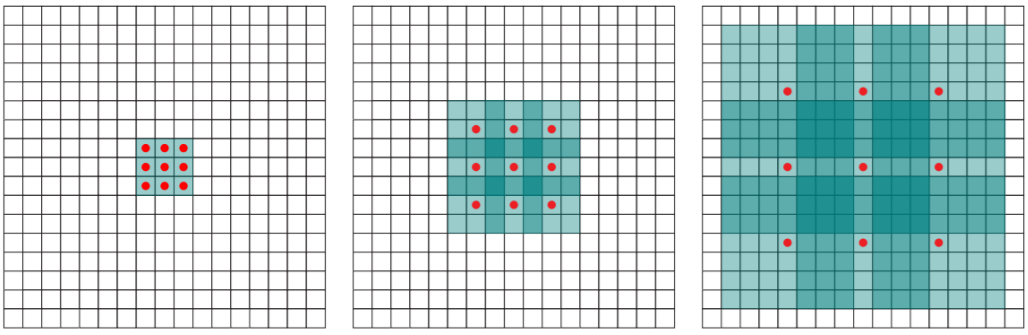
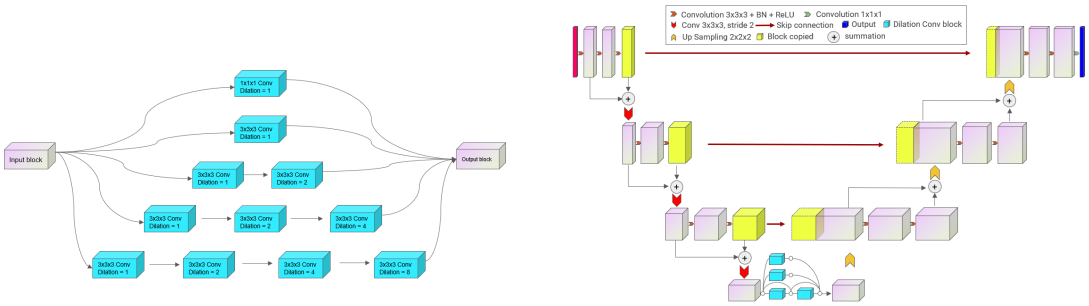


Figure 3.3: Dilation convolution with exponentially increased dilation factor. The receptive field in green color exponentially increases accordingly[38].



(a) Dilation Convolution net

(b) Dilation convolution with exponentially increased dilation factor [38]

Figure 3.4: 3D Dilation Res U-Net. The left image presents the dilation sub-network used at the bridge of Res U-Net. The right image shows the overview architecture of 3D Dilation Res U-Net.

3.1.4. Direct Concat U-Net (DC U-Net)

The idea behind shortcut connection from down-sampling path to up-sampling path is to enforce the low level information in feature maps to fuse with the high level information, so that more details of original image can be restored during decoding. In lower convolutional layers, more feature location information are kept with less convolution operator; while abstract global semantic features are extracted in higher layers. In other words, the original input without convolution and activation should have the most local feature details. Motivated by this theory, a simple modification was made on the basic 3D U-Net by concatenating the input image at each level to the up-sampling path in addition to original shortcuts. To keep the mapping size equal and the purity of raw information, only Max Pooling is applied before concatenation on each level to down-scale the input size. The architecture of this slight modified network is shown in Figure 3.5.

The three U-Net variants are designed in purpose of evaluating how accurate an individual CNN network can approach the segmentation mask reference. Moreover, they can be used as the backbone of longitudinal models proposed in the following section. The comparison of performance between CNN models, as well as

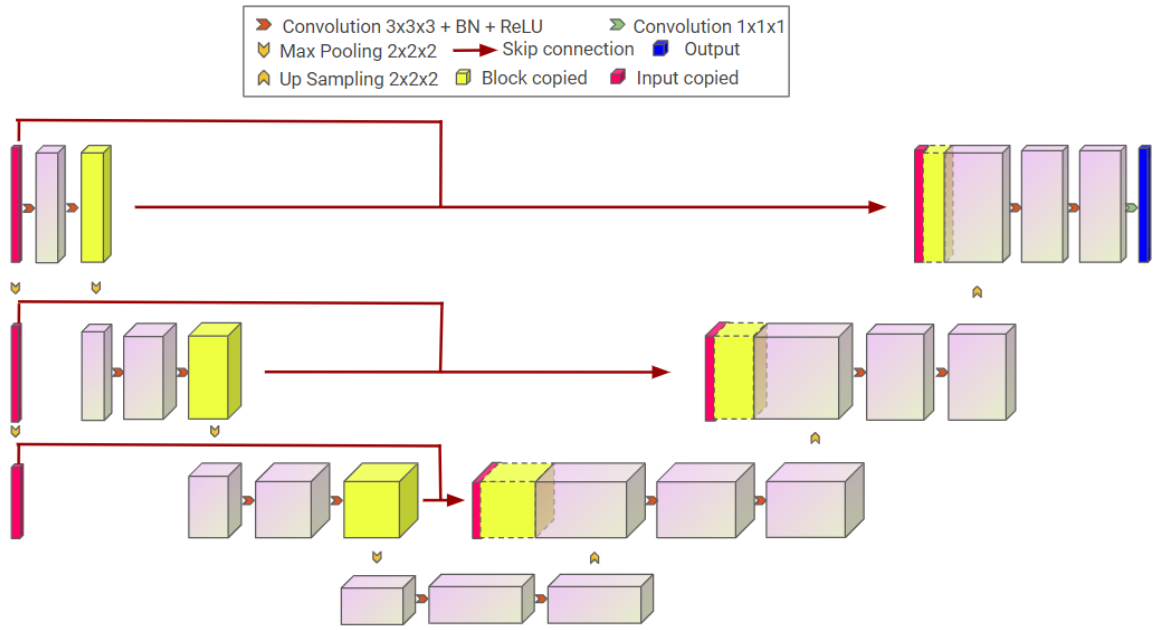


Figure 3.5: DC U-Net structure.

between CNNs and longitudinal models will be presented and discussed in the next chapter.

3.2. Proposed models: 4D longitudinal models

RNN is a type of Neural Network that allows previous outputs to serve as inputs in the next step while having hidden states. Due to this property, it is usually considered as the optimal candidate to process sequential data. The application of RNN in medical image segmentation is primarily combined with 2D CNN network to process slices or feature maps, as discussed in last chapter. Previous work indicates that introducing RNN framework appears to benefit the segmentation performance by incorporating the prior distribution knowledge existing in neighbour slices. This method is referred to 2.5D segmentation. However, information flow not only exists inside consecutive 2D slices, but also travels through temporal dimension. In the longitudinal framework, a single 3D image volume serves as the input at each time point, which is referred to so-named "4D segmentation". The focus of this project is attempting to utilize this temporal information to obtain a more close-to-reality segmentation result.

In this chapter, the basics of Long Short Term Memory Network(LSTM) will be introduced first, which is one of the most used RNN structure today due to its capability in solving gradient vanish or explosion problem of standard RNN. Based on LSTM networks, three types of longitudinal segmentation network are proposed and will be presented.

3.2.1. Long Short Term Memory Network

An individual LSTM unit consists of three gates: forget gate, input gate and output gate, and two states: cell state and hidden state. The architecture of a LSTM unit is shown in Figure. In the following mathematical expressions, h and c denote hidden state and cell state, respectively; W and b represent the weight and bias term, respectively.

- **Cell state:** the cell state maintains the filtered information along whole data flow.

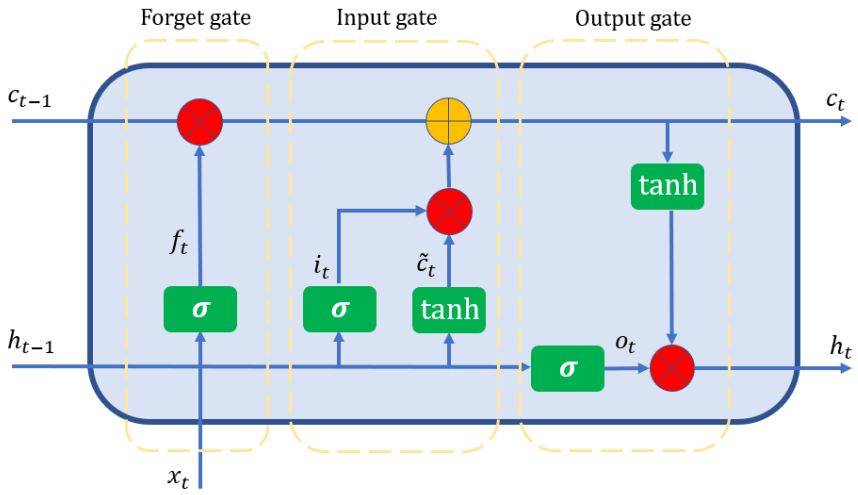


Figure 3.6: Overview of LSTM unit architecture. Figure adapted from [39].

- **Hidden state:** the hidden state provides the information from previous time step to next step.

- **Forget gate:** information from previous hidden state and present input will be activated by a sigmoid function σ to squash the value between 0 and 1 as the output of forget gate. This scaled value will be multiplied with cell state from last time step to determine how much memory should be dropped off from here. The smaller the value, the more discarded. Mathematical expression of forget gate at time t is:

$$f_t = \sigma(W_{fh}h_{t-1} + W_{fx}x_t + b_f) \tag{3.2}$$

- **Input gate:** input gate decides what new information should be stored in cell. As the manipulation in forget gate, the inputs from previous and current will be scaled into $0 \sim 1$ by sigmoid as the memory volume controller. The larger the value, the more new information will be extracted. In a parallel path, the two inputs are filtered by a tanh function and multiplied with memory volume controller simultaneously, leading to the new information which will be stored into cell state. Mathematical expression of input gate at time t is:

$$\begin{aligned} i_t &= \sigma(W_{ih}h_{t-1} + W_{ix}x_t + b_i) \\ \tilde{c}_t &= tanh(W_{\tilde{c}_t h}h_{t-1} + W_{\tilde{c}_t x}x_t + b_{\tilde{c}_t}) \end{aligned} \tag{3.3}$$

  After input gate and output gate, the cell state will be updated by dropping some old memory and adding some new information, which is mathematically expressed as following:

$$c_t = f_t c_{t-1} + i_t \tilde{c}_t \tag{3.4}$$

- **Output gate:** the output gate functions as a controller to make a decision that what information stored in current state will be transmitted into next unit. As before, sigmoid function is imposed on inputs to generate a scaling factor and cell state is activated by a tanh function. The product of this two parts defines the output of this unit, that is the current hidden state. Mathematically speaking:

$$\begin{aligned} o_t &= \sigma(W_{oh}h_{t-1} + W_{ox}x_t + b_o) \\ h_t &= o_t tanh(c_t) \end{aligned} \tag{3.5}$$

ConvLSTM Most LSTMs, or generally speaking, RNNs, are designed for dealing with temporally sequential data, like text and voice. To deal with spatial data, like image, convolution is more efficient than this type of fully connected structure(FC-LSTM). Therefore, convolutional LSTM (ConvLSTM) is proposed to extend capacity of traditional LSTM[36]. The primary advance in ConvLSTM is that the traditional inner products of parameters in FC-LSTM are replaced by convolution operator. As a result, corresponding expressions are modified into:

$$\begin{aligned} f_t &= \sigma(W_{fh} * h_{t-1} + W_{fx} * x_t + b_f) \\ i_t &= \sigma(W_{ih} * h_{t-1} + W_{ix} * x_t + b_i) \\ \tilde{c}_t &= tanh(W_{\tilde{c}_t h} * h_{t-1} + W_{\tilde{c}_t x} * x_t + b_{\tilde{c}_t}) \\ o_t &= \sigma(W_{oh} * h_{t-1} + W_{ox} * x_t + b_o) \end{aligned} \tag{3.6}$$

LSTM networks In addition to various inner structure of LSTM unit, the connection scheme of network is also a research topic. Since one-way LSTM usually cannot meet some complex requirements, some large scale networks are therefore proposed. The most common architectures are stacked LSTM and bidirectional LSTM, which are adopted as the network scheme in this project:

- **Stacked LSTM:** In deep learning, the simplest way to add depth of network is to stack layers. The Stacked LSTM is an extension to a single-layer LSTM that has multiple hidden layers where each layer contains multiple memory cells. The hidden state from previous layer will be the input to the next layer. The temporal information flow in each layer do not have interaction. Figure 3.7 presents the architecture of a M-layer stacked LSTM network with N time steps. $h_{t_n}^m$ and $c_{t_n}^m$ refer to the hidden state and cell state at time step t_n of layer m, respectively, where $n = 0, 1, 2, \ldots, N$, $m = 1, 2, \ldots, M$

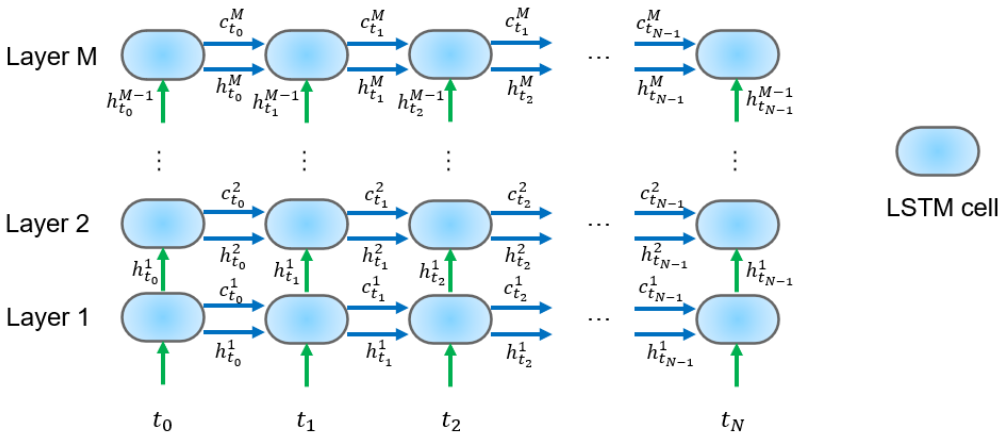


Figure 3.7: Architecture of stacked LSTMs.

- **Bidirectional LSTM (BiLSTM):** Typical one directional LSTM only take advantage of previous inputs and current input during processing, while for sequential input data future information may also be useful for analyzing current context or prediction. Bidirectional LSTM, introduced by Schuster and Paliwal(1997), can be trained using all available input information in the past and future of a specific time frame by configuring a time forward LSTM layer and backward layer [31]. The input at each time step will be given to both forward pass and backward pass. The architecture of a Bidirectional LSTM network is shown in Figure3.8, where $\overrightarrow{[]}$ and $\overleftarrow{[]}$ indicates the direction of states. Other notifications keep the same meaning as in stacked LSTMs.

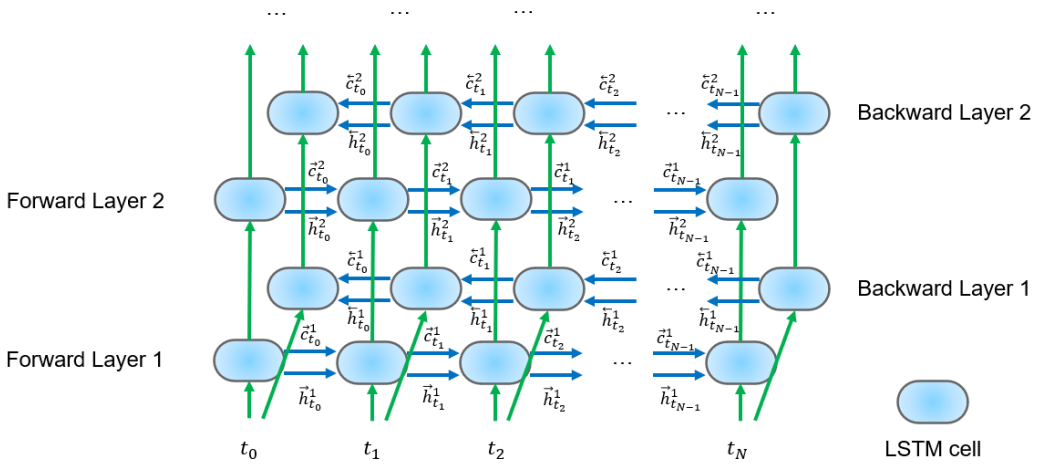


Figure 3.8: Architecture of stacked Bidirectional LSTMs.

3.2.2. Back-Connection longitudinal model

From the longitudinal viewpoint, the individual image within a series of chronological images is an input unit of LSTM. In [13], Chen et al. took the processed slices at the input of LSTM network to achieve 2.5D segmentation, where the output is a reconstructed 3D image from 2D slices. Inspired by this idea, the Back-Connection longitudinal model can be created as a 4D segmentation paradigm. The 3D U-Net, or any other 3D CNN network is adapted as the backbone to process the input 3D image in chronological order, and then the output layer of backbone is connected to 3D ConvLSTM networks in purpose of making communication between consecutive time points, which functions like a post-processing to smooth the the segmentation results over time. The LSTM part can be either stacked ConvLSTM or Bidirectional ConvLSTM. The architecture of Back-Connection longitudinal model is shown in Figure 3.9.

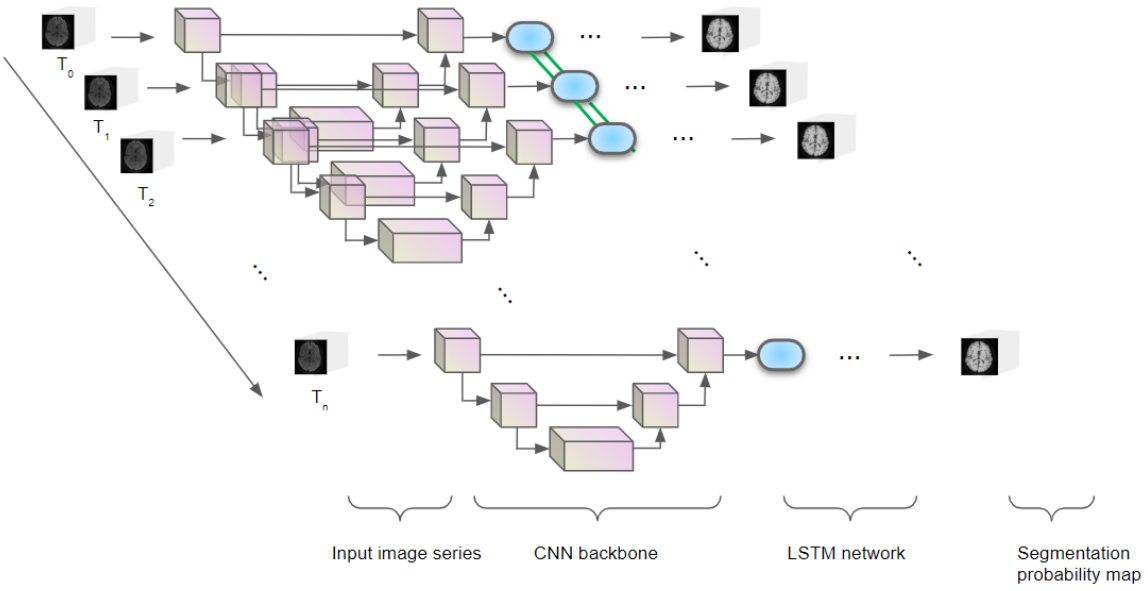


Figure 3.9: Architecture of Back-Connection longitudinal network.

3.2.3. Intermediate-Connection longitudinal model

From existing work, the temporal information communication can take place among a certain layer of CNN backbone as well[30]. Extended to a global longitudinal framework, 3D encoder-decoder networks are adopted as backbones and the bridge components are replaced by LSTM units to connect them together. The idea behind this manipulation is that the communication between high level features in bridge layers contain more general morphology information which might contribute to the improvement of segmentation consistency in a higher level. The architecture of Intermediate-Connection longitudinal network is presented in Figure 3.10.

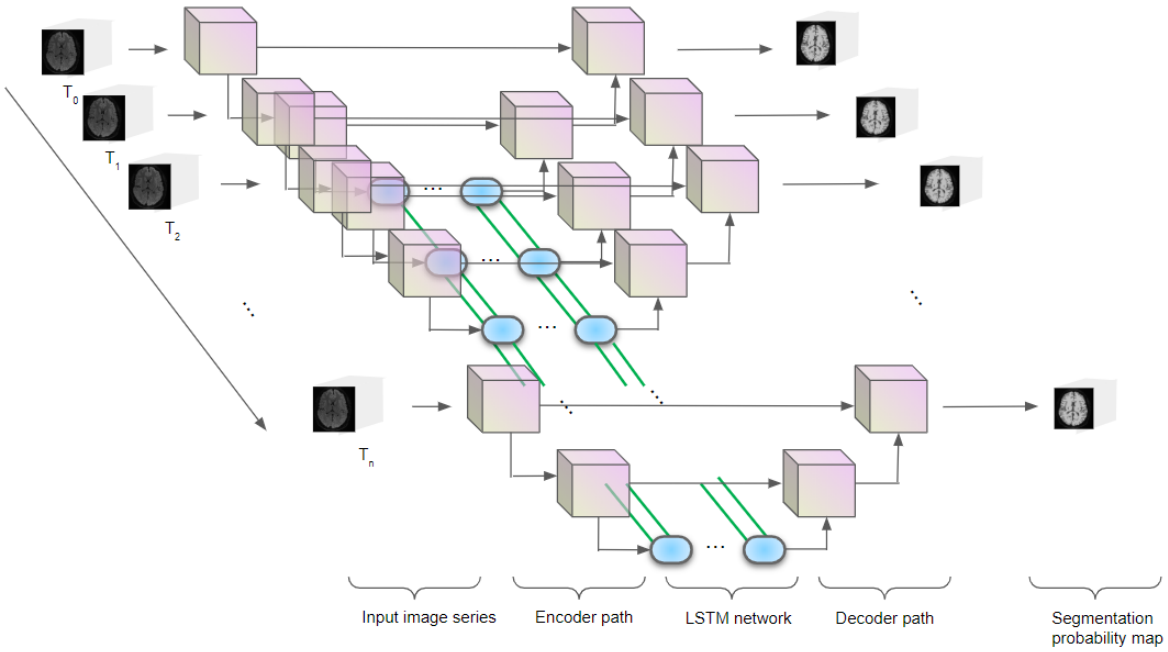


Figure 3.10: Architecture of Intermediate-Connection longitudinal network.

3.2.4. Shortcut-Connection longitudinal model

The shortcut connection between decoder and encoder path of U-Net is thought to help fuse the low level feature and high level feature to increase the segmentation accuracy. It is therefore reasonable to hypothesise the concatenation layers contain most abundant information of both sparse and dense features. To explore whether this mixed feature map could increase the communication effectiveness, the LSTM unit is inserted right after concatenation manipulation, that is, take the concatenated feature map as input of each LSTM unit and feed the output to the rest of up-sampling path. At the same time, the output is transferred to other time steps for the filtering of useful information. The insertion level and the number of insertion can be chosen arbitrarily, depending on the convolution level of backbone and the capacity of hardware. The structure of Shortcut-Connection longitudinal model is shown in Figure 3.11.

3.3. Experimental Settings

3.3.1. Dataset

The dataset used in this project is provided by Biomedical Image Group Rotterdam(BIGR) of Department of Radiology and Nuclear Medicine at Erasmus Medical Center. There are 79 patients with low grade glioma in total and each of them has multiple brain MRI scans, from at least 2 to at most 32 times. Due to the long

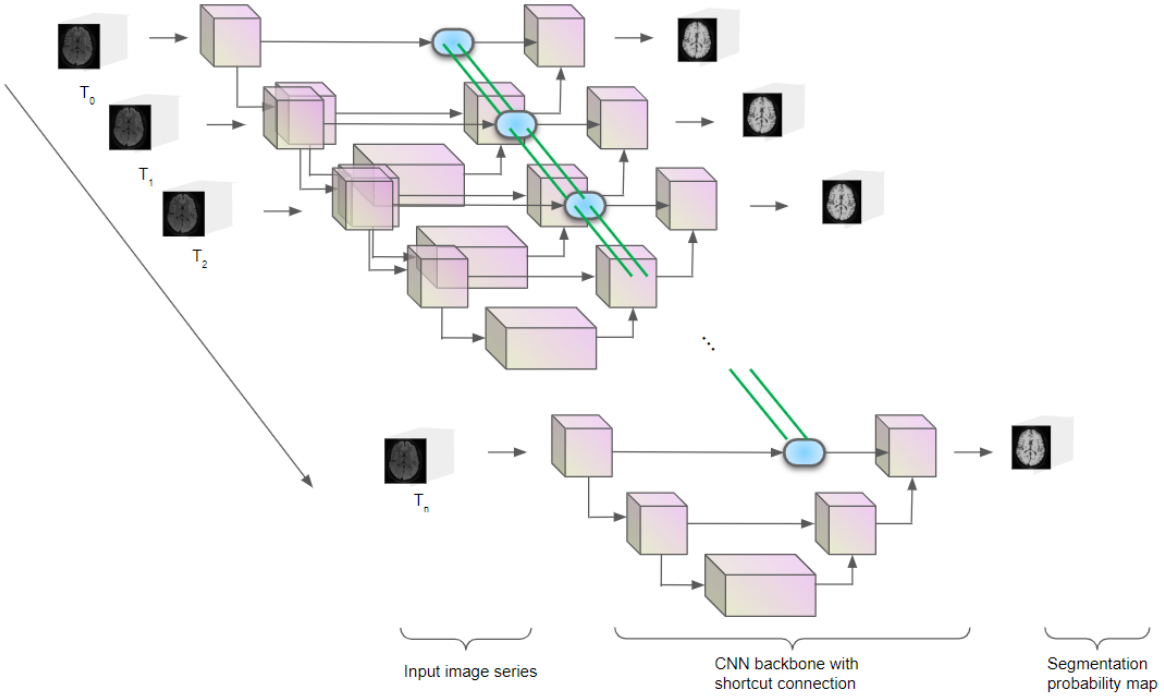


Figure 3.11: Architecture of Shortcut-Connection longitudinal network

time period, some old MRI images are in low resolution or even not intact. The MRI image series of each patient are co-registered in time dimension by Elastix[1] with only a rigid transformation. The time span of MRI scans per patient is different, from as short as within 1 years to as long as 16 years. The dataset was randomly divided into a training set and a testing set with the ratio of 9 : 1, in which the training set contains 72 patients and testing set contains 7 patients. The testing set retains untouched until evaluating the well-trained models to give final results. There might be zero, one or more times resections happened within this period. Each MRI images contains four MRI sequences: T1-Weighted, T2-weighted, T2W-FLAIR and T1-W + contrast images. The size of all the images in this dataset is $189 \times 233 \times 197$ voxels. The segmentation targets are background, CSF, grey matter, white matter and glioma, labeled with number 0, 1, 2, 3 and 4 in masks, respectively. The ground truth pixel-level marking masks used for training are generated in three stages. First, normal tissues(WM, GM and CSF) are segmented by FAST, an automatic image registration software based on a hidden Markov random field model and an associated Expectation-Maximization algorithm[40]. Second, tumor(TM) segmentation is obtained by HD-GLIO[4], a brain tumor segmentation tool developed by Heidelberg University Hospital, Germany and the Division of Medical Image Computing at the German Cancer Research Center (DKFZ) Heidelberg. Third, the mask of normal tissues and tumor are combined together. In addition, a brain mask generated by skull-striped and background subtraction with HD-BET[3] is provided under each patient folder to exclude background noises. Due to the long time span and unexpected imaging artefacts during scanning, some samples may look a bit different from others, which results in the inconsistent labeling. Moreover, since the masks are not generated from sophisticated delineation by experts, unreasonable errors are unavoidable. These are two primary defects in our dataset. A typical example data is presented in Figure 3.12.

3.3.2. Implementation Details

All the models were implemented in Pytorch framework[9] and trained with Adam optimizer[25] with initial learning rate 0.001, which will reduce by a factor of 0.1 once learning stagnates for 15 epochs. All learn-

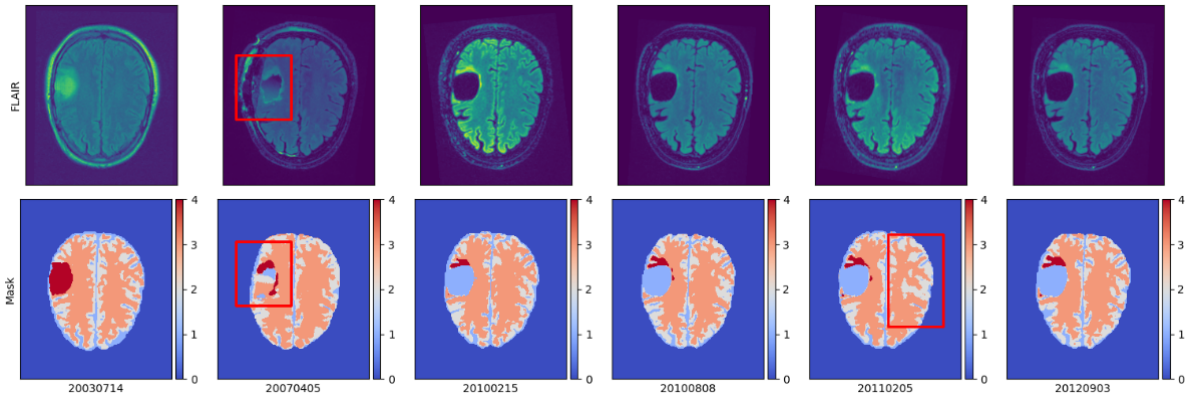


Figure 3.12: An example of patient data with 6 times MRI scans. The upper row is the FLAIR image modality and the lower row lists corresponding masks. The scanning date at the bottom is recorded in format "yyyymmdd". There was a resection during 2003-07-14 and 2007-04-05. After that, glioma grows up gradually again. Two types of defects is presented in this example: 1. A probable magnetic imaging artefact in 2007-04-05 results in inconsistent segmentations of tumor region compared to the following time points (marked by red squares); 2. Due to the limited accuracy of FAST, an unreasonable shrink of White Matter mask(marked by color label 3) in terms of area at time point 2011-02-06 can be noticed(marked by a red square).

able parameters, i.e., weights and biases of the models were initialized based on the Kaimin initialization method[20]. In each epoch, the training set will be divided again into training and validation subset with the ratio of 9 : 1, illustrated in Figure 3.13. This is a kind of data shuffle strategy to increase the sampling randomness during training. To save computational resources and avoid overfitting, early stopping strategy is used. The network will stop training if the evaluation metrics hasn't seen improvement for over 30 epochs. Data augmentation, including flip, rotation, Gaussian noise and elastic deformation, are applied to improve the generalization and avoid overfitting as well. The experiments are performed on Cartesius, the Dutch national supercomputer[6]. In GPU partition, each node is equipped with two NVIDIA Tesla K40m GPUs. To accelerate the training process and make more room for the gradients calculation on graph, distributed training strategy by splitting the input data batch into equally half size for each GPU is adopted. Training of CNN networks costs about 10 hours and longitudinal networks require at least 20 hours.

Different from normal 2D images, a major concern of 3D model training is the large memory cost. To fit the capacity of GPU, $64 \times 64 \times 64$ fixed-size cropped patches from original images are used for a mini-batch input as basic experimental condition, rather than the whole image. Another special point in this research is the input sampling strategy used for 3D and 4D models. Usually we shuffle the training data randomly to obtain a more generalized model, but in this project, the images are intra-subject, so a direct shuffle across subjects with random batch sampling will destroy the hierarchical data structure. Instead, the training and testing batch is selected based on patient unit.

Input for 3D models: The batch size used for training 3D models is set to 8, which means in each epoch 8 patients will be randomly chosen with a random time point MRI scan. The input dimension is hence $N \times C \times D \times H \times W$, where N refers to batch size(patient number), C is the number of modality types, in this case the value is 4. D, H, W are image depth, height and width of cropped patch, respectively. The patch crop of each image is randomly located.

Input for 4D models: The batch size used for training 4D models is set to 4, with at least 3 successive images are selected per patient. The cropped patch size is kept to be 64 , subject to memory restrictions of the GPU hardware. The MRI image series of each patient are sorted beforehand according to capturing time to ensure that the successive selection is in forward longitudinal order. As a result, the input tensor should be $N \times T \times C \times D \times H \times W$, where T is the time points and other notations are in consistent as before. The starting

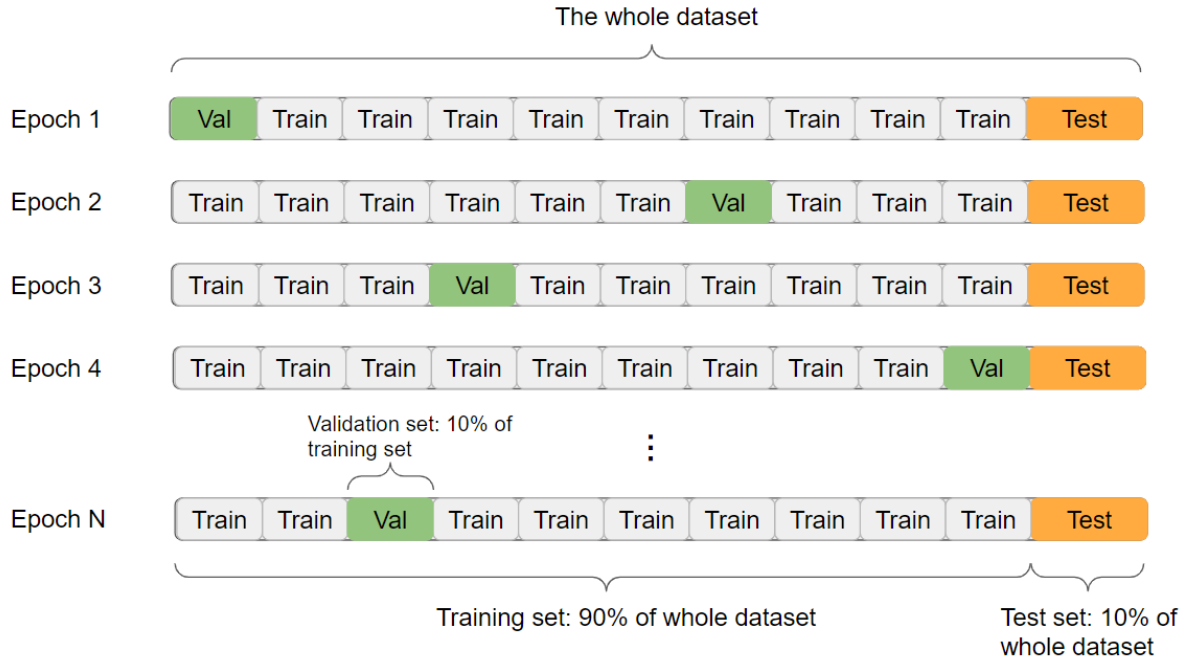


Figure 3.13: Illustration of dateset splitting strategy. 10% of the whole dataset is reserved as testing data and the rest is used in training phase(termed as training set in figure). In each epoch, random 10 % of training set is selected for validation.

time point of successive images is randomly selected in each epoch. The cropping location throughout all time points of the same patient is identical, while alters in different subjects. Compared to white matter, grey matter and CSF, the volume of glioma is quite small, which results in a low chance for a randomly-cropped patch to contain tumors region. To ease this trouble, the crop strategy is adjusted to increase the proportion of tumour-contained patch in the whole training phase. Worth to note, too high proportion will increase the bias to segment the tumor in testing phase. By evaluating several values, the threshold is empirically chosen to be 40%, which means at least 40% patches in each epoch definitely contain tumors while the others are randomly cropped, with a little probability to have tumour. A example segmentation results under different thresholds is presented in Figure 3.14.

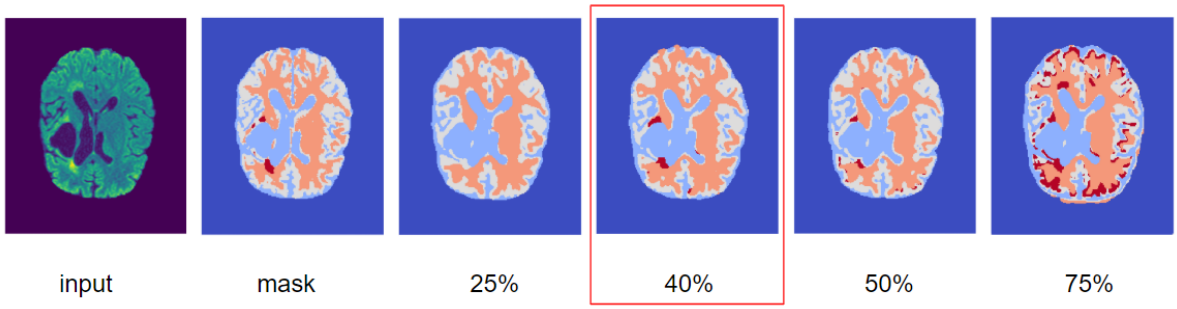


Figure 3.14: Segmentation results by 4D model under different proportions of tumor patches in a training epoch. If the threshold is less than 40%, tumor will be underestimated while overestimated if higher than 40%.

Before training of the longitudinal models, some preliminary experiments were conducted on the pure LSTM networks with single and bidirectional architecture (see Appendix A). The results show that the bidirectional ConvLSTMs with single layer provide better results than one-way stacked LSTMs, and repeating more stacked bidirectional layers does not necessarily improve the training effects but increases the memory cost. Thus,

the LSTMs component used in this project is one-layer bidirectional ConvLSTM network.

3.3.3. Pretraining strategy of longitudinal model

Although all the CNN networks are trained from scratch with identical parameters combination, the training strategy of 4D longitudinal models are different from each other. From preliminary experiments, training from scratch is not the optimal choice for 4D models. There are two major drawbacks of this strategy: 1. the parameter amount of longitudinal model increases linearly with the number of input time points, so a large amount of learnable parameters demands more GPU resources and required longer time to converge; 2. all three types of longitudinal networks failed to detect tumour in a series of input images, see figures in Appendix B. This phenomenon may attribute to the inconsistent size, morphology and location of glioma across time dimension. If the tumor information in last time step cannot help the segmentation in the following steps, the forget gate in LSTM will judge it as useless context and close corresponding weights response. To conquer these troubles, 4D networks are trained with pretrained weights learned from 3D network backbone. For back-connection model, all the parameters on CNN backbone are frozen after loading pretrained weights, leaving only convolution kernels in LSTM networks available for training. In the intermediate-connection model, not only LSTMs at bridge as the connection between backbones can be trained, the decoder path is also freed out to take part in the back propagation of gradients. The training of shortcut-connection network adapts similar strategy to intermediate connected model, no matter how many layers of shortcuts are connected by LSTMs.

The usage of partial pretrained weights on the same dataset will take an role of unusual regularization: it can introduce the bias towards configurations of the parameter space that are beneficial to longitudinal training by setting up a particular initialization point. Since the weights are obtained by well-trained CNN backbones, this particular initialization point imposes an implicit constraint on the weights where it indicates which minima of the cost function are allowed to reach[17]. Consequently, the convergence of 4D model will be faster with a lower possibility of divergence. On the other hand, the more trainable parameters we have, the computation resources is required. By freezing most pretrained parameters of backbone during training 4D model, we could save memory for increasing input batch size, patch size or time-point number.

3.3.4. Evaluation Metrics

As mentioned in Chapter 1, the question to be answered in this project is the feasibility of longitudinal model in terms of improving the segmentation accuracy and consistency compared to CNN models. Therefore, the evaluation of models should take both accuracy and consistency into consideration.

Accuracy Metrics

The segmentation accuracy can be evaluated based on either spatial overlap or boundary curve(surface in 3D) distance between the predicted and the ground truth, which are referred to area-based and distance-based metrics, respectively. The most representative area-based metrics in segmentation is Dice Similarity Coefficient(DSC), which calculates the ratio of the overlap between ground truth and prediction to the union of them. In 3D cases, the area is adapted accordingly to volume. The mathematical expression of DSC is:

$$DSC = 2\frac{|V_{groundtruth} \cap V_{predicted}|}{|V_{groundtruth}| + |V_{predicted}|} \tag{3.7}$$

Hausdorff Distance(HD) and Average symmetric surface distance(ASD) are two classical distance-based metrics. HD is defined as the maximum distance of a set to the nearest point in the other set[2], which measures

the mutual proximity of two sets. The HD from set A to set B is:

$$\tilde{\delta}_H(A,B) = \max_{a\in A}\min_{b\in B}\|a-b\| \tag{3.8}$$

Since in most cases, the number of points in the two sets are different, which results in the asymmetric property of HD. That is, $\tilde{\delta}_H(A,B)$ is not equal to $\tilde{\delta}_H(B,A)$. Therefore, a general definition to measure the HD between set A and B is:

$$\delta_H(A,B) = \max(\tilde{\delta}_H(A,B), \tilde{\delta}_H(A,B)) \tag{3.9}$$

Unlike HD measures the maximum distance, ASD provides an average distance measurement between two sets. It calculates the average of all the distances from points on the boundary of prediction(B_P) to the boundary of ground truth B_P and from points on B_G to B_P:

$$ASD = \frac{1}{|B_P|+|B_G|} \times \left(\sum_{x\in B_P} d(x,B_G) + \sum_{y\in B_G} d(y,B_P)\right) \tag{3.10}$$

Consistency Metrics

Normal brain tissues are assumed to be relatively stable in terms of volume and shape in a mature brain, while glioma will grow slowly over time, which might lead to corresponding changes in other tissue. A straightforward measurement with regard to consistency is the smoothness of volume development of normal tissues along the time dimension. In the ideal case, the development of normal tissue volume before and after resection should be smooth. Right after the operation, the volume of glioma will see a sharp drop and CSF will increase suddenly in comparable volume. During the growth of glioma, normal tissues should experience extrusion or tissue transformation, leading to a slight decline of their volumes. However, it is tricky to evaluate this type of smoothness. The standard deviation (σ) can reflect the stability of volume development with regard to a mean value, but the development trend of volume is ignored. For instance, if the volume of tissue is gradually decreasing, the σ would be the half of difference between volume at the ending and at the beginning. The development trend can be extremely smooth in fact but σ may give a high value indicating a large uncertainty. Therefore, the volume size development is not a fair metric to evaluation the overall smoothness of development.

Hereby we propose a more robust and reasonable consistency metric, *Tissue Transformation Rate(TTR)* and *Tissue Maintaining Rate(TMR)*. In common cases, the normal tissue in the mature cerebral should keep stable and unchanged. If lesions happen, some normal tissues will transformed to tumor gradually. For glioma, this type of transformation is even slow. Tissue transformation rate describe the proportion of a certain type of tissue transformed into another type from last time to next time. On the contrary, tissue maintaining rate represents how many tissues maintain itself from time to time. We could therefore expect a consistent segmentation over time to have most normal tissues staying unchanged over time, i.e. a high tissue maintaining rate, and the tissue transformation to happen as little as possible, i.e. a low tissue transformation rate. The average TTR and TMR across all scans of each patient can be formulated as following:

$$TTR_{ij} = \frac{1}{T-1}\sum_{t=0}^{T-1} F_{ij}^{t\to t+1} \tag{3.11}$$

$$TMR_{ii} = \frac{1}{T-1}\sum_{t=0}^{T-1} F_{ii}^{t\to t+1} \tag{3.12}$$

$$F_{ij}^{t\to t+1} = \frac{\Sigma v_t^i v_{t+1}^j}{\Sigma v_t^i} \tag{3.13}$$

where A and B are different tissue labels. $F_{ij}^{t\to t+1}$ refers the transformation rate from tissue i to tissue j, from time t to $t+1$. v_t^i is the indicator function defined to be 1 if the voxel is labeled with i at time t while,

and 0 otherwise. T is the total MRI scans of a patient. Our expectation is to make TMR as high as possible and TTR as low as possible over all scans of a patient. The consistency evaluation could only take place on normal tissues, i.e. WM, GM and CSF, because development of tumor is not consistent anymore if operation happening occasionally. Although we cannot evaluate a smoothness of tumor segmentation over time, we should expect a much lower $F_{TM,TM}^{t\rightarrow t+1}$ if any resection happens, for the judgement of a reasonable segmentation between time points. Since we use the proportion of volume as the attribute, TMR and TTR are more robust than tissue volume development which takes the absolute volume for evaluation. Even when operation takes place or tumor is growing, we can always expect a relatively stable proportion of changed volume. As a result, the curve of TMR of normal tissues over time should have a high mean value and a low standard deviation. Another benefit to use these metrics is that it reveals the temporal changes of local tissues instead of mere tissue volume. The individual tissue distribution may differ from time point to time point seriously but their volume remains unchanged. Definitely in this case tissues are inconsistent but volume development metric will give a fake perfect smooth measurement. Because there are four targets in addition to background in our dataset, 25 possible transformation may happen. The best tool we can think of to visualize the different tissue conversions from last time to the next time is the transition matrix heatmap.

Regardless of using accuracy metrics or consistency metrics, the final goal is to evaluate if the model could achieve a more realistic result. From this perspective, a higher accuracy value cannot support the claim that the result approaches the ground truth better since the training masks are just estimations of truth. However, higher accuracy metric can be an indicator that the model's capacity to fit the artefact labels is advanced. We expect an idea model to have strong fitting ability while realize a more consistent, i.e. a more realistic segmentation. Apparently, there is no benchmark to evaluate how high the TMR is could the result be defined as consistent, so this metric is a relative metric between models. Due to the imperfect dataset, it is probably that even the training masks cannot meet the consistency criteria. In this case, a prominent different between model's prediction and mask is probably not a negative result, but the most authoritative judgement on whether prediction is better or not is always determined by the experts.

3.3.5. Loss Function

The loss function used for training in this project is Dice loss, which is a popular choice among practitioners. The mathematical form of Dice loss in binary segmentation can be modified from DSC as following:

$$L_{Dice}(y,\tilde{y}) = 1 - DSC = 1 - \frac{2\sum_{i\in\mathbb{X}} y_i \tilde{y}_i}{\sum_{i\in\mathbb{X}} y_i + \sum_{i\in\mathbb{X}} \tilde{y}_i} \tag{3.14}$$

where $\tilde{y} \in [0,1]$ is a continuous value, referring to the predicted probability, and y is the truth voxel label with binary values 0 and 1.

3.4. Experiment workflow

Within multiple CNN backbones and multiple longitudinal architectures, we would like to find out the optimal 4D networks and test it on our dataset. This process can be hierarchical. First, the preliminary experiments are conducted to choose optimal ConvLSTM component. Second, U-Net variants are evaluated with regard to the segmentation accuracy to select the most powerful backbone. Third, taking one of U-Net backbone to construct the three longitudinal networks and evaluate their accuracy performance, in purpose of finding the optimal 4D architecture. This step can be carried out together with step two, since we assume the effect of longitudinal information is independent of similar U-Net backbone types but concerns primarily with different connection strategies. Finally, the optimal CNN backbone is combined with best longitudinal architecture to create the final 4D longitudinal network. The evaluation of longitudinal model is a relative

comparison with its backbone CNN and masks to explore if the temporal information is in favour of improving accuracy as well as consistency. The whole workflow is illustrated in diagram3.15.

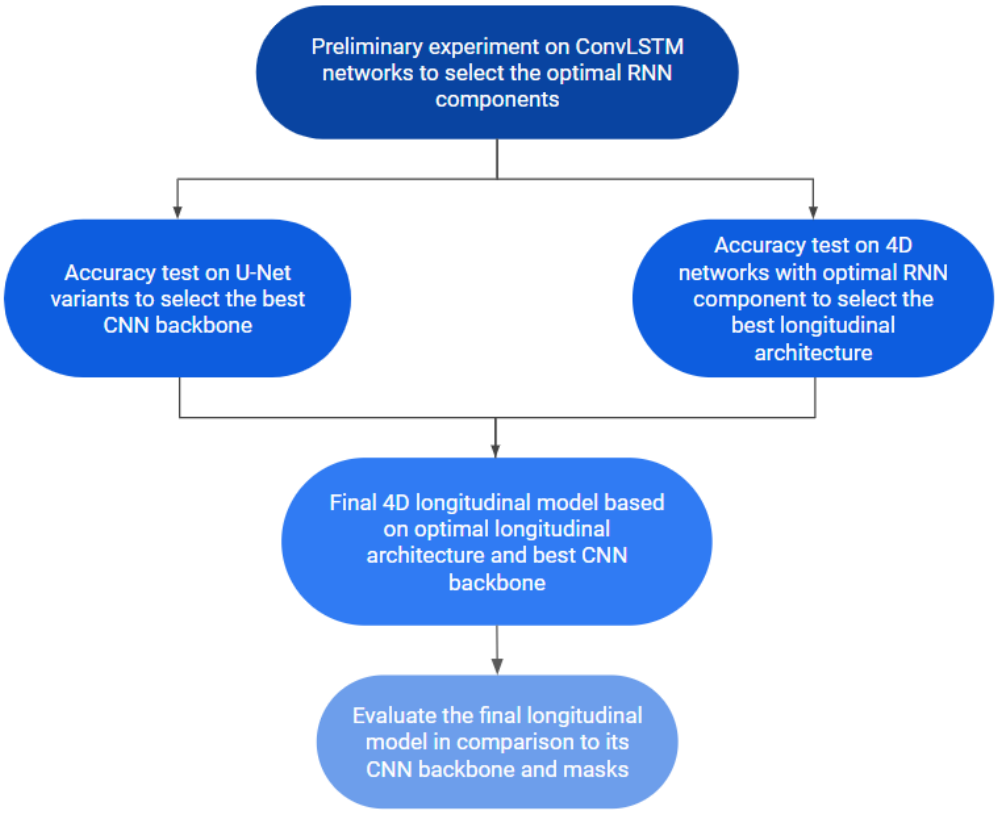


Figure 3.15: The experiment workflow of this project.

4

Results and Analysis

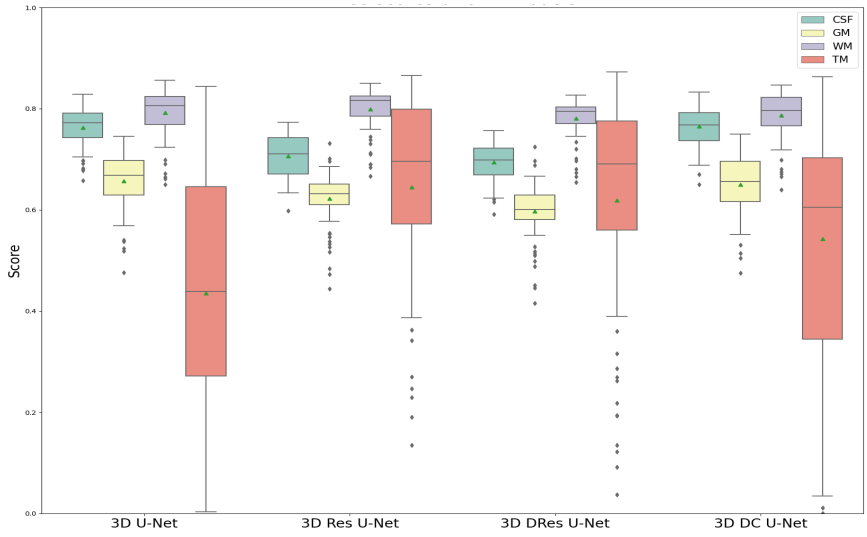
4.1. Results

All the models are tested on the testing dataset, which keeps untouched during training. Since the back propagation of network is deactivated when testing, the GPU memory is freed out. Therefore, instead of cropping out a patch to feed the model and stitching the predicted output up, a whole image is used as the input for testing. For CNN models, the batch size is 1 with 1 time-point, i.e. 1 image each time; for 4D models, the batch size is 1 as well, but the time-point can be any value as long as within the GPU capability. Here the time-point is maintained to be equal to the value used during training.

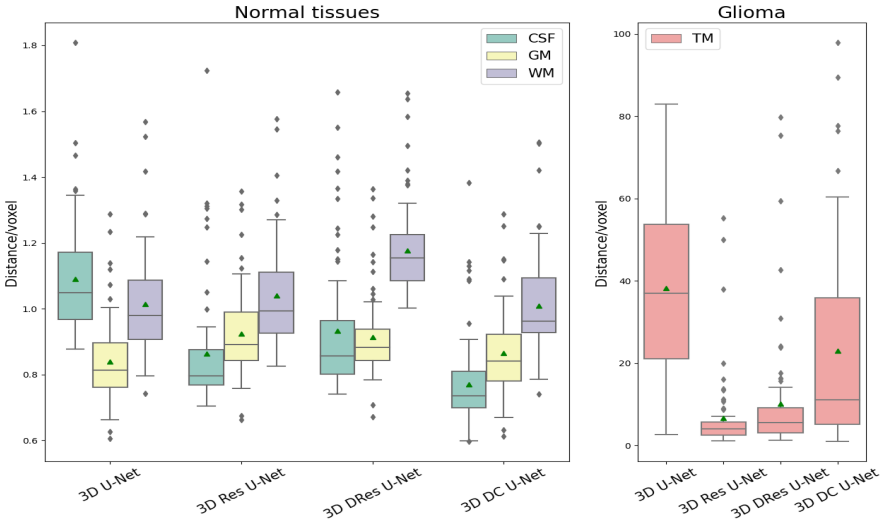
4.1.1. Results of accuracy

Optimal CNN backbone selection

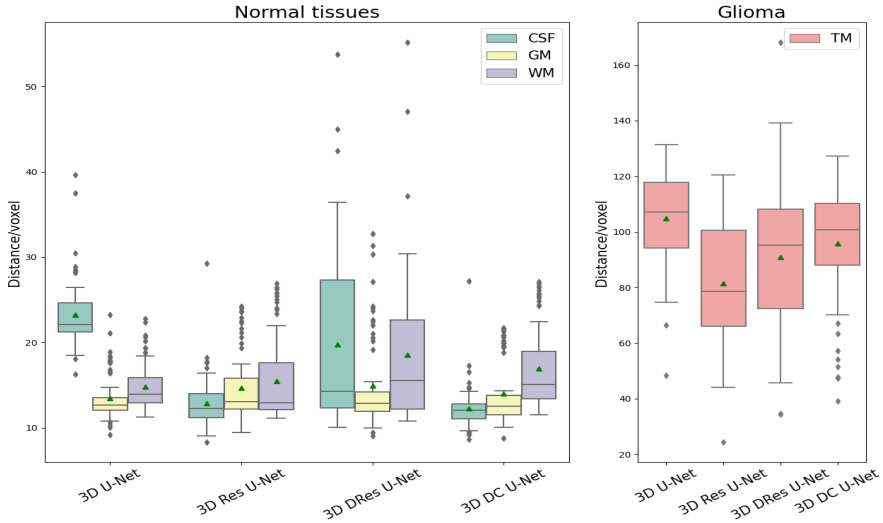
The accuracy evaluation based on DSC, ASD and HD of CNN models across 7 testing patients are presented in Figure 4.1. Looking at DSC boxplot, we could note that compared to baseline, the introduce of residual connection boosts Dice score on tumor segmentation significantly and slightly on WM, but meanwhile, the performance on other two normal tissues drops down a little. Surprisingly, replacing the bridge of 3D Res U-Net by the dilation convolution does not see positive effects on Dice coefficient for all the targets. The 3D DCU-Net improves the tumor segmentation accuracy to a degree while maintaining the performance on other normal tissues at the same time. Since the tumor is the most difficult target to be segmented due to its small and dramatically changing size and location, the variance of its DSC is consequently largest. With regard to ASD, all U-Net variants improve performance on CSF and tumor segmentation. 3D DC U-Net outperforms the other two variants on normal tissues but achieves worse score on tumor. Under the evaluation of HD, U-Net variants don't have apparent advantage when compared to the baseline on normal tissues, but they all perform better on tumor segmentation. It is hard to say which network absolutely outperforms others because under different metrics, every model may have advantages. However, it is apparent that residual connection do help improve tumor segmentation, for 3D Res U-Net always providing the best result on this target across all the metrics. The proposed 3D DC U-Net also strengthens the capability of segmenting small targets like tumor, although the result is less good as residual connection module. Nevertheless,its segmentation quality on normal tissues is better than Res U-Net as well as DRes U-Net, and almost equally good as baseline. That means the "hard-code" concatenation from pooled raw input image indeed contains more detailed location information benefiting the segmentation of small targets.



(a) DSC



(b) ASD



(c) HD

Figure 4.1: From up to down presents the three accuracy metrics for 4 CNN model variations. The model variants are list on the x-axis, and the metric value is on y-axis. For distance-based metrics, the results of normal tissues are separated from glioma since the value scale of tumor is much larger than other tissues. Different face colors of boxes represents different target regions. The horizontal line across each box is the median value and the green triangle is the mean value. The gray dots outside the maximum and minimum boundary caps are extreme values. For two distance-based metrics, since the scale of results of tumor is largely different from normal tissues, their boxplots are separately shown.

To investigate how good the CNN model is able to perform on the segmentation accuracy, I choose the best three networks, U-Net baseline, Res U-Net and DC U-Net, to compare them with different block channel numbers and cropped input patch sizes. The results (see Table 4.1) shows that increasing the number of convolution channels in blocks has no difference in performance of DSC for all variants, but utilizing larger image patch as input results better outcomes than small patch size. However, larger patch size costs more GPU memory.

Table 4.1: The table shows the mean DSC of different experiment setup of CNN variants on testing data

Methods (base channels / patch size)	CSF	GM	WM	TM
U-Net (16 / 64)	0.772	0.658	0.792	0.436
U-Net (32 / 64)	0.764	0.663	0.789	0.442
U-Net (16 /128)	0.798	0.695	**0.825**	0.694
Res U-Net (16 / 64)	0.710	0.625	0.801	0.647
Res U-Net (32 / 64)	0.721	0.618	0.795	0.672
Res U-Net (16 /128)	0.724	0.669	0.819	0.705
DC U-Net (16 / 64)	0.788	0.648	0.792	0.580
DC U-Net (32 / 64)	0.770	0.651	0.810	0.611
DC U-Net (16 / 128)	**0.805**	**0.702**	0.820	**0.747**

These results indicate that the 3D DC U-Net presents the best overall performance with regard to the segmentation accuracy if the patch size is large enough. In basic condition where $64 \times 64 \times 64$, the performance of 3D DC U-Net and 3D Res U-Net is comparable.

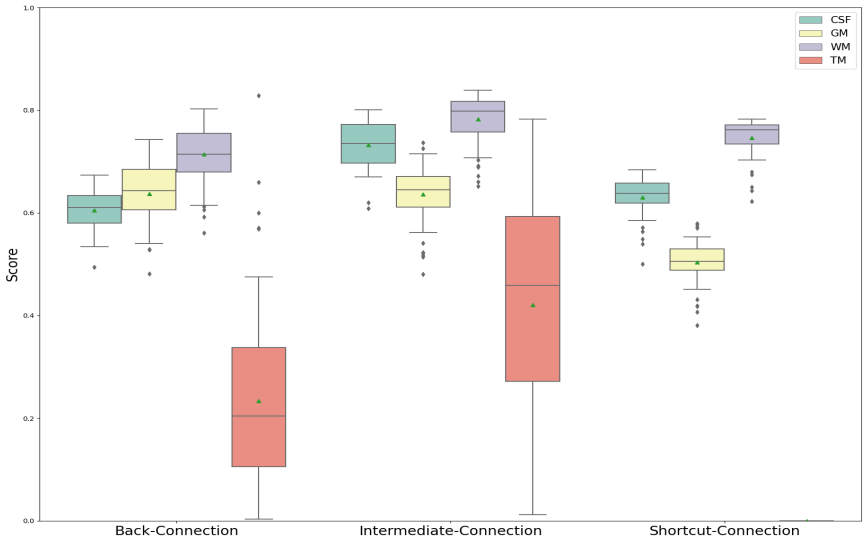


Figure 4.2: The accuracy results for different 4D longitudinal model structures, taking 3D U-Net as example backbone.

Optimal longitudinal architecture selection

Any U-Net variants introduced before can be used as the backbone of longitudinal models, but we assume that the additional effects brought by longitudinal connection are independent on the type of backbone. Since the focus of this section is on the comparison between different 4D architectures, the simplest 3D U-Net backbone is taken as an example. As introduced in previous chapter, the RNN architecture used in longitudinal networks is 1-layer bidirectional LSTM based on preliminary experimental results. Figure 4.2 presents DSC of three proposed 4D models. Except for the comparable DSCs of GM between Back-connection and Intermediate-connection type, the intermediate-connection model overwhelmingly outperforms the other two. Interestingly, shortcut connection type even failed to segment tumors. This failure leads to infinite val-

ues on distance metrics which is meaningless, so only DSC metric is provided here.

Comparison between optimal longitudinal type and corresponding backbone

To the utmost extent to evaluate the longitudinal information influence on accuracy, we should select the best 4D architecture to compare with its backbone model. Based on the accuracy evaluation of three longitudinal models, the Intermediate-connection one demonstrates the best learning ability to mimic the masks. Here we take the 3D DC U-Net and its Intermediate-connection longitudinal model as an example to show their disparity in terms of accuracy(shown in Figure 4.3). We could notice that the overall accuracy of longitudinal network is slightly superior to its backbone on the three metrics. However, this tiny superiority cannot be seen in the Intermediate-connection with 3D Res U-Net and 3D U-Net backbone(See the results in Appendix C), where the performance of longitudinal model is even a little bit worse than backbone. In fact, the pre-trained weights from CNN backbone plays a major role in maintaining 4D model's segmentation accuracy, since if the longitudinal network training from scratch, the result is even worse. The accuracy comparison of 3 types of longitudinal models taking 3D U-Net as backbone training with and without pretrained weights refers to Appendix B.1.

The segmentation accuracy of optimal 4D model

Based on previous experimental results, the best CNN backbone is between 3D DC U-Net and 3D Res U-Net, depends on the input patch size; the optimal longitudinal architecture is Intermediate-connection type. The final 4D model is the combination of these two components. We therefore compare the segmentation accuracy between the optimal longitudinal type with 3D U-Net, 3D Res U-Net and 3D DC U-Net to obtain the 4D model with best accuracy performance. To avoid hardware memory issue, the crop patch size used in this experiment is $64 \times 64 \times 64$. The results are shown in 4.4. Except for HD metric, the Intermediate-connection type with 3D DC U-Net backbone outperforms the other two significantly. Especially for DSC metric, this model produces not only higher mean and median score, but also a more concentrated distribution of prediction.

4.1.2. Results of consistency

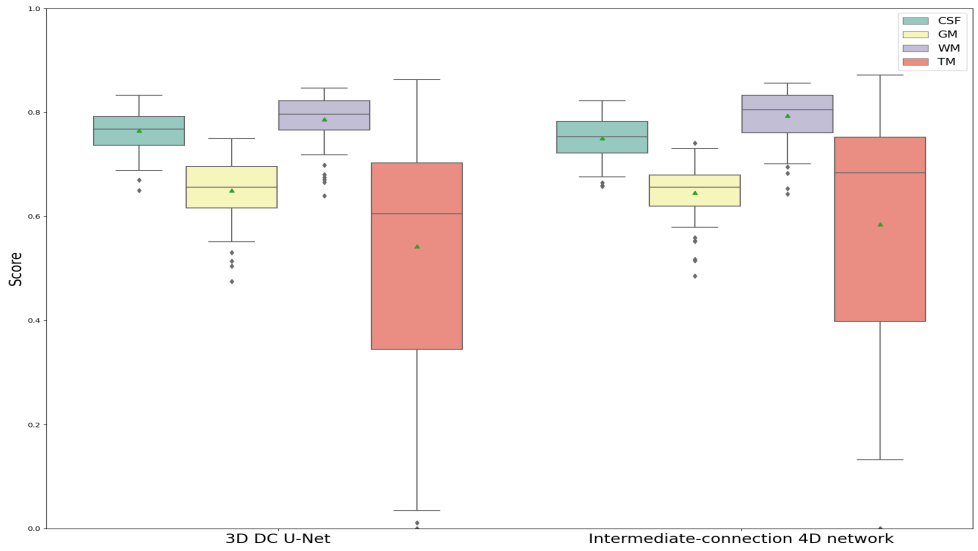
The major question to be answered in this project is whether the longitudinal information help increase the segmentation consistency. The 4D model used in this experiment is the Intermediate-connection longitudinal architecture with 3D DC U-Net backbone, for sake of its best performance on accuracy metrics.

TMR of normal tissues over time

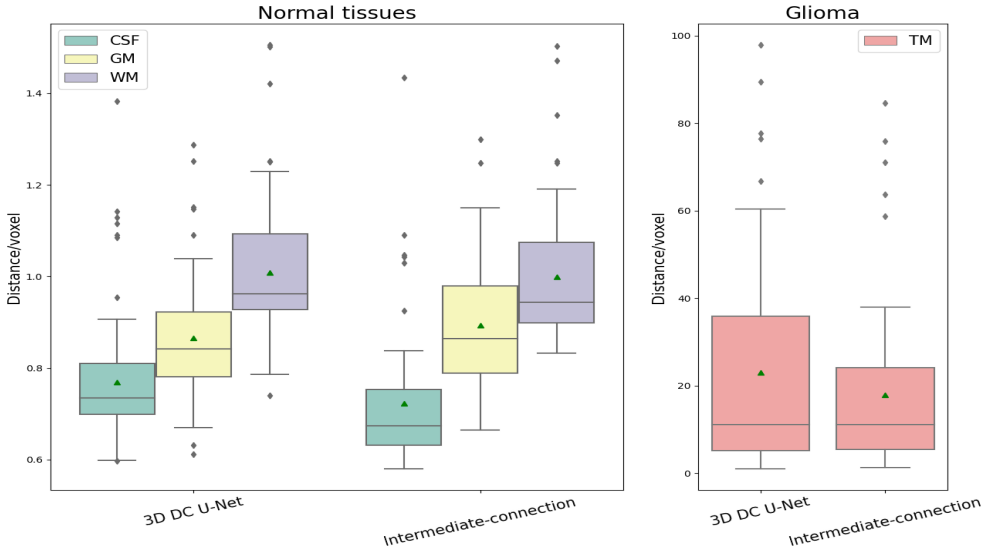
Since the TMR and TTR are dual metrics, to have a intuitive feeling of the segmentation consistency, we plot the TMR development curve over MRI scans of each testing patient, as shown in Figure 4.5, where 3 of them are listed as examples. It is apparent that the results from longitudinal model have highest mean TMR across scans among three predictions, meaning most normal tissues are maintained itself on average in longitudinal prediction. However, the standard deviations of longitudinal results are not always lowest.

Tissue transformation rate

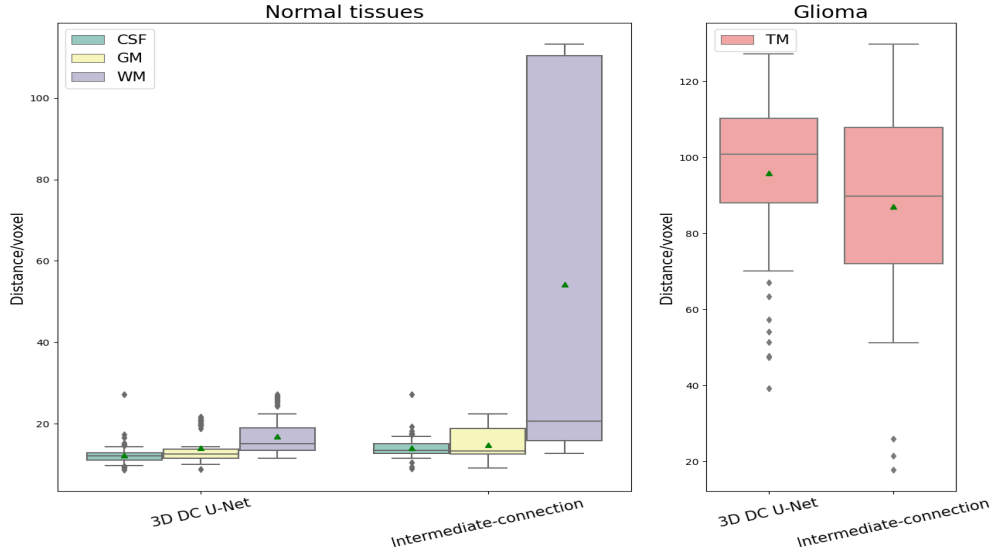
TMR development curve gives a straightforward overview of segmentation consistency. We also interested in the exact transformation of tissue from one type to another, i.e TTR performance. The transition matrix heatmap of average TTR across MRI scans of the three testing patients is presented in Figure 4.6. Here we could see the tissue transformation condition over time. While normal tissues keep high TMR(diagonal of heatmap) and low TTR in all these three patients, the glioma TMR of mask in patient EGD-0125 and EGD-0265 is much lower, which indicates there are resection happens at some time points. We could also notice that the TMR and TTR of glioma in patient EGD-0125 and EGD-0505 from mask differ significantly from deep learning models, which demonstrates a large discrepancy between mask and prediction.



(a) DSC

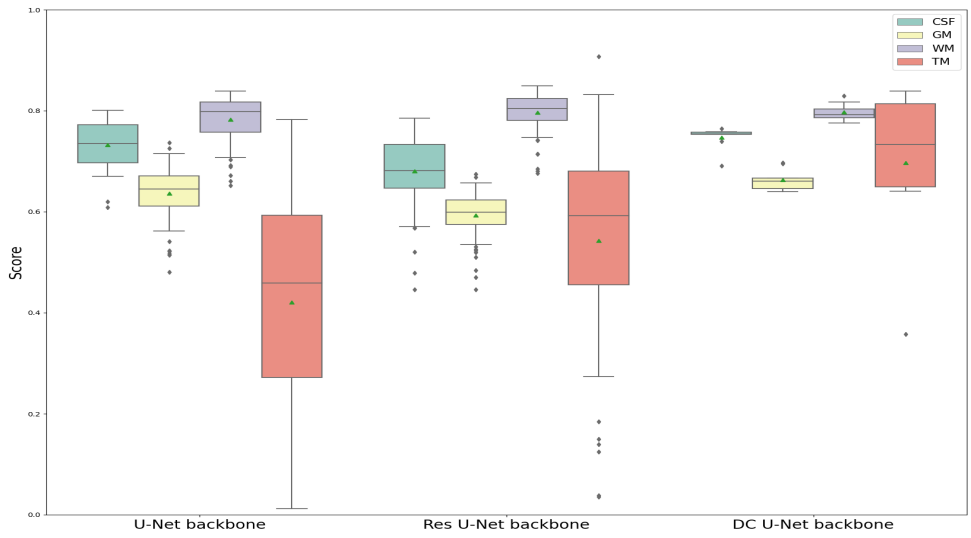


(b) ASD

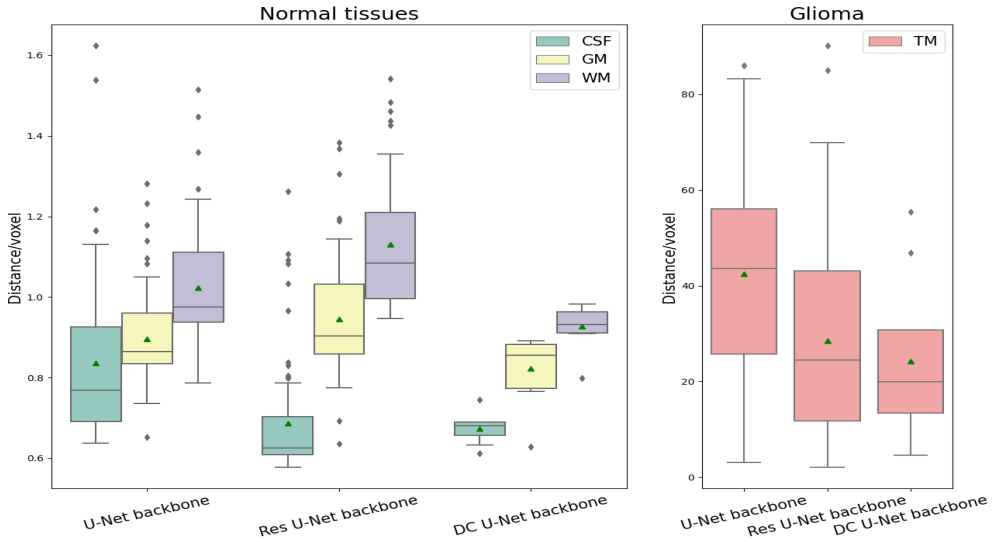


(c) HD

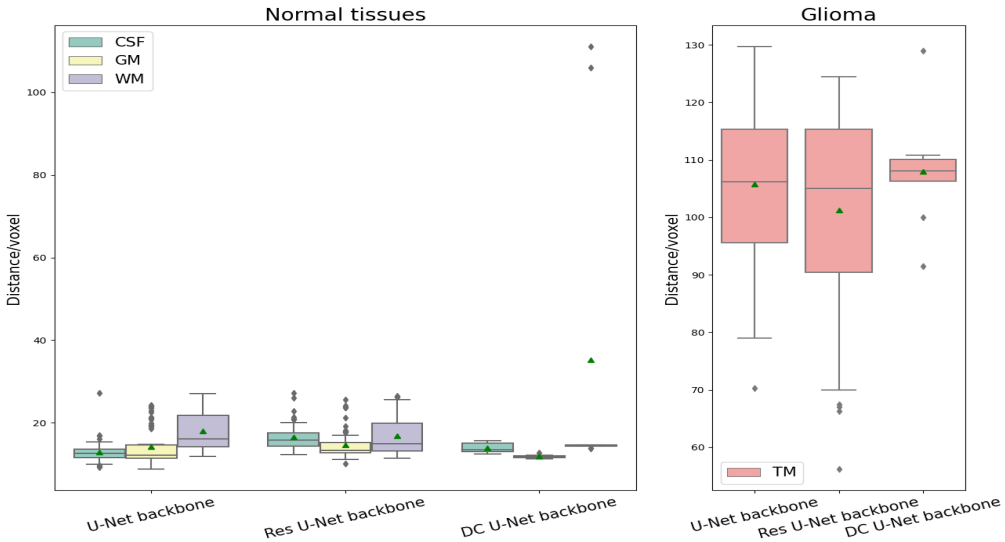
Figure 4.3: Accuracy comparison between 3D DC U-Net and intermediate-connection 4D model.



(a) DSC

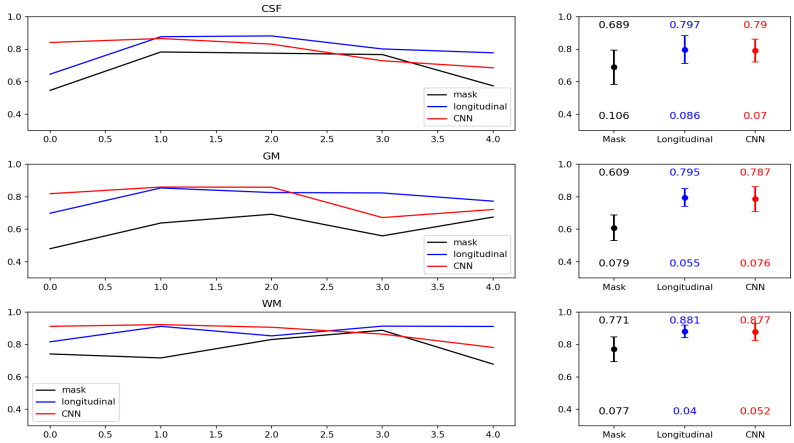


(b) ASD

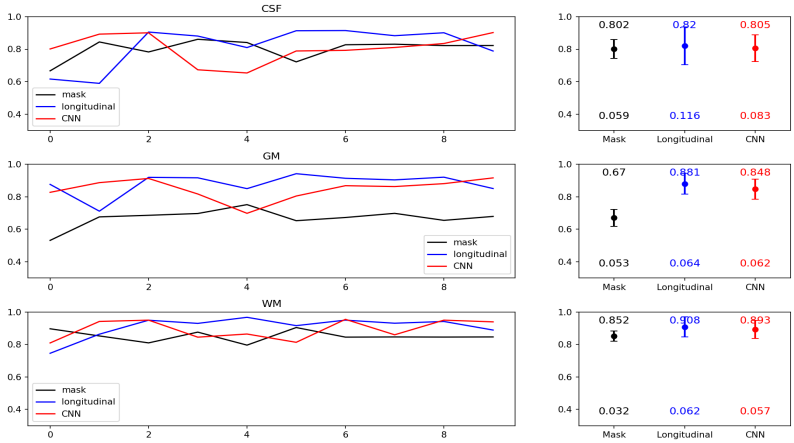


(c) HD

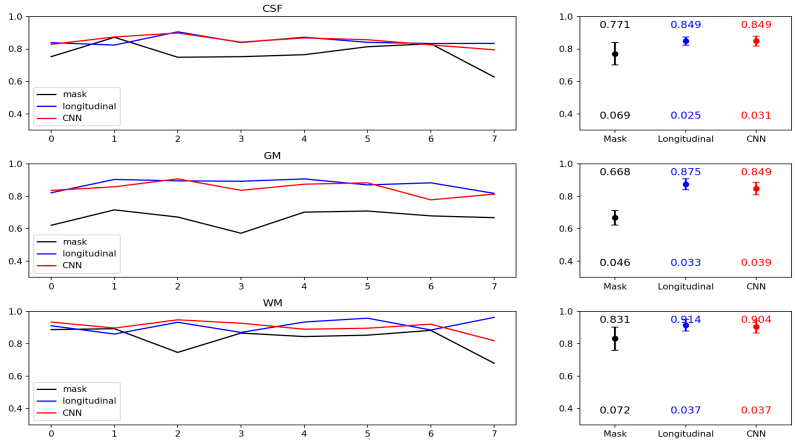
Figure 4.4: Accuracy comparison between optimal longitudinal architecture with 3 different backbones. The one with 3D DC U-Net backbone outperforms than the other two.



(a) EGD-0125



(b) EGD-0265



(c) EGD-0505

Figure 4.5: Left: TMR development curve over time of 3 testing patients. Right: mean and standard deviation of TMR curve. The values above segments are mean and below are standard deviation. A mean value of 1 means that, with respect to the previous time-point, all voxels of that label remained the same.

4.2. Discussion

The results of CNN networks from three metrics suggest that the primary improvement of model variants is on tumor segmentation. Surprisingly, compared to pure residual connected 3D Res U-Net, the mixed dilation convolution and residual connection structure in 3D DRes U-Net harms the segmentation accuracy of tumor marginally, instead of improving it further, which is contrary to our initial expectation that dilation convolution can better capture multi-scale contextual information of glioma. The underlying cause of this unexpected phenomenon remains unclear. Another important finding is that the designed hard-code identical mapping in 3D DC U-Net could help improve tumor segmentation as well, and maintain a fairly good quality on normal tissues segmentation simultaneously. When the patch size is increased to 128, this variant even achieved best performance on CSF, GM and TM, with only a tiny distance from best WM score obtained by 3D U-Net. It is therefore likely that such a raw information mapping from input is indeed helpful in extracting location information.

The comparison between three types of longitudinal networks is interesting as well. Especially for Shortcut-connection type, the segmentation of tumor totally failed. A possible explanation might be that the identity mapping from encoder path is designed to provide detailed location information in favor of feature recovering, but the shape and position of tumor probably vary significantly along chronological axis, which will mess up information and mislead the network to regard it as the least consistent context waiting to be abandoned. The bridge part of backbone stores most abstract feature maps which only focus on high level semantic information, hence the inter-communication at this level won't affect much the details expression in final segmentation. As for back-connection model, since the weights of backbone are unable to update, the training only takes place within the one-layer BiConvLSTM network. The information flow on time dimension does not contribute effectively to the improvement of each individual backbone, which results in the degraded performance compared to single CNN network. Interestingly, if the decoder path is freed out for training, like the intermediate-connection model does, although the segmentation of normal tissues maintains good results as obtained from individual backbone, the tumor region is unable to be detected any more, referring to Shortcut-connection's performance. The reason accounts for this phenomenon could be attributed to the low-level features' interaction as well. The connection position is the output of backbone, where detail features are already recovered. Recurrent convolution at this point is likely to mess up the individual segmentation features and properties again. Inconsistent targets(glioma) will be dropped out while consistent ones be kept(WM, GM, CSF). To sum up, in order to force the communication along temporal dimension while prevent from losing individual details, the connection should take place at deepest part where highest level features exist, and low level features keep separated from each other.

Experimental results show that the longitudinal information do has some impacts on the segmentation quality. First, the upper limits of segmentation accuracy of optimal longitudinal architecture is somehow determined by its 3D backbone, although for 3D DC U-Net we noticed a slight improvement on its corresponding Intermediate-connection 4D network, but that is too little to be a valuable improvement. Second, the superiority on segmentation accuracy of CNN backbone cannot definitely transmitted to its corresponding 4D model. As we see that the longitudinal DC U-Net overwhelmingly outperforms longitudinal Res U-Net on all four targets, while among individual 3D models, the Res U-Net produces best TM segmentation accuracy. Third, with regard to the consistency, the findings cannot support the hypothesis that the longitudinal network could absolutely approximate the ground truth better than CNN networks. Although the longitudinal model provides the highest TMR of normal tissues over time, the stability of prediction in successive time points is not always better than CNN backbone. Sometimes the standard deviation is even a bit higher than mask. This phenomenon may partly be explained by the imperfect dataset. As discussed in chapter 3, er-

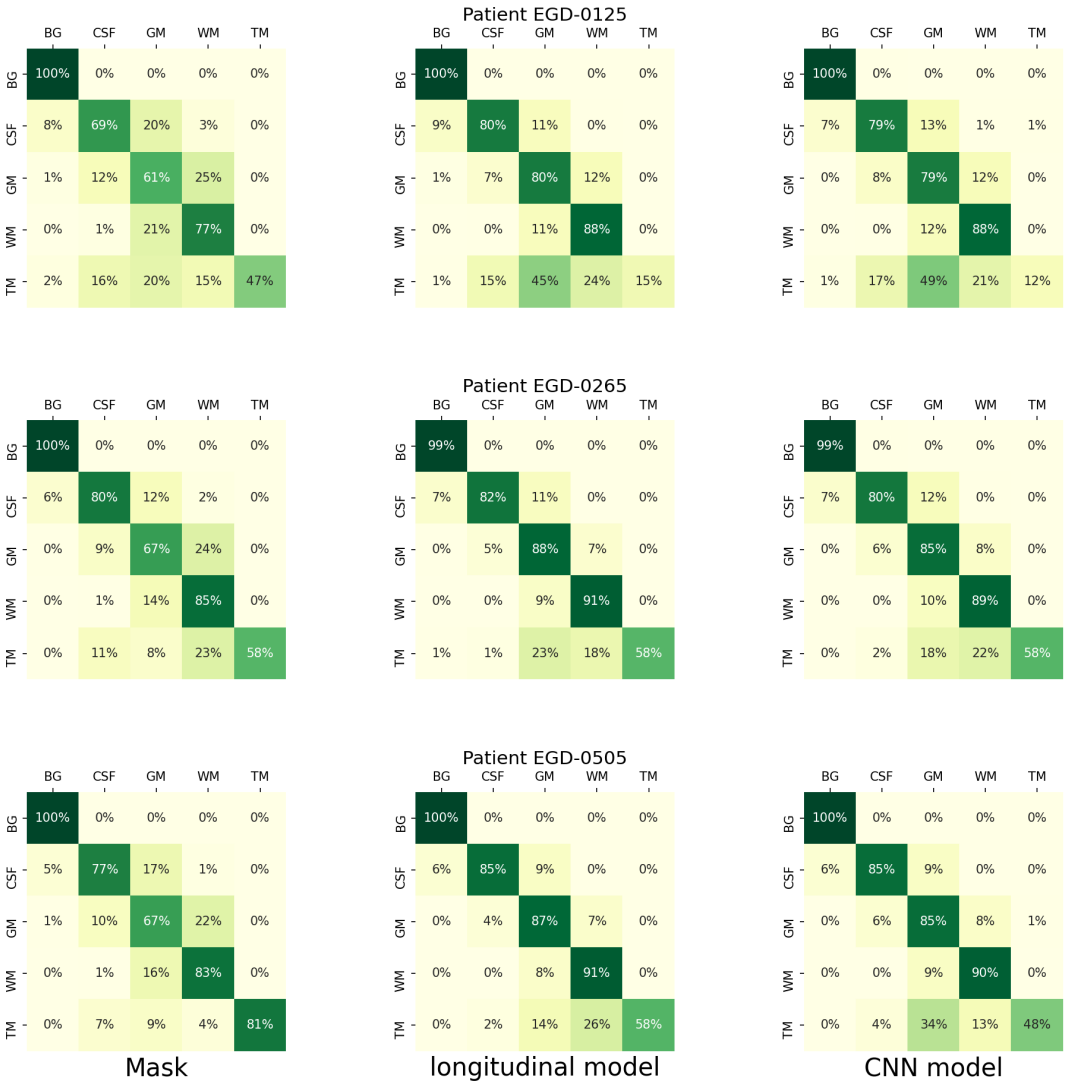


Figure 4.6: Average transition rate matrix over time of 3 testing patients. From left to right column: ground truth mask, longitudinal model, 3D DC U-Net. The x-axis of transition heatmap refers to the label of t_i and the y-axis is the label of t_{i+1}. The value in each square is the averaged transformation rate across the whole time span.

rors existing in masks break the consistency in training samples, like the example shown in Figure 3.12, while unfortunately the longitudinal model is likely to strengthen consistent context and abandon inconsistent contents. If the training examples fail to provide consistent enough information, the fusion of temporal contexts might even impose negative influence. Another possible explanation is that the widely variant scanning interval in dataset. Since this is real medical data, we could not expect the patient to have uniform visits for MRI scans. Further more, some scans might be lost or unavailable at some time points. This inherent property of data may cause the violation of stable tissue transformation assumption in training samples, since the growth of glioma after a enough long time will prominently affect other normal tissues no matter how slow it progresses.

5

Conclusions and Future Work

5.1. Conclusions

In this project, we created longitudinal networks to evaluate the influence of temporal information on segmentation quality from accuracy and constancy perspective by a hierarchical design flow. First, a 3D U-Net and its variants 3D Res U-Net and 3D DRes U-Net by introducing the ideas of residual connection and dilation convolution were implemented. A new hard-coded identity mapping idea is proposed to develop the 3D DC U-Net. After comparing their performance on brain tissues and glioma semantic segmentation tasks, the 3D DC U-Net is selected to be the optimal backbone of longitudinal model creation as it is proven to possess the capacity of improving tumor segmentation accuracy while maintaining the quality of other normal tissue segmentations at the same time. Further experiments on CNN suggest that a larger input patch size leads to a significant boost on segmentation accuracy. Second, three types of longitudinal models by combining CNN with LSTM networks differently were proposed. The results indicate that the intermediate-connection type provides best segmentation accuracy among these three. The final longitudinal model is generated based on the optimal longitudinal architecture and CNN backbone. Third, the evaluation of final 4D longitudinal model is conducted on both accuracy and consistency. We found that the longitudinal segmentation accuracy is probably limited by the CNN capacity. As for the consistency, even though the longitudinal segmentation achieves best mean TMR, the standard deviation of TMR curve shows no superiority to individual segmentation by CNN. These findings, however, may be somewhat limited by the imperfect data.

Back to the research questions proposed at the beginning, the answers are presented as following based on the experimental results:

- **How to design longitudinal multi-target segmentation model with limited memory usage?**
  The longitudinal segmentation network can be created by combining 3D U-Net backbone and LSTM networks in different approaches with the number of output feature channels equal to the number of segmenting targets plus one (background). In this project, three types of longitudinal models are implemented as examples. To conquer the GPU memory limitation, pretraining of CNN backbone strategy is adopted.

- **How to evaluate the segmentation consistency?**
  Based on the fact that the normal brain tissues of an adult should keep relatively stable in terms of morphology over time, the evaluation of consistency can therefore based on the Tissue Transformation

Rate(TTR) and Tissue Maintaining Rate(TMR). We expect a stable and high value for average TMR of normal tissues across MRI scans to show a consistent segmentation result. This metric cannot be used on glioma segmentation since this target is inconsistent at all.

- **Can longitudinal information help improve segmentation accuracy of each target region?**
  Compared to the CNN backbone, the introduce of longitudinal information does not contribute positively to the segmentation accuracy in present study. Instead, the accuracy depends on the applied CNN backbone to a certain degree.

- **Can longitudinal information help improve segmentation consistency of each target region?**
  At least on our dataset, the answer is partial improvement. The longitudinal model improves the mean TMR over time, the standard deviation, which describes the stable property of tissue transformation, is however not always better than CNN and masks. These finds should be interpreted with caution since the dataset we are using is not so accurate.

5.2. Future Work

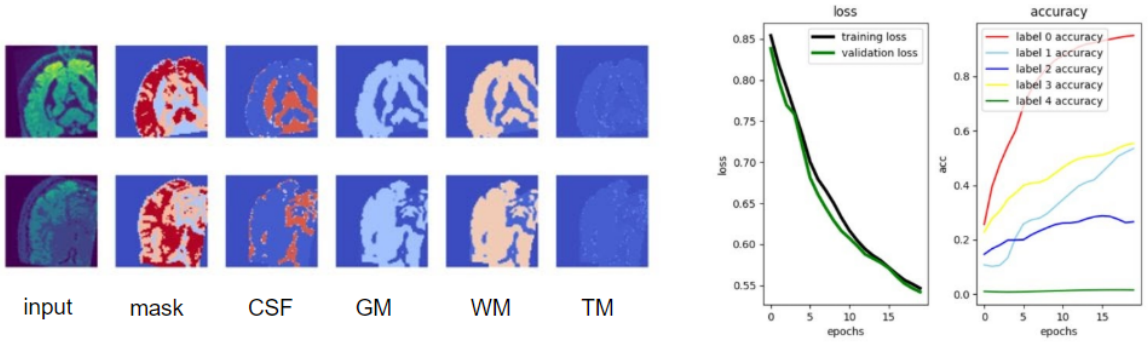
Time, data and computation resource are three major concerns in developing deep learning networks. There are also many tricks for training network to obtain higher scores on evaluation metrics. Although no benchmark in this exploratory project can be used to evaluate whether the generated segmentation are good enough or not, it is worth to note a better result is always there. Due to the time limitation, no exploration of best hyper parameters combination was conducted since the absolute highest accuracy is not the research topic of this project, not to mention the dataset is imperfect. Some experimental parameter setups like the optimizer, learning rate decay strategy and so on are chosen empirically. A more sophisticated finetune can be done if a better accuracy is desired and a dataset with better segmentation labels is available.

With regard to the longitudinal model, the LSTM networks used in research are basic ones. There are many other innovative architectures available in literature which can be arbitrarily inserted as a part of CNN backbone, like listed in [39]. Similarly, it is possible to replace the CNN backbone from U-Net to other non encoder-decoder segmentation networks as well, for instance Deeplab family[14, 15]. The integration strategy between CNN and RNN is also worth to explore further to find the most suitable 4D architecture.

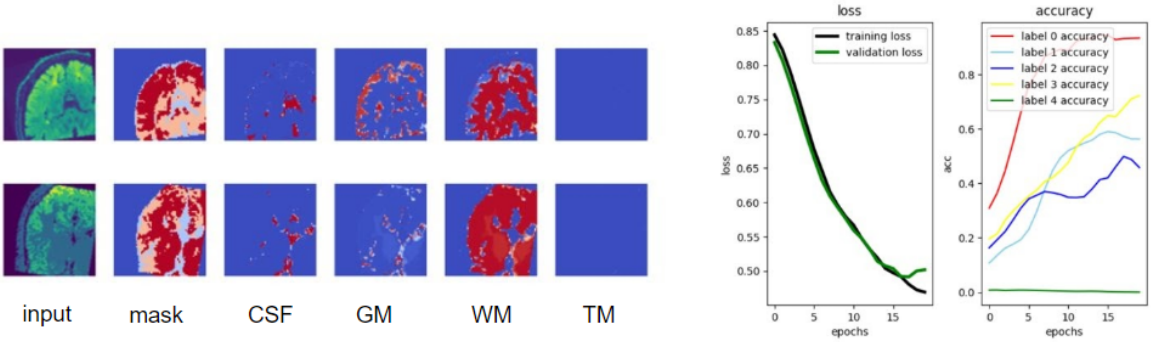
For the training strategy of longitudinal model, the input time step number used in this study is kept unchanged. A more common training paradigm of RNN network would use different data length in each input to capture more general information over time.

A

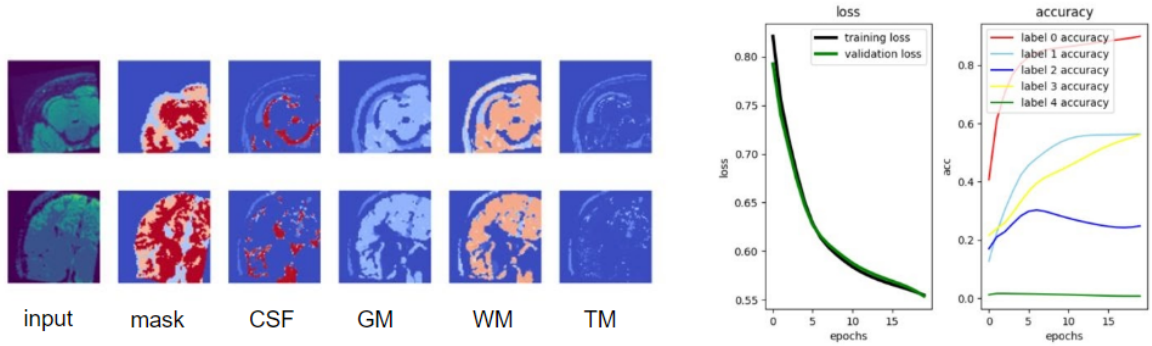
Preliminary Experiments on Pure LSTM Networks



(a) Two-layers Stacked ConvLSTM



(b) One-layers BiconvLSTM

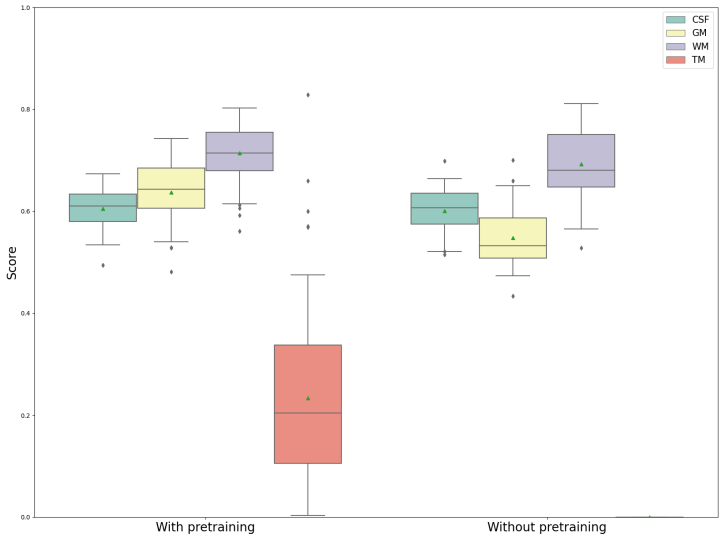


(c) Three-layers BiConvLSTM

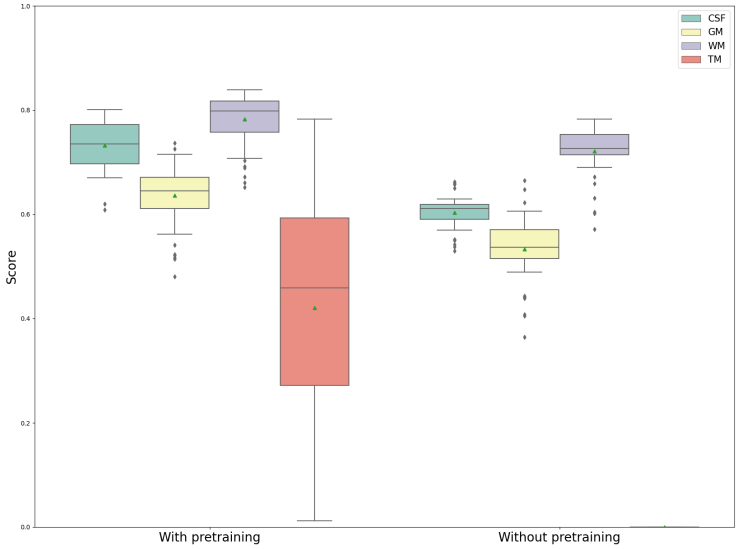
Figure A.1: Left: binary segmentation map per channel. Right: training loss and accuracy curves. The bidirectional convolutional LSTM with one layer gives fastest convergence.

B

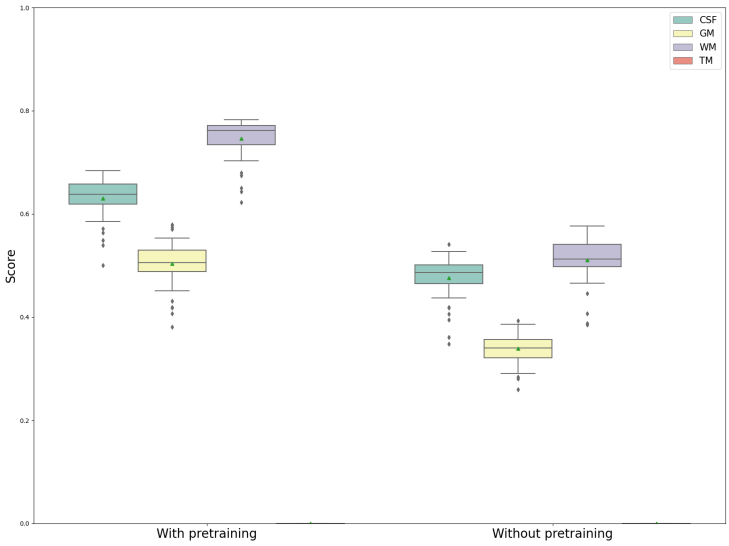
Effect of Pretrained Weights on Longitudinal Networks



(a) Back-Connection



(b) Intermediate-Connection

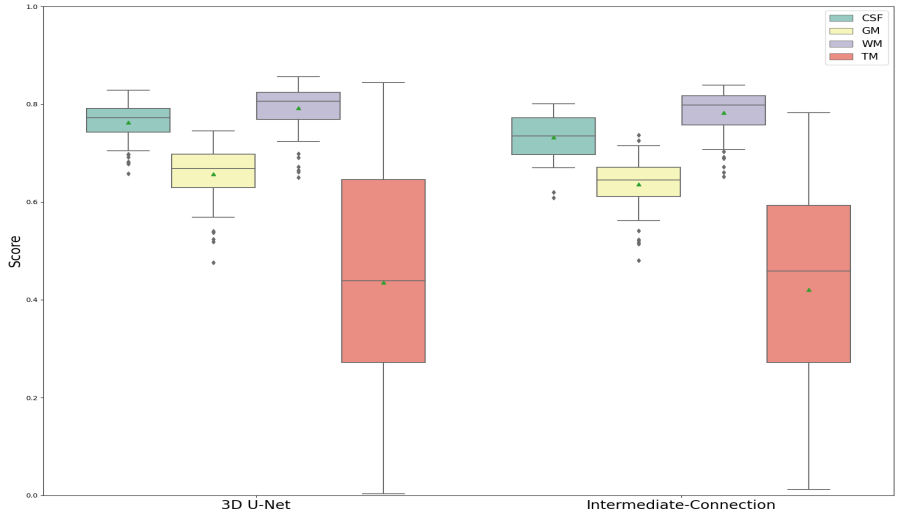


(c) Shortcut-Connection

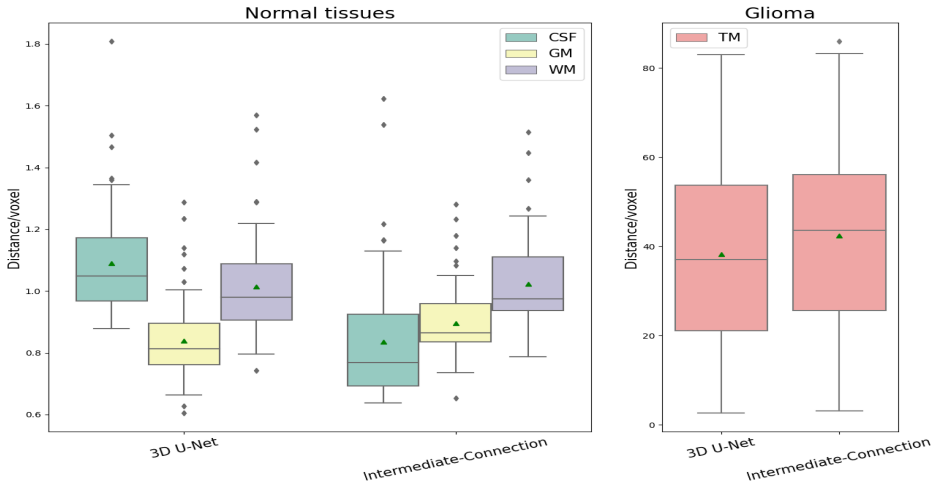
Figure B.1: The accuracy comparison with regard to DSC from three types of longitudinal networks. Without pretraining weights on corresponding CNN backbone, all the longitudinal models provide worse performance.

C

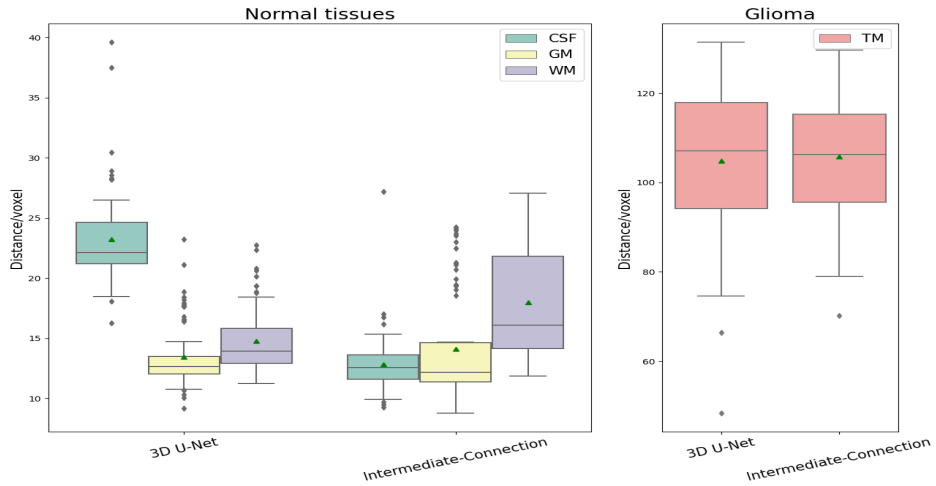
Accuracy Comparison Between Longitudinal Networks and Corresponding 3D Backbones



(a) DSC

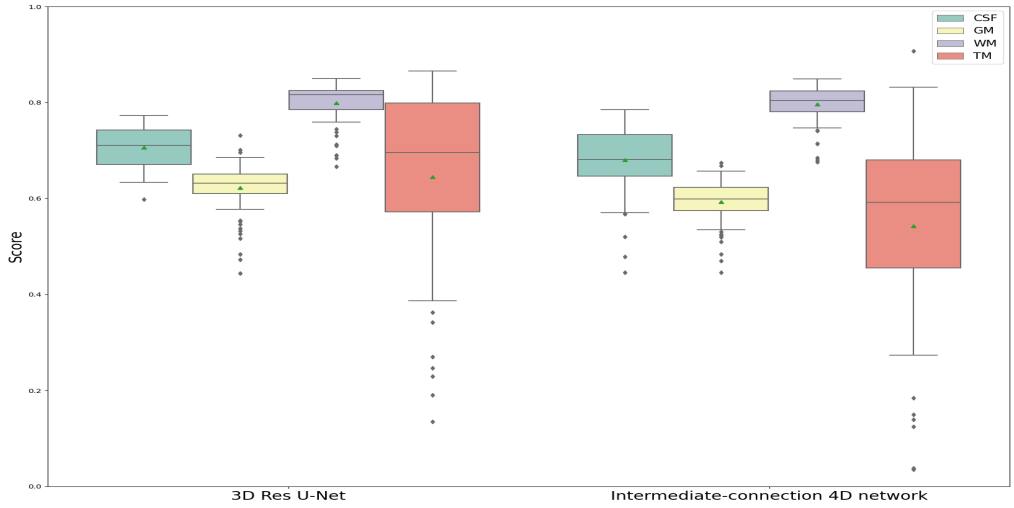


(b) ASD

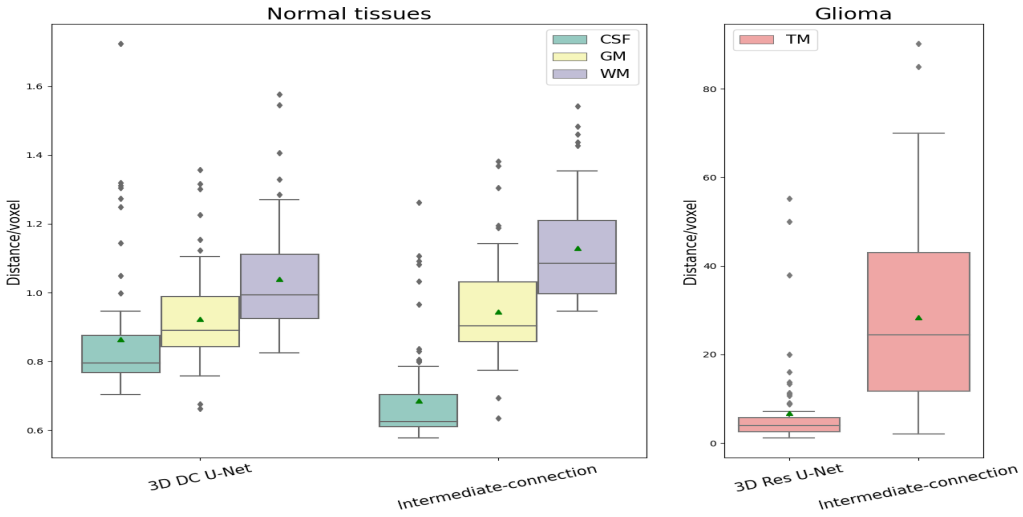


(c) HD

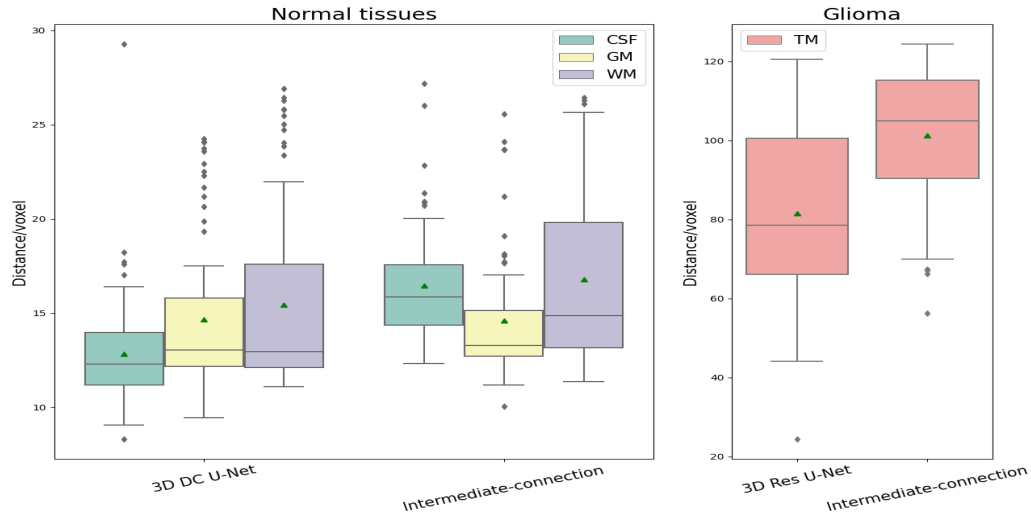
Figure C.1: The accuracy results with regard to DSC, ASD and HD of 3D U-Net backbone and its Intermediate-connection type 4D model. Longitudinal network does not show superiority to 3D backbone.



(a) DSC



(b) ASD



(c) HD

Figure C.2: The accuracy results with regard to DSC, ASD and HD of 3D Res U-Net backbone and its Intermediate-connection type 4D model. Longitudinal network does not show superiority to 3D backbone.

Bibliography

[1] Elastix. URL https://elastix.lumc.nl/.

[2] Hausdorf distance. URL http://www-cgrl.cs.mcgill.ca/~godfried/teaching/cg-projects/98/normand/main.html.

[3] Hd-bet, . URL https://github.com/NeuroAI-HD/HD-BET.

[4] Hd-glio, . URL https://github.com/NeuroAI-HD/HD-GLIO.

[5] Mri modalities. URL https://casemed.case.edu/clerkships/neurology/Web.

[6] Cartesius cluster. URL https://userinfo.surfsara.nl/systems/cartesius/getting-started.

[7] Glioma treatment, . URL https://www.urmc.rochester.edu/neurosurgery/services/brain-spinal-tumor/conditions/low-grade-glioma.aspx.

[8] Glioma class, . URL https://www.uptodate.com/contents/low-grade-glioma-in-adults-beyond-the-basics.

[9] Pytorch. URL https://pytorch.org/.

[10] Esther Alberts, Guillaume Charpiat, Yuliya Tarabalka, Thomas Huber, Marc-André Weber, Jan Bauer, Claus Zimmer, and Bjoern H Menze. A nonparametric growth model for brain tumor segmentation in longitudinal mr sequences. In *BrainLes 2015*, pages 69–79. Springer, 2015.

[11] Reza Azad, Maryam Asadi-Aghbolaghi, Mahmood Fathy, and Sergio Escalera. Bi-directional convlstm u-net with densley connected convolutions. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[12] Yuri Boykov and Gareth Funka-Lea. Graph cuts and efficient nd image segmentation. *International journal of computer vision*, 70(2):109–131, 2006.

[13] Jianxu Chen, Lin Yang, Yizhe Zhang, Mark Alber, and Danny Z Chen. Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation. In *Advances in neural information processing systems*, pages 3036–3044, 2016.

[14] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[15] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[16] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[17] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 201–208, 2010.

[18] Yang Gao, Jeff M Phillips, Yan Zheng, Renqiang Min, P Thomas Fletcher, and Guido Gerig. Fully convolutional structured lstm networks for joint 4d medical image segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1104–1108. IEEE, 2018.

[19] Amir Gholami, Shashank Subramanian, Varun Shenoy, Naveen Himthani, Xiangyu Yue, Sicheng Zhao, Peter Jin, George Biros, and Kurt Keutzer. A novel domain adaptation framework for medical image segmentation. In *International MICCAI Brainlesion Workshop*, pages 289–298. Springer, 2018.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[22] Yipeng Hu, Marc Modat, Eli Gibson, Nooshin Ghavami, Ester Bonmati, Caroline M Moore, Mark Emberton, J Alison Noble, Dean C Barratt, and Tom Vercauteren. Label-driven weakly-supervised learning for multimodal deformable image registration. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1070–1074. IEEE, 2018.

[23] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[24] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, 24(7):881–892, 2002.

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[26] Bo Li, Wiro J Niessen, Stefan Klein, Marius de Groot, M Arfan Ikram, Meike W Vernooij, and Esther E Bron. A hybrid deep learning framework for integrated segmentation and registration: evaluation on longitudinal white matter tract changes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 645–653. Springer, 2019.

[27] Yuan Liang, Weinan Song, JP Dym, Kun Wang, and Lei He. Comparenet: Anatomical segmentation network with deep non-local label fusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 292–300. Springer, 2019.

[28] 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.

[29] Alexey A Novikov, David Major, Maria Wimmer, Dimitrios Lenis, and Katja Bühler. Deep sequential segmentation of organs in volumetric medical scans. *IEEE transactions on medical imaging*, 38(5):1207–1215, 2018.

[30] Rudra PK Poudel, Pablo Lamata, and Giovanni Montana. Recurrent fully convolutional neural networks for multi-slice mri cardiac segmentation. In *Reconstruction, segmentation, and analysis of medical images*, pages 83–94. Springer, 2016.

[31] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[32] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

[33] Liesbeth Vandewinckele, Siri Willems, David Robben, Julie Van Der Veen, Wouter Crijns, Sandra Nuyts, and Frederik Maes. Segmentation of head-and-neck organs-at-risk in longitudinal ct scans combining deformable registrations and convolutional neural networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 1–10, 2019.

[34] Shuang Wang, Xiuli Ma, Xiangrong Zhang, and Licheng Jiao. Watershed-based textural image segmentation. In *2007 International Symposium on Intelligent Signal Processing and Communication Systems*, pages 312–315. IEEE, 2007.

[35] Jiong Wu, Yue Zhang, and Xiaoying Tang. A joint 3d+ 2d fully convolutional framework for subcortical segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 301–309. Springer, 2019.

[36] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.

[37] Zhenlin Xu and Marc Niethammer. Deepatlas: Joint semi-supervised learning of image registration and segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 420–429. Springer, 2019.

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

[39] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270, 2019.

[40] Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57, 2001.

[41] 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.