Joint Sequence Learning and Cross-Modality Convolution for 3D Biomedical Segmentation

Kuan-Lun Tseng[1], Yen-Liang Lin[2], Winston Hsu[1] and Chung-Yang Huang[1]

[1]National Taiwan University, Taipei, Taiwan
[2]GE Global Research, NY, USA
tkuanlun350@gmail.com, yenlianglintw@gmail.com, whsu@ntu.edu.tw, cyhuang@ntu.edu.tw

Abstract

Deep learning models such as convolutional neural network have been widely used in 3D biomedical segmentation and achieve state-of-the-art performance. However, most of them often adapt a single modality or stack multiple modalities as different input channels. To better leverage the multi-modalities, we propose a deep encoder-decoder structure with cross-modality convolution layers to incorporate different modalities of MRI data. In addition, we exploit convolutional LSTM to model a sequence of 2D slices, and jointly learn the multi-modalities and convolutional LSTM in an end-to-end manner. To avoid converging to the certain labels, we adopt a re-weighting scheme and two-phase training to handle the label imbalance. Experimental results on BRATS-2015 [13] show that our method outperforms state-of-the-art biomedical segmentation approaches.

1. Introduction

3D image segmentation plays a vital role in biomedical analysis. Brain tumors like gliomas and glioblastomas have different kinds of shapes, and can appear anywhere in the brain, which make it challenging to localize the tumors precisely. Four different modalities of magnetic resonance imaging (MRI) image are commonly referenced for the brain tumor surgery: T1 (spin-lattice relaxation), T1C (T1-contrasted), T2 (spin-spin relaxation), and FLAIR (fluid attenuation inversion recovery). Each modality has distinct responses for different tumor tissues. We leverage multiple modalities to automatically discriminate the tumor tissues for assisting the doctors in their treatment planning.

Recently, deep learning methods have been adopted in biomedical analysis and achieve the state-of-the-art performance. Patch-based methods [19, 10] extract small patches

of an image (in a sliding window fashion) and predict the label for each central pixel. These methods suffer from slow training, as the features from the overlapped patches are recomputed. Besides, they only take a small region into a network, which ignores the global structure information (*e.g.*, image contents and label correlations). Some methods apply 2D segmentation to 3D biomedical data [20] [5]. They slice a 3D medical image into several 2D planes and apply 2D segmentation for each 2D plane. 3D segmentation is then generated by concatenating the 2D results. However, these 2D approaches ignore the sequential information between consecutive slices. For examples, there may have rapid shape changes in the consecutive depths. 3D-based approaches [14] use 3D convolution to exploit different views of a 3D image. But, they often require a larger number of parameters and are prone to overfitting on the small training dataset. The above methods often stack modalities as different input channels for deep learning models, which do not explicitly consider the correlations between different modalities. To address these problems, we propose a new deep encoder-decoder structure that incorporates spatial and sequential information between slices, and leverage the responses from multiple modalities for 3D biomedical segmentation.

Figure 1 shows the system overview of our method. Given a sequence of slices of multi-modal MRI data, our method accurately predicts the different types of tumor issues for each pixel. Our model consists of three main parts: multi-modal encoder, cross-modality convolution and convolutional LSTM. The slices from different modalities are stacked together by the depth values (b). Then, they pass through different CNNs in the multi-modal encoder (each CNN is applied to a different modality) to obtain a semantic latent feature representation (c). Latent features from multiple modalities are effectively aggregated by the proposed cross-modality convolution layer (d). Then, we leverage convolutional LSTM to better exploit the spatial and se-
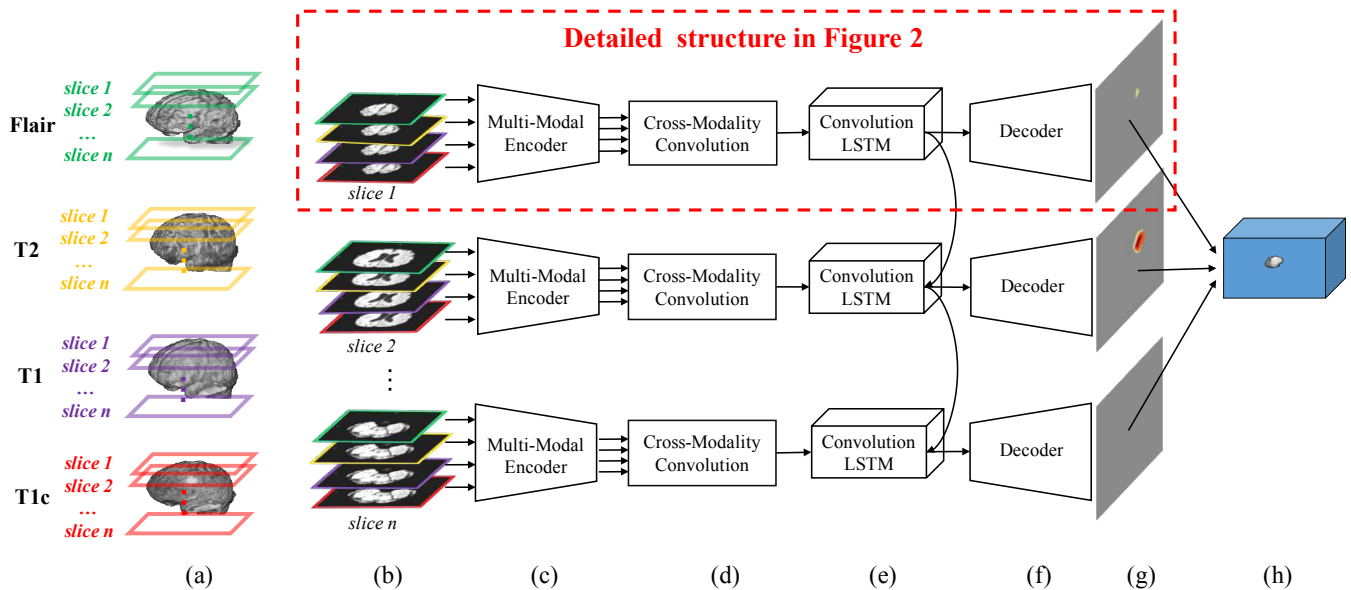
Figure 1. System overview of proposed method for 3D biomedical segmentation. (a) We extract the slices (according to the depth values) from different modalities (i.e., Flair, T2, T1, T1c) of a 3D MRI image. (b) The slices of the same depth from different modalities are combined together. (c) Each slice in a stack is fed into the multi-modal encoder for learning a latent semantic feature representation. (d) A cross-modality convolution is utilized to aggregate the information from different modalities. (e) We leverage convolutional LSTM to better model the correlations between consecutive slices. (f) A decoder network is used to up-sample the outputs of convolutional LSTM to the original resolution as the input slices. (g) The final results (i.e., different types of tumor issues) are predicted at each pixel location. (h) We stack the sequence of 2D prediction results into a 3D segmentation.

quential correlations of consecutive slices (e). A 3D image segmentation is generated (h) by concatenating a sequence of 2D prediction results (g). Our model jointly optimizes the slice sequence learning and multi-modality fusion in an end-to-end manner. The main contributions of this paper are summarized as following:

- We propose an end-to-end deep encoder-decoder network for 3D biomedical segmentation. Experimental results demonstrate that we outperform state-of-the-art 3D biomedical segmentation methods.

- We propose a new cross-modality convolution to effectively aggregate the multiple resolutions and modalities of MRI images.

- We leverage convolution LSTM to model the spatial and sequential correlations between slices, and jointly optimize the multi-modal fusion and convolution LSTM in an end-to-end manner.

2. Related Work

Image Semantic Segmentation. Various deep methods have been developed and achieve significant progress in image segmentation [15, 2, 7, 18]. These methods use convolution neural network (CNN) to extract deep representations and up-sample the low-resolution feature maps

to produce the dense prediction results. SegNet [2] adopts an encoder-decoder structure to further improve the performance while requiring fewer model parameters. We adopt the encoder-decoder structure for 3D biomedical segmentation and further incorporate cross-modality convolution and convolutional LSTM to better exploit the multi-modal data and sequential information for consecutive slices.

3D Biomedical Image Segmentation. There have been much research work that adopts deep methods for biomedical segmentation. Havaei et al. [10] split 3D MRI data into 2D slices and crop small patches at 2D planes. They combine the results from different-sized patches and stack multiple modalities as different channels for the label prediction. Some methods utilize full convolutional network (FCN) structure [15] for 3D biomedical image segmentation. U-Net [20] consists of a contracting path that contains multiple convolutions for downsampling, and a expansive path that has several deconvolution layers to up-sample the features and concatenate the cropped feature maps from the contracting path. However, depth information is ignored by these 2D-based approaches.

To better use the depth information, Lai et al. [14] utilize 3D convolution to model the correlations between slices. However, 3D convolution network often requires a larger number of parameters and is prone to overfitting on small

dataset. kU-Net [6] is the most related to our work. They adopt U-Net as their encoder and decoder and use recurrent neural network (RNN) to capture the temporal information. Different from kU-Net, we further propose a cross-modality convolution to better combine the information from multi-modal MRI data, and jointly optimize the slice sequence learning and cross-modality convolution in an end-to-end manner.

Multi-Modal Images. In brain tumor segmentation, multi-modal images are used to identify the boundaries between the tumor, edema and normal brain tissue. Cai et al. [4] combine MRI images with diffusion tensor imaging data to create an integrated multi-modality profile for brain tumors. Their brain tissue classification framework incorporates intensities from each modality into an appearance signature of each voxel to train the classifiers. Menze et al. [17] propose a generative probabilistic model for reflecting the differences in tumor appearance across different modalities. In the process of manual segmentation of a brain tumor, different modalities are often cross-checked to better distinguish the different types of brain tissue. For example, according to Menze et al. [16], the edema is segmented primarily from T2 images and FLAIR is used to cross-check the extension of the edema. Also, enhancing and non-enhancing structures are segmented by evaluating the hyper-intensities in T1C.

Existing CNN-based methods (e.g., [19, 10]) often treat modalities as different channels in the input data. However, the correlations between them are not well utilized. To our best knowledge, we are the first to jointly exploit the correlations between different modalities, and the spatial and sequential dependencies for consecutive slices.

3. Method

Our method is composed of three parts, multi-modal encoder and decoder, cross-modality convolution, and convolution LSTM. Encoder is used for extracting the deep representation of each modality. Decoder up-samples the feature maps to the original resolution for predicting the dense results. Cross-modality convolution performs 3D convolution to effectively combine the information from different modalities. Finally, convolutional LSTM further exploits the sequential information between consecutive slices.

3.1. Encoder and Decoder

Due to the small size of BRATS-2015 training dataset [13], we want the parameter space of our multi-modal encoder and decoder to be relatively small for avoiding the overfitting. Figure 2 shows our multi-modal encoder and decoder structure. We adopt the similar architecture as in SegNet [2] for our encoder, which comprises four convolution layers and four max pooling layers. Each convolution layer uses the kernel size 3×3 to produce a set of feature

maps, which are further applied by a batch normalization layer [12] and an element-wise rectified-linear non-linearity (ReLU). Batch normalization layer is critical for training our network, as the distributions of tumor and non-tumor tissues can vary from one slice to another even in the same brain. Then, a max pooling layer with size 2 and stride 2 is applied and the output feature maps are down-sampled by a factor of 2.

For the decoder network, each deconvolution layer performs the transposed convolution. Then, a convolution and batch normalization are applied. After up-sampling the feature maps to the original resolution, we pass the output of the decoder to a multi-class soft-max classifier to produce the class probabilities of each pixel.

3.2. Multi-Resolution Fusion (MRF)

Recent image segmentation models [8, 15, 9, 20] fuse multi-resolution feature maps with the concatenation. Feature concatenation often requires additional learnable weights because of the increase of channel size. In our method, we use the feature multiplication instead of concatenation. The multiplication does not increase feature map size and therefore no additional weights are required. We combine the feature maps from the encoder and decoder networks, and train the whole network end-to-end. The overall structure of cross-modality convolution (CMC) and multi-resolution fusion are shown in Figure 3. We perform CMC after each pooling layer in the multi-modal encoder, and multiply it with the up-sampled feature maps from the decoder to combine multi-resolution and multi-modality information. We will explain the details of CMC layer in the next section.

3.3. Cross-Modality Convolution (CMC)

We propose a cross-modality convolution (CMC) to aggregate the responses from all modalities. After the multi-modal encoder, each modality is encoded to a feature map of size $h \times w \times C$, where w and h are feature dimensions, and C is the number of channels. We stack the features of the same channels from four modalities into one stack. After reshaping, we have $C \times 4 \times h \times w$ feature maps. Our cross-modality convolution performs 3D convolution with the kernel size $4 \times 1 \times 1$, where 4 is the number of modalities. As the 3D kernel convolves across different stacks, it assigns different weights to each modality and sums the feature values in the output feature maps. The proposed cross-modality convolution combines the spatial information of each feature map and models the correlations between different modalities.

3.4. Slice Sequence Learning

We propose an end-to-end slice sequence learning architecture to capture the sequential dependencies. We use im-
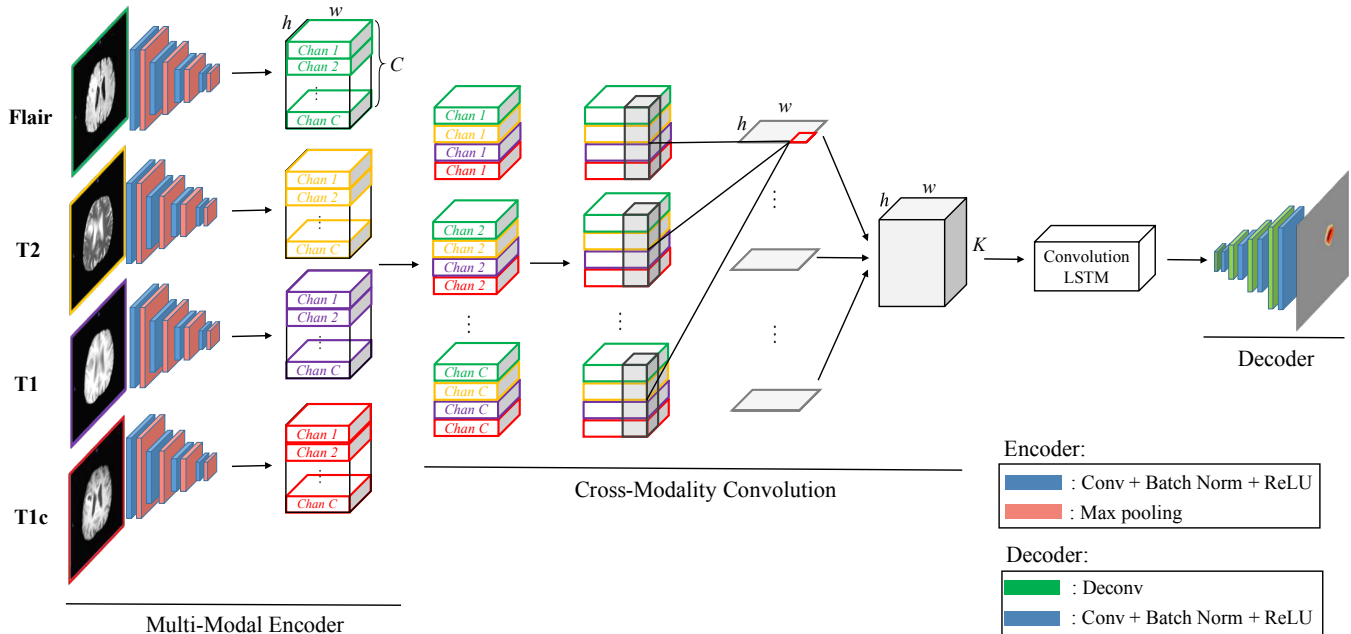
Figure 2. Cross-modality convolution is a weighting function across four different modalities. The features extracted by the multi-modal encoder have the shape size $4 \times h \times w \times C$. We reshape the input features to $C \times h \times w \times 4$ and apply cross-modality convolution with kernel size $4 \times 1 \times 1$ to produce a set of feature maps. The feature maps are fed into convolutional LSTM and decoder to generate the final dense predictions.

age depths as a sequence of slices and leverage convolutional LSTM [22] (convLSTM) to model the slice dependencies. Different from traditional LSTM [11], convLSTM replaces the matrix multiplication by the convolution operators in state-to-state and input-to-state transitions, which preservers the spatial information for long-term sequences.

Convolutional LSTM (convLSTM). The mechanism is similar to the traditional LSTM except that we replace the matrix multiplication by a convolution operator $*$. The overall network is defined as following:

$$i_t = \sigma(x_t * W_{xi} + h_{t-1} * W_{hi} + b_i)$$
$$f_t = \sigma(x_t * W_{xf} + h_{t-1} * W_{hf} + b_f)$$
$$c_t = c_{t-1} \circ f_t + i_t \circ \tanh(x_t * W_{xc} + h_{t-1} * W_{hc} + b_c)$$
$$o_t = \sigma(x_t * W_{xo} + h_{t-1} * W_{ho} + b_o)$$
$$h_t = o_t \circ \tanh(c_t)$$

Where σ is a sigmoid function and $\tanh$ is a hyperbolic tangent function. There are three gates, namely input gate i_t, forget gate f_t and output gate o_t. The forget gate controls whether to remember previous cell state c_{t-1} by the output of activation function σ. Similarly, input gate controls whether new candidate value should be added into new cell state c_t. Finally, the output gates controls which parts we want to produce. The output size of feature maps depends on the kernel size and padding methods.

Our slice sequence learning architecture combines the

encoder-decoder network with a convLSTM to better model a sequence of slices. ConvLSTM takes a sequence of consecutive brain slices encoded by multi-modal encoder and CMC (Figure 1 (e)). The weights in convLSTM are shared for different slices, therefore the parameter size does not increase linearly as the length of slice sequence growths. The output feature maps of convLSTM are up-sampled by the decoder network (Figure 1 (f)).

4. Experiments

We conduct experiments on two datasets to show the utility of our model. We compare our cross-modality convolution with the traditional methods that stack modalities as different channels. We evaluate our sequence learning scheme on typical video datasets to verify our method for modeling the temporal dependency. We also evaluate our methods on a 3D biomedical image segmentation dataset.

4.1. Dataset

CamVid dataset. Cambridge-driving labelled video database (CamVid) dataset [3] is captured from the perspective of a driving automobile with fixed-position CCTV-style cameras. CamVid provides videos with object class semantic labels. The database provides ground truth labels that associate each pixel with one of 32 semantic classes. We split the CamVid dataset with 367 training, 100 vali-
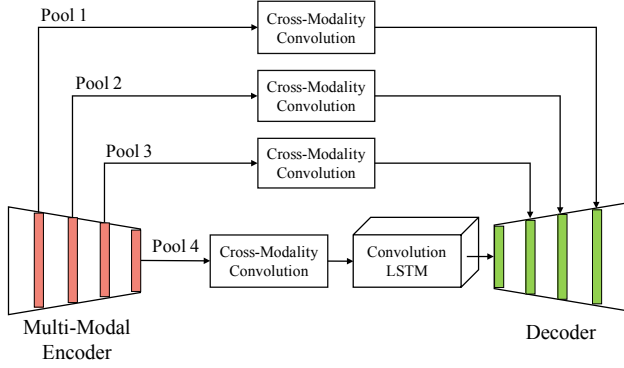
Figure 3. System overview of our multi-resolution fusion strategy. In the multi-modal encoder, the feature maps generated after each pooling layer are applied with cross-modality convolution to aggregate the information between modalities. Following that, those feature maps are multiplied with the up-sampled feature maps from the decoder to combine the multi-resolution information.

dation and 233 testing images. The evaluation criteria is the mean intersection over union (Mean IU). Mean IU is a commonly-used segmentation performance measurement that calculates the ratio of the area of intersection to the area of unions. Selected images are sampled from the original videos and down-sampled to 640×480. The length between two consecutive image is 30 frames long.

BRATS-2015 dataset. BRATS-2015 training dataset comprises of 220 subjects with high grade gliomas and 54 subjects with low grade gliomas. The size of each MRI image is $155 \times 240 \times 240$. We use 195 high grade gliomas and 49 low grade gliomas for training, and the rest 30 subjects for testing. We also conduct five-fold evaluation by using BRATS-2015 online judge system for avoiding overfitting. All brain in the dataset have the same orientation and the four modalities are synchronized. The label image contains five labels: non-tumor, necrosis, edema, non-enhancing tumor and enhancing tumor. The evaluation system separates the tumor structure into three regions due to practical clinical applications.

- Complete score: it considers all tumor areas and evaluates all labels 1, 2, 3, 4 (0 for normal tissue, 1 for edema, 2 for non-enhancing core, 3 for necrotic core, and 4 for enhancing core).

- Core score: it only takes tumor core region into account and measures the labels 1, 3, 4.

- Enhancing score: it represents the active tumor region, i.e., only containing the enhancing core (label 4) structures for high-grade cases.

There are three kinds of evaluation criteria: Dice, Positive Predicted Value and Sensitivity.

$$Dice = \frac{P_1 \bigcap T_1}{(P_1 + T_1)/2}$$

$$PPV = \frac{P_1 \bigcap T_1}{P_1}$$

$$Sensitivity = \frac{P_1 \bigcap T_1}{T_1},$$

where T is ground truth label and P is predicted result. T_1 is the true lesion area and P_1 is the subset of voxels predicted as positives for the tumor region.

4.2. Training

Single Slice Training. The critical problem in training a fully convolutional network in BRATS-2015 dataset is that the label distribution is highly imbalanced. Therefore, the model easily converges into local minimum, i.e., predicting every pixel as non-tumor tissue. We use *median frequency balancing* [8] for handling the data imbalance, where the weight assigned to a class in the cross-entropy loss function is defined as:

$$\alpha_c = median\_freq/freq(c)$$

where $freq(c)$ is the number of pixels of class c divided by the total number of pixels in images where c is present, and $median\_freq$ is the median of all class frequencies. Therefore the dominant labels will be assigned with the lowest weight which balances the training process.

In our experiments, we find that weighted loss formulation will overly suppress the learning of dominant labels (e.g., label 0 for normal tissues) and wrongly predict the labels.

Two-Phase Training. In the first phase, we only sample the slices that contain tumor tissues and use *median frequency balancing* method for de-weighting the losses of the dominant classes (e.g., normal tissue and background). In the second phase, we replace the *median frequence balancing* and use a lower learning rate (*i.e.*, 10e-6 in our experiments) for training the model. With the true distribution reflected on loss function, we can train our model in a more balanced way to preserve diversity and adjust the output prediction to be closer to the real distribution of our data. Two-phase training alleviates the problem of overly suppressing the dominant classes and learn much better results.

Slice Sequence Training. We avoid sampling the empty sequences (all the slices in the sequence are normal brain tissues) in the first training phase to prevent the model from getting trapped into local minimum and apply two-phase training scheme for slice sequence leanring.

Method	label 0	label 1	label 2	label 3	label 4	MeanIU
U-Net [20]	92.3	42.9	73.6	45.3	62.0	54.2
U-Net + two phase	98.6	43.8	67.4	24.0	60.5	59.3
MME + MRF + CMC	98.2	47.0	72.2	41.0	72.2	61.8
MME + MRF + CMC + two-phase	**99.1**	48.8	63.8	36.2	76.9	64.0
MME + MRF + CMC + convLSTM	96.6	**94.3**	71.2	32.8	**96.0**	62.5
MME + MRF + CMC + convLSTM + two-phase	98.5	92.1	**77.3**	**55.9**	78.6	**73.5**

Table 1. MeanIU on our BRATS-2015 testing set with 30 unseen patients. The results show that the proposed model: MME + MRF + CMC and MME + MRF + CMC + convLSTM achieve the best performance and outperform the baseline method U-Net [20]. Two-phase training better handles the label imbalance problem and significantly improves the results.

Method	MeanIU
SegNet [2]	47.85
Encoder-Decoder + convLSTM	48.16
Encoder-Decoder + MRF	49.91
Encoder-Decoder + convLSTM + MRF	**52.13**

Table 2. MeanIU on CamVid test set. Our encoder-decoder model with convolutional LSTM and multi-resolution fusion achieve the best results.

Method	MeanIU
U-Net [20]	54.3
U-Net+two-phase	59.3
Encoder-Decoder	44.14
MME + CMC	45.80
Encoder-Decoder + MRF	55.37
MME + CMC + MRF	61.83
MME + CMC + MRF + two-phase	64.02
MME + CMC + MRF + convLSTM	62.15
MME + CMC + MRF + convLSTM + two phase	**73.52**

Table 3. Segmentation results of our models on BRATS-2015 testing set with 30 unseen patients.

For training the convolutional LSTM, we adopt the orthogonal initialization [21] for handling the missing gradient issue. For CamVid dataset, we use the batch size of 5 and sequence length of 3. For BRATS-2015 dataset, we use the batch size of 3 and sequence length of 3. The initial learning rate is 10e-4 and we use Adam optimizer for training the network.

4.3. Baseline

The most relevant work to our method are kU-Net [6] and U-Net [20]. Both models achieve the state-of-the-art results in 3D biomedical image segmentation. However, kU-Net is not originally designed for brain tumor segmentation, and the source code is not publicly available. Therefore, we compare our method with U-Net, which shows competitive performance with kU-Net. Original implementation of U-Net does not adopt batch normalization. However, we find that it can not converge when training on BRATS-2015 dataset. Thus, we re-implement their model with Tensorflow [1] and incorporate batch normalization layer before every non-linearity in the contracting and expansive path of U-Net model. We use orthogonal initialization and set the batch size to 10. The inputs for U-Net is 4-channel MRI slices that stack four modalities into different channels. We also investigate two-phase training and re-weighting for U-Net.

4.4. Experimental Results

We conduct the experiments to evaluate cross-modality convolution (CMC). We compare the performance of our multi-modal encoder and CMC layers with an encoder-decoder model (see Table 3). The encoder-decoder model refers to a single encoder and decoder network without fusion. The input of the encoder-decoder model stacks different modalities as different channels, while the input of our MME+CMC is illustrated in Figure 1(b). The performance of our MME+CMC outperforms the basic encoder-decoder structure by approximately two percent in Mean IU. Currently, the feature maps extracted by MME are down-sampled to a lower resolution, thus a certain amount of spatial information is lost. We conjecture that the performance of our CMC could be further improved by incorporating higher resolution feature maps.

We conduct another experiment by using multi-resolution feature maps with CMC to verify whether multiple resolution helps. In Table 3, we can observe that MRF significantly improves the performance, e.g., encoder-decoder with MRF improves the basic encoder-decoder by 10 percent. We also evaluate our feature multiplication and feature concatenation used by U-Net, and find that they achieve similar performance.

Table 1 shows that MME+CMC+MRF outperforms U-Net (similar to our encoder-decoder+MRF) on Mean IU (almost every label except for label 0 (normal tissues)). Because of the number of normal tissues are much larger than the other labels, the accuracy of label 0 is critical in Mean IU metric. As a result, we use two-phase training to refine our final prediction. After the second phase of training, the accuracy for label 0 is improved and our model shows much clean prediction results (cf. Figure 4).

To verify the generalizability of our sequence learning method, we further perform the slice-to-sequence experi-
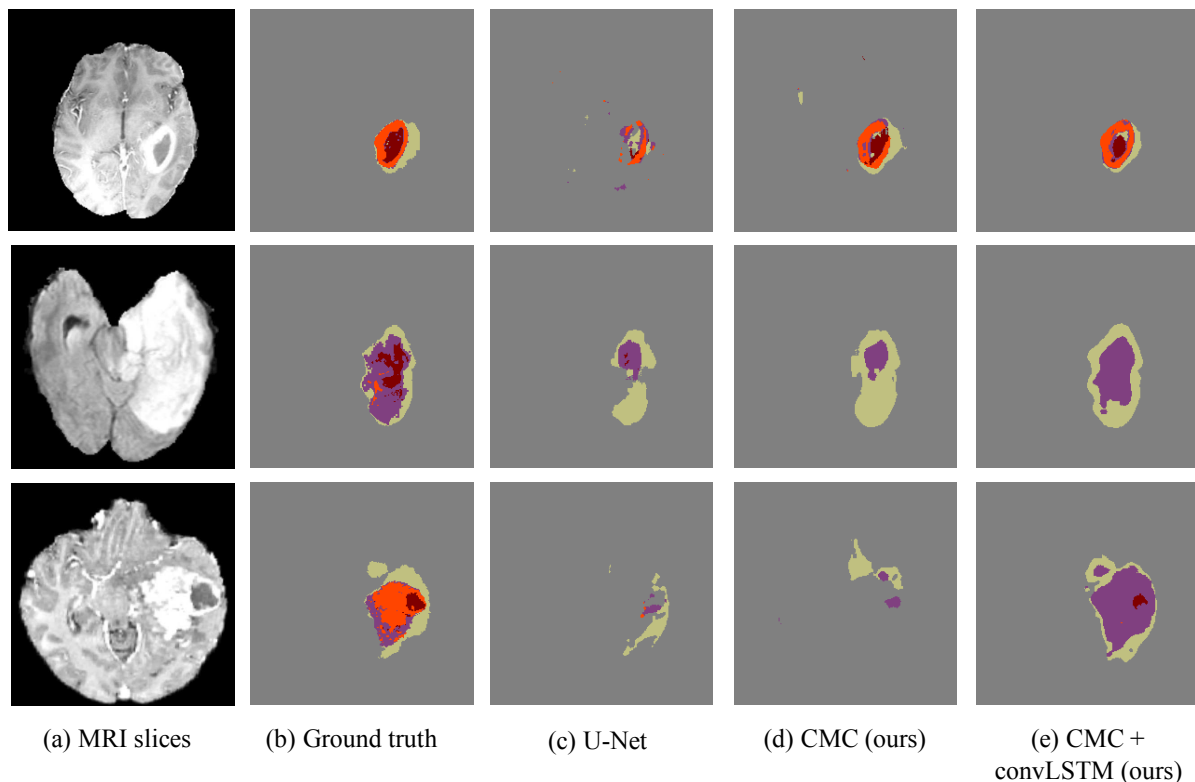
| | (a) MRI slices | (b) Ground truth | (c) U-Net | (d) CMC (ours) | (e) CMC + convLSTM (ours) |

Figure 4. Segmentation results of variants of our method and U-Net [20] baseline.

Method	Dice			Sensitivity			PPV		
	Complete	Core	Enhancing	Complete	Core	Enhancing	Complete	Core	Enhancing
U-Net [20]	0.8504	0.6174	0.6793	0.8727	0.5296	0.7229	0.8376	0.7876	0.7082
Ours	0.8522	0.6835	0.6877	0.8741	0.6545	0.7735	0.9117	0.7945	0.7212

Table 4. Five-fold cross validation segmentation results on 30 unseen patient evaluated by BRATS-2015 online system. Our model uses MME + MRF + CMC + convLSTM + two-phase settings, and outperforms U-Net [20] in different measurements.

ments on CamVid dataset. The sequence length used in the CamVid experiment is three and the settings for encoder-decoder are the same as BRATS dataset. We incorporate convolutional LSTM with both basic encoder-decoder model and encoder-decoder+MRF. Results show that convolutional LSTM consistently improves the performance for both settings (cf. Table 2) and outperforms SegNet [2]. Due to the ability of convolutional LSTM for handling long-short term sequences and preserving the spatial information at the same time, the dependencies between slices are well learned.

Our slice-to-sequence also improves the results of BRATS dataset. In Table 1, we can see that the accuracy of label 1 to 4 (row 5 and 6) is much better than the single slice training (row 3 and 4). After the second training phase, the accuracy of label 0 is improved and the model achieves the Mean IU 73.52, which outperforms single slice training model by a large margin.

In Table 4, we compare our slice-to-sequence model with U-Net on BRATS-2015 online system (we upload the five fold cross validation results to BRATS-2015 online evaluation system). Two-phase training is applied to both methods and trained with the same epochs (without post-processing). Our slice-to-sequence model outperforms U-Net in different measurements. The visualized results also show that sequential information improves the predictions for detailed structures (cf. Figure 4).

Experimental results show that the proposed cross-modality convolution can effectively aggregate the information between modalities and seamlessly work with multi-resolution fusion. The two components are combined to achieve the best results. The slice-to-sequence architecture further utilizes the sequential dependencies to refine the segmentation results. This end-to-end trainable architecture shows many potentials since it provides consistent improvement on every configurations of our model.

5. Conclusions

In this paper, we introduce a new deep encoder-decoder architecture for 3D image biomedical segmentation. We present a new cross-modality convolution to better exploit the multi-modalities, and a sequence learning method to integrate the information from consecutive slices. We jointly optimizing sequence learning and cross-modality convolution in an end-to-end manner. Experimental results on BRATS-2015 dataset demonstrate that our method improves the state-of-the-art methods.

6. Acknowledgement

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 6

[2] V. Badrinarayanan, A. Handa, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv preprint arXiv:1505.07293*, 2015. 2, 3, 6, 7

[3] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV*, 2008. 4

[4] H. Cai, R. Verma, Y. Ou, S.-k. Lee, E. R. Melhem, and C. Davatzikos. Probabilistic segmentation of brain tumors based on multi-modality magnetic resonance images. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2007. 3

[5] H. Chen, X. Qi, L. Yu, and P.-A. Heng. Dcan: Deep contour-aware networks for accurate gland segmentation. *arXiv preprint arXiv:1604.02677*, 2016. 1

[6] J. Chen, L. Yang, Y. Zhang, M. Alber, and D. Z. Chen. Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation. *NIPS*, 2016. 3, 6

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 2

[8] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, 2015. 3, 5

[9] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*, 2015. 3

[10] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 2016. 1, 2, 3

[11] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 1997. 4

[12] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 3

[13] M. Kistler, S. Bonaretti, M. Pfahrer, R. Niklaus, and P. Büchler. The virtual skeleton database: An open access repository for biomedical research and collaboration. *Journal of Medical Internet Reserach*, 2013. 1, 3

[14] M. Lai. Deep learning for medical image segmentation. *arXiv preprint arXiv:1505.02000*, 2015. 1, 2

[15] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2, 3

[16] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 2015. 3

[17] B. H. Menze, K. Van Leemput, D. Lashkari, M.-A. Weber, N. Ayache, and P. Golland. A generative model for brain tumor segmentation in multi-modal images. In *MICCAI*, 2010. 3

[18] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *ICCV*, 2015. 2

[19] S. Pereira, A. Pinto, V. Alves, and C. Silva. Brain tumor segmentation using convolutional neural networks in mri images. *IEEE Transactions on Medical Imaging*, 2016. 1, 3

[20] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 1, 2, 3, 6, 7

[21] A. M. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013. 6

[22] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NIPS*, 2015. 4