Focus, Segment and Erase: An Efficient Network for Multi-Label Brain Tumor Segmentation

Xuan Chen[1*][0000−0002−6570−1049], Jun Hao Liew[1*][0000−0002−7538−6759], Wei Xiong[2], Chee-Kong Chui[1][0000−0001−9463−4781], and Sim-Heng Ong[1][0000−0003−2766−8150]

[1] National University of Singapore
{xuan.chen,liewjunhao}@u.nus.edu {mpecck,eleosh}@nus.edu.sg
[2] Institute for Infocomm Research
{wxiong@i2r.a-star.edu.sg}

Abstract. In multi-label brain tumor segmentation, class imbalance and inter-class interference are common and challenging problems. In this paper, we propose a novel end-to-end trainable network named FSENet to address the aforementioned issues. The proposed FSENet has a tumor region pooling component to restrict the prediction within the tumor region ("focus"), thus mitigating the influence of the dominant non-tumor region. Furthermore, the network decomposes the more challenging multi-label brain tumor segmentation problem into several simpler binary segmentation tasks ("segment"), where each task focuses on a specific tumor tissue. To alleviate inter-class interference, we adopt a simple yet effective idea in our work: we erase the segmented regions before proceeding to further segmentation of tumor tissue ("erase"), thus reduces competition among different tumor classes. Our single-model FSENet ranks 3^{rd} on the multi-modal brain tumor segmentation benchmark 2015 (BraTS 2015) without relying on ensembles or complicated post-processing steps.

Keywords: Brain Tumor Segmentation · Convolutional Neural Network · Class Imbalance · Inter-Class Interference

1 Introduction

Brain tumor, though not a common disease, severely harms the health of patients and causes high mortality. Automatic brain tumor segmentation would greatly assist medical diagnosis and treatment planning, since manual segmentation is time-consuming and requires a high degree of professional expertise. The segmentation task is very challenging due to the diversity of the tumors in terms of their location, shape, size and contrast, which restrict the application of strong priors. Hence, researchers have spent much time and effort in studying this topic.

*Authors contributed equally

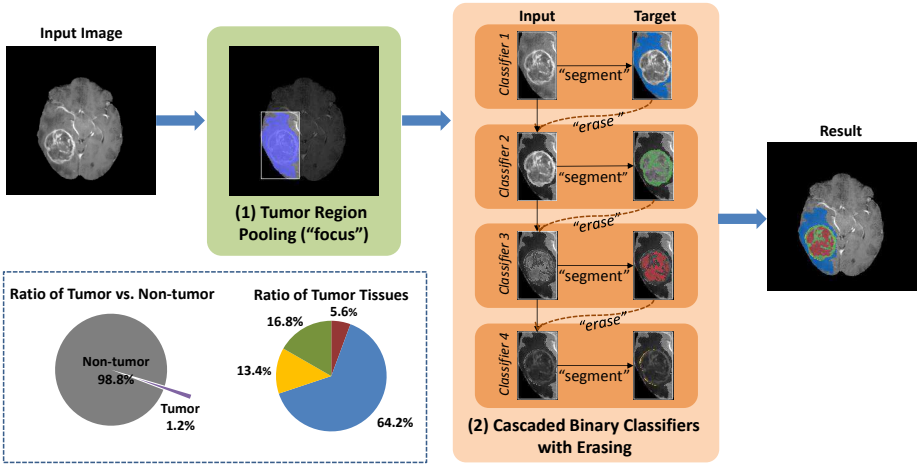

Fig. 1. Overview of FSENet. The operations are shown on the input image for illustrative purpose. (1) The tumor region pooling first extracts the tumor region ("focus"). (2) Within the tumor region, each classifier sequentially segments a target tissue ("segment"), and erases it before proceeding to the next classifier ("erase"), forming a cascaded framework. Finally, all segmented tissues are combined to produce the multi-label result. Blue, green, red and yellow indicate edema, enhancing core, necrosis and non-enhancing core respectively. The darkened regions represent the erased areas. The pie charts illustrate the class imbalance problem (Better viewed in color)

The approaches for brain tumor segmentation can be generally categorized into two classes, *i.e.,* generative methods and discriminative methods. Generative methods [19, 23], which model tumor anatomy and appearance statistics explicitly, usually have better generalization ability, but require more professional knowledge and elaborate pre-processing steps. Discriminative methods [1, 4, 7, 11, 22], though relying heavily on the quality of training data, can learn task-relevant demands from human-labeled data directly.

An example of discriminative methods is machine learning, which has been successfully applied in this field. Before the advent of the deep learning era, traditional machine learning approaches typically rely on the dedicated selection of hand-crafted features, for example, first-order textures [1], histogram and spatial location[4], and a mixture of high dimensional multi-scale features[7] to achieve good performance. However, searching exhaustively for the best combination of features by trial-and-error is not feasible. Deep convolutional neural networks (DCNNs), on the other hand, are able to extract more suitable features for the task on their own by updating the networks gradually with gradient back-propagation, and thus have gained popularity in the medical image processing community [2, 5, 6, 10, 11, 14, 20, 22, 24, 25].

Common problems faced in multi-label brain tumor segmentation are class imbalance and inter-class interference. The class imbalance problem exhibits two

aspects. First, the non-tumor region may be tens, or even hundreds of times the size of a tumor lesion. Second, some tumor tissues are much larger compared to others, for example, edema *vs.* necrotic core. We plot the statistics of each class in the training set of BraTS 2015 [15, 18] as pie charts in Fig. 1 to show the class imbalance problem. The inter-class interference is caused by similar features shared among different tumor tissues, leading to difficulties in differentiating each class and also interfering with their predictions.

In this paper, we propose a novel network named FSENet that aims to address these problems. Fig. 1 shows an overview of the proposed FSENet. While segmenting each tumor tissue is highly challenging, separating the entire tumor from the non-tumor region is relatively easy. Thus, we first identify the tumor region with a whole-tumor classifier and then extract features of the tumor region with a tumor region pooling component, such that the influence of the large non-tumor region can be alleviated by discarding a large portion of negative sample features. This step tells the network where it should pay attention to, demonstrating the "focus" feature of our FSENet.

In order to reduce the inter-class interference, a simple yet effective idea is adopted: the previously segmented tumor tissues are erased before proceeding to the segmentation of the next tumor label. We first decompose the multi-label segmentation problem into several binary sub-problems, which are more specialized in discriminating specific tumor tissues. Taking this step further, we cascade our binary classifiers sequentially in an "outer-to-inner" manner according to the typical brain tumor structure, *i.e.,* edema first, followed by enhancing core, necrosis and non-enhancing core. Furthermore, an erasing process is introduced between the classifiers to erase features from the feature maps if they are confidently classified as foreground by the previous classifiers. Usually, the inner tissues, like necrosis, are more irregular in size, shape, contrast and distribution, and thus more difficult to segment compared to the outer tissues. As a result, erasing the segmented outer tissue class would reduce their interference with the prediction of the remaining more challenging labels. This step demonstrates the "segment" and "erase" features of our FSENet. To summarize, our contributions are fourfold:

- We propose a tumor region pooling component to force a prediction to be made only on the extracted tumor region, in order to suppress the negative influence from the dominant non-tumor region.
- We propose to replace one-stage multi-label segmentation with a component that consists of cascaded binary classifiers with erasing to simplify and specialize the problem, and to avoid inter-class interference.
- We develop an end-to-end training pipeline which achieves significantly performance boost over the baseline (without the proposed components) with only $\sim$1.7% overhead.
- Our single-model FSENet achieves 3^{rd} place performance on the BraTS 2015 leaderboard without heavy model ensembles or complicated post-processing steps.

2 Related Work

Class imbalance is a common problem in medical image analysis. For example, in liver computed tomography (CT) images, the lesions are several times smaller than the liver, and may only occupy a few pixels.

Various approaches have been proposed to address the class imbalance problem. One approach is to keep a reasonable ratio of positive samples and negative samples by manual oversampling or undersampling [2, 9, 22]. However, as in multi-label segmentation, this method is only applicable to a patch-based framework, but not to one that takes the entire image as input. Another typical approach is to modify the loss function [5, 20], such that the network is less sensitive to the class imbalance problem. Although [20] claims the effectiveness of using the Dice loss, it is only suitable for a binary segmentation problem. The weighted cross-entropy loss [5], unlike the Dice loss, is more flexible in that it is suitable for both binary and multi-label segmentation. However, it suffers from the elaborate selection of weighting factors.

In our approach, we use the coarse binary segmentation result to locate the tumor region, and then extract the region for fine multi-label segmentation. By extracting the tumor region, the non-tumor samples are naturally reduced and hence have less influence on the fine-grained prediction. The proposed method is implemented as a region pooling component, which is able to work within an image-based framework and requires no hyperparameter.

Sequential prediction is also one plausible way to deal with the class imbalance problem, as well as simplifying difficult one-stage multi-label segmentation by using several specialized classifiers. Its effectiveness in multi-label segmentation has been widely reported. Sequential prediction is usually implemented by cascading multiple models [2, 3, 5, 8, 10, 11, 25]. The first model performs a similar function to our proposed region pooling component, which is to identify the region of interest (RoI). The following models are trained to handle more difficult tasks with the help of the identified RoI. For example, in [2], one 3D-UNet that is used to separate the whole tumor from the non-tumor region is cascaded with a second 3D-UNet which discriminates the different brain tumor tissues. One obvious disadvantage of cascading multiple models is that the overall framework may be sub-optimal since end-to-end training is inapplicable. In addition, the deep convolution features extracted by each CNN could not be fully utilized, thus reducing computational efficiency.

In our paper, we implement sequential prediction by cascading classifiers, rather than cascading models, in such a way that the proposed FSENet can perform end-to-end training. All classifiers share features that are first extracted by a fully convolutional network, instead of having their own networks as in the case of cascading models, so that the deep convolutional features can be well utilized. Similarly, the cascaded classifiers solve the difficult multi-label segmentation problem in a more specialized and effective way. The main difference lies in the novel erasing operation introduced between classifiers, which is able to alleviate the inter-class interference that is common in medical images due to the similar features shared by different tissues. This erasing operation suppresses

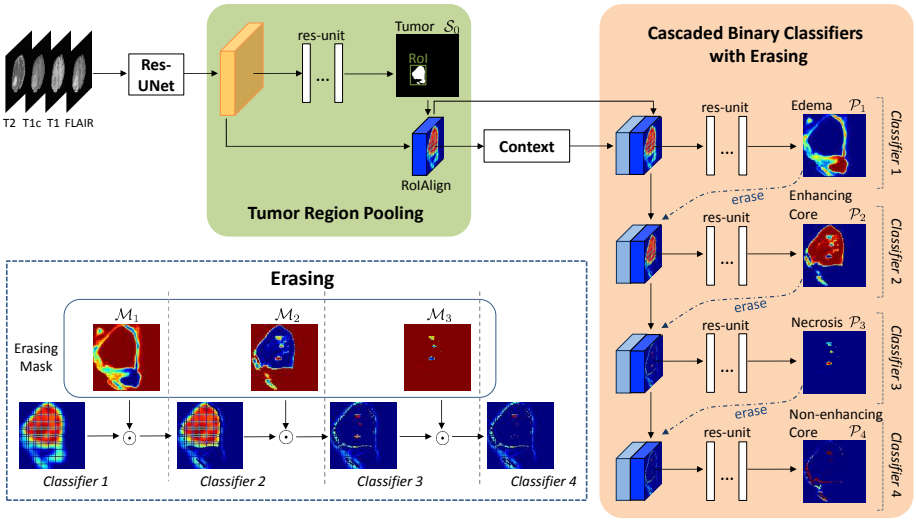

Fig. 2. Architecture of the proposed FSENet. The input goes through Res-UNet for the extraction of deep convolutional features. A whole-tumor classifier produces binary tumor/non-tumor segmentation to locate the tumor, and then the tumor region pooling component extracts the valid region from the feature maps accordingly. The extracted feature maps pass through the cascaded binary classifiers with erasing component for the segmentation of each tumor tissue in an "outer-to-inner" order. An example of the erasing process is visualized in the bottom left box. Feature multiplying the erasing mask with Hadamard product gives the erased feature maps.

the responses of regions that correspond to the confident foreground prediction produced by the previous classifiers. The classifiers are cascaded in an "outer-to-inner" manner according to the typical brain tumor structure, and this is done in such a way that the outer tissues would not interfere with the segmentation of the inner tissues.

3 FSENet

The proposed FSENet includes two novel components, *i.e.*, tumor region pooling and cascaded binary classifiers with erasing. The architecture of FSENet is shown in Fig. 2.

Following convention, the input of the proposed network is the concatenation of all four available channels, *i.e.*, contrast-enhanced T1-weighted (T1c) image, T1-weighted (T1) image, T2-weighted (T2) image and FLuid-Attenuated Inversion Recovery (FLAIR) image of each brain magnetic resonance (MR) slice to fully utilize the multi-modal information. We feed forward the input to a fully-convolutional network to extract deep convolutional features. The feature maps pass through a whole-tumor classifier, which separates the tumor and non-

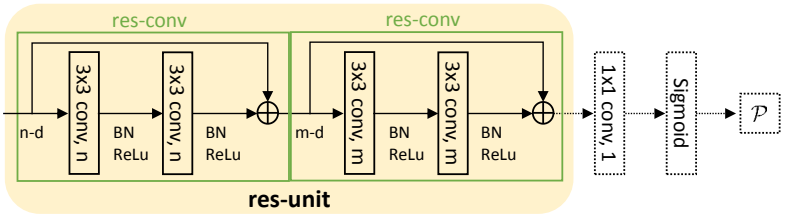

Fig. 3. Configuration of the residual unit and the classifier in the FSENet. The parts inside the green rectangle are the residual convolution block (res-conv). A residual unit (res-unit) consists of two res-convs. The parts indicated in dash line are only included in the classifier

tumor regions. According to the binary tumor/non-tumor segmentation result, the tumor region pooling module knows where to focus to extract the tumor region from the feature maps, such that the negative influence from the dominant non-tumor region can be alleviated in the subsequent predictions. The extracted feature maps are sequentially fed to the cascaded binary classifiers with erasing component such that more discriminative representation related to a specific tissue is emphasized, thus favors more accurate pixel-wise classification. The erasing operation helps to suppress inter-class interference, and hence assists the prediction of the latter class to improve overall performance. The final multi-label segmentation result is the fusion of the predictions given by the four binary classifiers, as well as the whole-tumor classifier.

3.1 Res-UNet

UNet [24] has wide applications in medical image processing [5, 6, 14]. Features and ground truth information of small and scattered tissues can totally disappear in a network whose output stride is larger than 1. Thus, UNet is an appropriate choice since the generated feature maps share the size of the input image. In order to increase the network capability without hindering the gradient back-propagation, we replace each convolution layer in UNet with a residual convolution block (res-conv) (Fig. 3) as proposed in [13], which turns UNet into its residual counterpart (Res-UNet). Res-UNet is adopted as our backbone architecture to extract deep convolutional features. However, the proposed components can be generalized to any fully convolutional network easily, and is not limited to this specific Res-UNet.

A whole-tumor binary classifier is attached to the Res-UNet to segment the entire tumor from the non-tumor region. By thresholding the prediction $\mathcal{P}_0 \in \mathbb{R}^{H\times W}$ of the whole-tumor classifier with a constant value of 0.5, the binary tumor/non-tumor segmentation result $\mathcal{S}_0 \in \mathbb{R}^{H\times W}$ can be obtained. The feature maps extracted with Res-UNet and the binary tumor/non-tumor segmentation result $\mathcal{S}_0$ are further utilized in the following components of the FSENet, which are discussed in detail below.

3.2 Tumor Region Pooling

Brain tumor tissues usually occupy a small number of pixels in the MR image, while the non-tumor region is several times larger than the tumor (Fig 1), which causes a severe class imbalance problem and hence leads to difficulties in learning. To address this problem, we propose to use RoIAlign [12] to extract the features of the tumor region from the original feature maps, so that the following classifiers only need to focus on the tumor region for subsequent fine-grained segmentation. Extracting the tumor region has two benefits. First, since the non-tumor region contains a large no-measurement area (the black region), computational resources can be saved on its segmentation since this is a relatively easy task. Second, the following fine-grained multi-label segmentation would not be hindered by the presence of the large non-tumor region.

RoIAlign locates the tumor region according to an RoI proposal, and then converts the valid region in the feature maps into RoI feature maps with fixed spatial extent $H_{\mathrm{RoI}} \times W_{\mathrm{RoI}}$. The RoI proposal is produced based on the binary tumor/non-tumor segmentation result $\mathcal{S}_0$ generated by the whole-tumor classifier. To avoid the warping problem, the RoI proposal is set to the smallest square bounding box that contains the tumor region. With the consideration of not losing too much detail during the pooling operation from a large spatial dimension to a small one, we empirically set both H_{RoI} and W_{RoI} equal to 100 in our experiments.

The RoI feature maps, which mainly contain features related to the tumor region, are then fed to the four cascaded binary classifiers with erasing to classify each pixel to its correct target class.

3.3 Cascaded Binary Classifiers with Erasing

Inter-class feature similarity and class imbalance (Fig. 1) are commonly exhibited among different tumor tissues. It would be challenging to achieve optimal multi-label segmentation in one stage according to our observation (Model 2 in Table 1). Instead of considering all the labels at the same time, we propose to divide the multi-label segmentation problem into several binary ones, thus turning the difficult one-stage task into a more tractable multi-stage task. Unlike a multi-label classifier, each binary classifier is able to learn more discriminative task-relevant representation of the target class for more accurate binary segmentation. The configuration of a binary classifier is shown in Fig. 3.

However, simply decomposing a multi-label classifier into several binary counterparts may not necessarily lead to improvement in performance. This is because the prediction of relatively small and scattered tissues, like necrotic tissue, would still be difficult due to the scarcity of positive samples and competition from other classes. Therefore, the overall performance is sub-optimal and should be improved. To address the problem, we first cascade the classifiers in an "outer-to-inner" fashion according to the typical brain tumor structure, *i.e.,* edema first, followed by enhancing core, necrosis and non-enhancing core. We introduce an

erasing process between the classifiers to erase the responses of previously segmented tissues, such that the remaining classes, which are usually more irregular in sizes, shapes, contrast and distributions, are free from the competition and interference of the earlier class. The erasing operation is multiplying, element-wise, the RoI feature maps with an erasing mask:

$$\mathcal{F}' = \mathcal{F} \odot \mathcal{M} \tag{1}$$

where $\mathcal{F}$ and $\mathcal{F}'$ are the RoI feature maps and the erased RoI feature maps respectively, $\mathcal{M}$ the erasing mask, and $\odot$ the Hadamard product.

An example demonstrating the erasing process is shown in the bottom left box in Fig. 2. The responses in the RoI feature maps are gradually erased after each binary segmentation stage, thus leading to fewer features and hence less competition and interference in subsequent segmentation of more difficult tumor tissues.

Suppose the prediction produced by each binary classifier is denoted as $\mathcal{P}_i \in \mathbb{R}^{H_{\text{RoI}} \times W_{\text{RoI}}}$, $i \in \{1, \ldots, 4\}$ with its value in the range $[0, 1]$, indicating the confidence of classifying the corresponding pixel to a target class. The erasing mask $\mathcal{M}_i$ is generated based on the prediction $\mathcal{P}_i$:

$$\mathcal{M}_i(x, y) = 1 - \mathcal{P}_i(x, y) \tag{2}$$

where $\mathcal{M}_i(x, y) \in [0, 1]$. The erasing mask is a reverse attention mask that focuses on the unsegmented regions, while suppressing the responses of the confident foreground regions predicted by previous classifiers.

Instead of totally removing the segmented region by thresholding the prediction with a specific constant value, the proposed erasing mask only suppresses their response to a certain extent according to the prediction confidence. This avoids the selection of the threshold value, and allows regions with not very high prediction confidence (usually along the boundary) to partially pass through the mask as supporting context in the following segmentations.

To provide the classifiers with a better understanding of the overall tumor structure, we introduce a context branch in our FSENet (Fig. 2), which contains the pyramid pooling module proposed in [26]. The multi-scale context information is concatenated with the erased RoI feature maps as the input of the classifier, thus providing additional reference to assist the segmentation.

3.4 Loss and Final Result

The network contains one whole-tumor classifier to identify the tumor region and four class-specific classifiers to segment different tumor tissues. For each classifier, we adopt both cross-entropy loss and Dice loss. Dice loss, which is widely used in the medical image processing community [20], can be defined as:

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2 \sum_{k=1}^{K} p_k g_k}{\sum_{k=1}^{K} p_k + \sum_{k=1}^{K} g_k} \tag{3}$$

where p_k is the prediction of a pixel, g_k, the corresponding ground truth, and K, the total number of pixels. The total loss is the weighted sum of all the losses:

$$\mathcal{L}_{\text{total}} = \sum_i \alpha_i \mathcal{L}^i + \sum_i \beta_i \mathcal{L}^i_{\text{dice}} \tag{4}$$

where $i \in \{0, \ldots, 4\}$ refers to the whole-tumor classifier and the four class-specific classifiers respectively, $\mathcal{L}^i$ the cross-entropy loss of the i^{th} classifier, and α_i and β_i the hyperparameters to emphasize or mitigate a certain loss. We assign equal importance to all losses, and hence α_i and β_i, $\forall i \in \{0, \ldots, 4\}$ are set to 1.

To generate the final multi-label segmentation result, the prediction $\mathcal{P}_i \in \mathbb{R}^{H_{\text{RoI}} \times W_{\text{RoI}}}, i \in \{1, \ldots, 4\}$ is first projected to have its original scale $\mathcal{P}'_i \in \mathbb{R}^{H \times W}, i \in \{1, \ldots, 4\}$, which represents the probability of each pixel belonging to the class. As for the non-tumor class, we have $\mathcal{P}'_0 = J - \mathcal{P}_0$, where $J \in \mathbb{R}^{H \times W}$ is an all-one matrix. Fusing the five predictions by the argmax function gives us the final multi-label segmentation result.

4 Experiment

4.1 Dataset and Experiment Settings

Dataset and Evaluation Metrics We evaluate the proposed FSENet on the multi-label brain tumor segmentation benchmark 2015 (BraTS 2015) [15, 18], which includes 4 tumor tissue categories and one non-tumor category (label=0). The 4 types of tumor tissues are necrosis (label=1), edema (label=2), non-enhancing core (label=3) and enhancing core (label=4). BraTS 2015 contains 220 high-grade glioma (HGG) cases and 54 low-grade glioma (LGG) cases in the training set, and 110 mixture cases of HGG and LGG in the test set. Each case includes four volumes, which correspond to the four modalities, *i.e.*, T1, T1c, T2 and FLAIR. A volume consists of 155 MR images of size 240×240. The performance is evaluated in terms of the Dice similarity score (Dice), positive prediction value (PPV), and sensitivity (Sens) over three predefined regions, *i.e.* whole tumor (label 1+2+3+4), tumor core (label 1+3+4) and active tumor (label 4). Dice score, PPV and sensitivity are defined respectively as:

$$\begin{aligned} \text{Dice}(P, T) &= \frac{2|P_1 \cap T_1|}{|P_1| + |T_1|} \\ \text{PPV}(P, T) &= \frac{|P_1 \cap T_1|}{|P_1|} \\ \text{Sens}(P, T) &= \frac{|P_1 \cap T_1|}{|T_1|} \end{aligned} \tag{5}$$

where $P \in \{0, 1\}$ is the prediction, $T \in \{0, 1\}$ the ground truth, P_1 and T_1 the sets of pixels where $P = 1$ and $T = 1$ respectively, and $|\cdot|$ the size of the set.

Training/Testing Settings In the training phase, only the slices that contain tumor tissue labels are used (19676 slices). We train our FSENet from

scratch with mini-batch size equals to 2. The T1, T1c, T2 and FLAIR images that correspond to the same brain MR slice are concatenated, forming a 4-channel input to the model. All patient cases in the training and test sets are pre-processed to correct for intensity inhomogeneity with a learning based two-step standardization [21]. In the test phase, for each patient case, all 155 slices are fed into the network for inference.

Our implementation is based on the PyTorch[1] platform using a NVIDIA GeForce TITAN Xp GPU with 12GB memory. The initial learning rate is set to 1×10^{-3} and decreased by a factor of 10 after 15 epochs. We train the FSENet for 25 epochs in total before deployment. To facilitate learning, we use ground truth to generate RoI proposal during training. However, the masks are always generated based on the predictions of the network stated in Equation (2). We use stochastic gradient descent (SGD) with momentum and weight decay set as 0.9 and 0.0005 respectively. The input images are horizontally flipped with probability of 0.5 during training. No other data augmentation is used.

In the test phase, we adopt horizontal flip as data augmentation. Simple connected component analysis is applied as the post-processing step to remove noise. We also experimented with more complicated post-processing steps such as 3D denseCRF [16] but only observe a marginal improvement. In the consideration of trade-off between marginal gain and heavy computational cost, we do not use any complicated post-processing techniques in the remaining experiments. It is also worth mentioning that all experimental results presented are generated by a single model without heavy model ensembles.

4.2 Ablation Analysis

We conducted a systematic ablation study using 220 out of the 274 patient cases (220 HGG cases and 54 LGG cases) from the training set for training and the remaining 54 cases for validation.

We present quantitative and qualitative analysis in Table 1 and Fig. 4 respectively. To better demonstrate the effectiveness of the cascaded classifier with erasing module in mitigating the inter-class interference and benefiting the prediction of the difficult class, we additionally report the mean intersection over union (IoU) score over the non-enhancing core category, which is the most difficult class to predict because of its irregularity and dispersibility.

Tumor Region Pooling We first study the effect of the tumor region pooling component in the FSENet. As discussed previously, tumor region pooling helps to ease the class imbalance problem, so that the model can focus on learning useful task-relevant representations for multi-label segmentation without the interference of the dominant non-tumor region. Firstly, we notice that Model 1 without the region pooling component fails to learn and predict almost all pixels as non-tumor category, if normal cross-entropy loss is applied. Instead, we use

[1] http://pytorch.org/

Table 1. Quantitative comparison among baselines and our models

Model	Methods				Dice			PPV			Sens			mean
No.	TRP	MultiS	Context	Erase	W	T	A	W	T	A	W	T	A	IoU
1					0.744	0.686	0.721	0.616	0.577	0.689	**0.982**	**0.910**	**0.805**	22.1
2	✓				0.751 ↑ 9.4%	0.735 ↑ 7.1%	0.688 ↓ 4.6%	0.628 ↑ 1.9%	0.686 ↑ 18.9%	0.774 ↑ 12.3%	0.965 ↓ 1.7%	0.837 ↓ 8.0%	0.659 ↓ 18.1%	24.6 ↑ 11.3%
3	✓	✓			0.890 ↑ 19.6%	0.776 ↑ 13.1%	0.708 ↓ 1.8%	0.897 ↑ 45.6%	0.831 ↑ 44.0%	0.788 ↑ 14.4%	0.892 ↓ 9.2%	0.768 ↓ 15.6%	0.677 ↓ 15.9%	28.1 ↑ 27.1%
4	✓	✓		✓	0.891 ↑ 19.8%	0.775 ↑ 13.0%	0.711 ↓ 1.4%	**0.912** ↑ **48.1%**	**0.843** ↑ **46.1%**	**0.792** ↑ **14.9%**	0.878 ↓ 10.6%	0.763 ↓ 16.2%	0.675 ↓ 16.2%	**28.3** ↑ **28.1%**
5	✓	✓	✓	✓	**0.892** ↑ **19.9%**	**0.782** ↑ **14.0%**	**0.734** ↑ **1.8%**	0.902 ↑ 46.4%	0.817 ↑ 41.6%	0.766 ↑ 11.2%	0.891 ↓ 9.3%	0.790 ↓ 13.2%	0.745 ↓ 7.5%	**28.3** ↑ **28.1%**

The three columns under each metric section correspond to the scores achieved over whole tumor (W), tumor core (T) and active tumor (A) respectively (percentages are the relative changes compared to Model 1). "TRP" indicates whether the tumor region pooling component is used. "MultiS" indicates whether the multi-label segmentation is done in multiple stage. "Context" means that if the context branch is included. "Erase" represents the erasing process.

a weighted cross-entropy loss, where the weighting factor for each class is the normalized inverse frequency of the corresponding class.

We apply the same weighted cross entropy loss to Model 2 for fair comparison. Obviously, even with a weighted cross-entropy loss function, Model 1 still gives unsatisfactory results (second row in Fig. 4)). Despite its high sensitivity score due to excessively predicting pixels as foreground classes, the generated result is undesirable. Model 2 generates more accurate segmentation results, and hence outperforms Model 1 in most of the evaluation categories as shown in Table 1. This shows the effectiveness of the "focus" step.

One-stage *vs.* Multi-stage To simplify multi-label segmentation task, we propose to decompose the one-stage multi-label segmentation problem into several binary segmentations. We expect these more specialized classifiers would perform better in differentiating each class, and hence boost overall results. To examine this, we additionally train a model which feeds the RoI feature map to 4 binary classifiers for individual tumor tissue segmentation (Model 3) as opposed to Model 2 that applies a softmax layer for one-stage multi-label segmentation. We find that Model 3 significantly outperforms Model 2 in most categories of the metrics, which endorses our assumption.

Erasing and Contextual Compensation In Model 3, the inter-class interference problem remains unsolved. Taking this one step further, we introduce the proposed erasing process to Model 4 to study its effectiveness. However, Model 4 only achieves slightly better performance compared to Model 3 which may due to the loss of context information. Therefore, a context branch is added to form Model 5. The additional context information provides reference for the classifiers

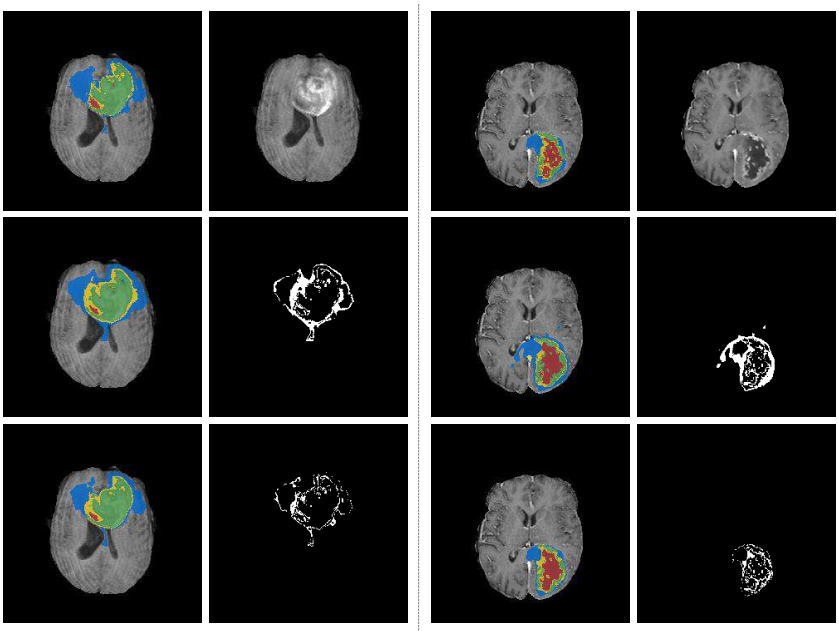

Fig. 4. Qualitative comparison among baselines and FSENet. The ground truths and T1c images of two examples are shown in the first row. The segmentation results generated by the baseline model (Model 1) and their corresponding error images are shown in the second row. The segmentation results produced by FSENet (Model 5) and their corresponding error images are shown in the third row. Color code is the same as that in Fig. 1 (Better viewed in color)

to understand the structure of the tumor. Together with the context branch and the erase process, Model 5 outperforms Model 3 in most evaluation metrics.

The Proposed FSENet The proposed FSENet achieves top performance in terms of the Dice similarity score which is a very important evaluation metric in medical image segmentation, and the prediction of the most difficult non-enhancing core class (Table 1). On top of Res-UNet ($\sim$ 65.5 million parameters), the FSENet only introduces $\sim$1.1 million extra parameters ($\sim$1.7% overhead) to achieve this significant boost in performance compared to the baseline (Model 1). We find that our FSENet can accurately identify and segment each tumor tissue (Fig. 4). The error images in the third row of Fig. 4 indicate that the prediction errors usually occur along the boundary.

A noteworthy advantage of our FSENet is that no hyperparameter is required. On the whole, the proposed pipeline is simple yet effective in dealing with problems of class imbalance and inter-class interference in multi-label brain tumor segmentation.

Table 2. Evaluation results on the test set of BraTS 2015

Network	Dice			PPV			Sens			Rank [2]
	W	T	A	W	T	A	W	T	A	
zhouc1 [17]	0.87	0.75	0.64	0.87	0.81	0.61	0.89	0.75	0.72	1
isenf1 [17]	0.85	0.74	0.64	0.83	0.80	0.63	0.91	0.73	0.72	2
Pereira *et al.*[22]	0.78	0.65	0.75	-	-	-	-	-	-	-
Kamnitsas *et al.*[14]	0.84	0.63	0.63	0.82	0.85	0.64	0.89	0.62	0.66	-
FSENet	0.85	0.72	0.61	0.86	0.83	0.66	0.86	0.68	0.63	3

4.3 Comparison with State-of-the-art Methods

We evaluate the performance of our FSENet by submitting our test set results to the official BraTS 2015 online evaluation platform. The results are reported in Table 2. We also compare our proposed FSENet with several state-of-the-art methods. The two methods "zhouc1" and "isenf1" currently rank 1^{st} and 2^{nd} on the leader-board respectively. However, since BraTS 2015 does not require participants to substantiate their achievements in peer-reviewed publications, we are unable to identify the authors and the details of their methods. The proposed FSENet ranks 3^{rd} on the learder-board. In addition, we also show the performance of two state-of-the-art CNN-based approaches that are evaluated on the same data set. In [22], two patch-based frameworks are trained separately for multi-label segmentation of HGG and LGG case respectively, considering their different characteristics. A multi-scale 3D CNN named DeepMedic proposed by [14] has two convolutional pathways, in order to better utilize multi-scale features for prediction. As shown in Table 2, the proposed FSENet achieve competitive single-model performance.

We also present several examples of the segmentation results generated by our FSENet in Fig. 5, showing the effectiveness of the proposed pipeline.

We are currently unable to report the performance of FSENet on the BraTS 2017 challenge, since access to the dataset is restricted to the challenge participants. A performance analysis based on the BraTS 2017 dataset will be conducted in future when the dataset is publicly available.

5 Conclusion

In this paper, we propose an end-to-end pipeline named FSENet for the challenging multi-label brain tumor segmentation task, which follows the "focus, segment and erase" approach. To address the common class imbalance and inter-class interference problems, two novel components are introduced, which are tumor region pooling and cascaded binary classifiers with erasing. We demonstrate the

[2] The rank is according to the leader-board by the time of paper submission.

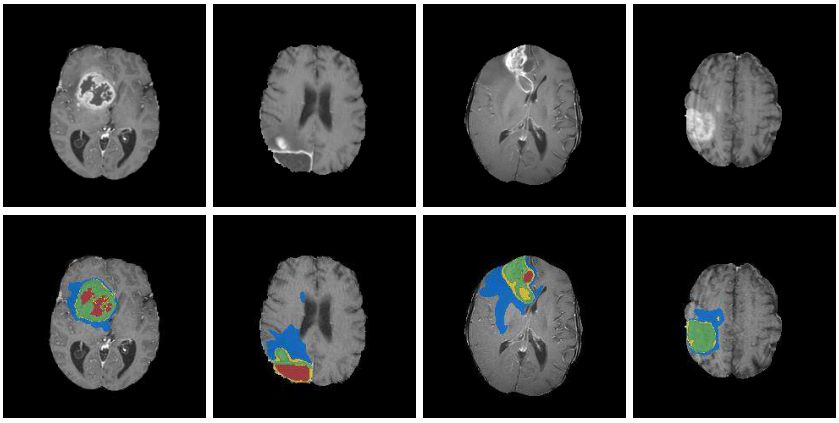

Fig. 5. Examples of the multi-label segmentation results produced by the proposed FSENet. First row shows T1c images from the test set of BraTS 2015. Results generated by the our FSENet are shown in the second row. Color code is the same as that in Fig. 1 (Better viewed in colors)

effectiveness of the tumor region pooling component, and also discuss its advantages compared to other techniques in terms of its flexibility for image-based multi-label segmentation framework and no restriction by the elaborate selection of hyperparamenter. The cascaded binary classifiers with erasing component divides difficult one-stage multi-label segmentation into multiple binary ones for capturing more discriminative task-relevant features. In addition, to suppress the competition and interference from easier to be segmented categories in the prediction of tougher ones, the binary classifiers are cascaded in the "outer-to-inner" manner and possess an erasing processing in between. We show the advantages of the proposed FSENet over the baseline models, demonstrating the effectiveness of the proposed pipeline. Besides, our FSENet achieves 3^{rd} place single-model performance on the BraTS 2015 leader-board without relying on heavy model ensembles or complicated post-processing techniques.

Other applications, like liver tumor segmentation and whole heart segmentation, share similar characteristics and challenges to that of multi-label brain tumor segmentation. We intend to investigate the performance of FSENet to these applications in future.

6 Acknowledgment

We appreciate the support of NVIDIA Corporation with the donation of the Pascal Titan Xp GPU used in this study.

References

1. Bauer, S., Fejes, T., Slotboom, J., Wiest, R., Nolte, L.P., Reyes, M.: Segmentation of brain tumor images based on integrated hierarchical classification and regularization. In: 2012 International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) BraTS Workshop (2012)
2. Beers, A., Chang, K., Brown, J., Sartor, E., Mammen, C., Gerstner, E., Rosen, B., Kalpathy-Cramer, J.: Sequential 3D U-Nets for biologically-informed brain tumor segmentation. arXiv preprint arXiv:1709.02967 (2017)
3. Casamitjana, A., Catà, M., Sánchez, I., Combalia, M., Vilaplana, V.: Cascaded V-Net using ROI masks for brain tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) Brain Lesion Workshop. pp. 381–391. Springer (2017)
4. Chen, X., Nguyen, B.P., Chui, C.K., Ong, S.H.: Automated brain tumor segmentation using kernel dictionary learning and superpixel-level features. In: IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 2547–2552 (2016)
5. Christ, P.F., Elshaer, M.E.A., Ettlinger, F., Tatavarty, S., Bickel, M., Bilic, P., Rempfler, M., Armbruster, M., Hofmann, F., DAnastasi, M., et al.: Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 415–423 (2016)
6. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 424–432. Springer (2016)
7. Cobzas, D., Birkbeck, N., Schmidt, M., Jagersand, M., Murtha, A.: 3D variational brain tumor segmentation using a high dimensional feature set. In: IEEE International Conference on Computer Vision (ICCV). pp. 1–8. IEEE (2007)
8. Colmeiro, R.R., Verrastro, C., Grosges, T.: Multimodal brain tumor segmentation using 3D convolutional networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 226–240. Springer (2017)
9. Feng, X., Meyer, C.: Patch-based 3D U-Net for brain tumor segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2017)
10. Gadermayr, M., Dombrowski, A.K., Klinkhammer, B.M., Boor, P., Merhof, D.: CNN cascades for segmenting whole slide images of the kidney. arXiv preprint arXiv:1708.00251 (2017)
11. Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.M., Larochelle, H.: Brain tumor segmentation with deep neural networks. Medical Image Analysis **35**, 18–31 (2017)
12. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. IEEE International Conference on Computer Vision (ICCV) (2017)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
14. Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical Image Analysis **36**, 61–78 (2017)

15. Kistler, M., Bonaretti, S., Pfahrer, M., Niklaus, R., Büchler, P.: The virtual skeleton database: An open access repository for biomedical research and collaboration. Journal of Medical Internet Research **15**(11), e245 (Nov 2013). https://doi.org/10.2196/jmir.2930, http://www.jmir.org/2013/11/e245/
16. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: Advances in Neural Information Processing Systems. pp. 109–117 (2011)
17. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: BraTS 2015 online evaluation platform. https://www.virtualskeleton.ch/BRATS/Start2015
18. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (2015)
19. Menze, B.H., Van Leemput, K., Lashkari, D., Weber, M.A., Ayache, N., Golland, P.: A generative model for brain tumor segmentation in multi-modal images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 151–159. Springer (2010)
20. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: Fourth International Conference on 3D Vision. pp. 565–571. IEEE (2016)
21. Nyúl, L.G., Udupa, J.K., Zhang, X.: New variants of a method of MRI scale standardization. IEEE Transactions on Medical Imaging **19**(2), 143–150 (2000)
22. Pereira, S., Pinto, A., Alves, V., Silva, C.A.: Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Transactions on Medical Imaging **35**(5), 1240–1251 (2016)
23. Prastawa, M., Bullitt, E., Ho, S., Gerig, G.: A brain tumor segmentation framework based on outlier detection. Medical Image Analysis **8**(3), 275–283 (2004)
24. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2015), http://arxiv.org/abs/1505.04597
25. Wang, G., Li, W., Ourselin, S., Vercauteren, T.: Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. arXiv preprint arXiv:1709.00382 (2017)
26. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)