

**UCSF**

**UC San Francisco Electronic Theses and Dissertations**

**Title**

Improving the generalizability of convolutional neural networks for brain tumor segmentation in the post-treatment setting

**Permalink**

<https://escholarship.org/uc/item/8b36c5gs>

**Author**

Ellison, Jacob Charles

**Publication Date**

2020

Peer reviewed|Thesis/dissertation

Improving the generalizability of convolutional neural networks for brain tumor segmentation in the post-treatment setting

by  
Jacob Ellison

THESIS

Submitted in partial satisfaction of the requirements for degree of  
MASTER OF SCIENCE


in

Biomedical Imaging

in the

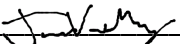
GRADUATE DIVISION  
of the  
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:  
  
F1E4E52A4E3D4D8... Janine Lupo  
Chair

DocuSigned by:  
  
Valentina Padoia

DocuSigned by:  
  
Yan Li

DocuSigned by:  
  
DEA76591766D41E... Javier Villanueva-Meyer

Committee Members



# **Improving the generalizability of convolutional neural networks for brain tumor segmentation in the post-treatment setting**

Jacob Ellison

## **Abstract**

Current encoder-decoder convolutional neural networks (CNN) used in automated glioma lesion segmentation and volumetric measurements perform well on newly diagnosed lesions that have not received any treatment. However, there are challenges in generalizability for patients after treatment, including at the time of suspected recurrence. This results in decreased translation to clinical use in the post-treatment setting where it is needed the most. A potential reason is that these deep learning models are primarily trained on a singular curated dataset and demonstrate decreased performance when they are tested in situations with unseen variations to disease states, scanning protocols or equipment, and operators. While using a highly curated dataset does have the benefit of standardizing comparison of models, it comes with some significant drawbacks to generalizability. The primary source of images used to train current models for glioma segmentation is the BraTS (Multimodal Brain Tumor Image Segmentation Benchmark) dataset. The image domain of the BraTS dataset is large, including high- and low-grade tumors, varying acquisition resolution, and scans from multi-center studies. Despite this, it may still lack sufficient feature representation in the target clinical imaging domain. Here we address generalizability to the disease state of post-treatment glioma. The current BraTS dataset consists entirely of images obtained from newly diagnosed patients who have not undergone surgical resection, received adjuvant treatment, or shown significant disease progression, all of which can greatly alter the characteristics of these lesions. To improve the clinical utility of deep learning models for glioma segmentation, they must accommodate variations in signal intensity that may arise as a result of resection, tissue damage (treatment induced or otherwise), or progression. We compared models trained on either BraTS data, UCSF acquired post-treatment glioma data, UCSF acquired newly diagnosed glioma data, and various combinations of these data, to determine the effect of including images with features unique

to treated gliomas into training the networks on segmentation performance in the post-treatment domain. Although an absolute threshold training inclusion value for generalization of segmentation networks to post-treatment glioma patients has not been established, we found that with 200 total training volumes, models trained with greater than or equal to 30% of the training images from patients with prior treatment received the greatest performance gains when testing in this domain. Additionally, we found that after this threshold is met, additional images from newly diagnosed patients did not negatively impact segmentation performance on patients with treated gliomas. We also developed a pre-processing pipeline and implemented a loss penalty term that incorporates cavity distance relationships to the tumor into weighting a cross entropy loss term. The aim of this was to bias the network weights to morphological features of the image relevant to pathologies that are prevalent post-treatment. This may either be used as an initialization for training with an available larger dataset such as BraTS or used to finetune a transferred network that has not seen sufficient post-treatment glioma images during training in order to allow domain adaptation with fewer training data from this disease state. Preliminary results show qualitatively more desirable segmentations of tumor lesions with respect to cavities and small disconnected components in selected examples that are worthy of further analysis with alternate training configurations, more focused performance assessments, and larger cohorts. Here, we will evaluate these techniques as potential solutions to improve the generalizability of CNN tumor segmentation to post-treatment glioma, as well as provide a framework for further data augmentation based on augmenting the boundary of these lesions.

## Acknowledgements

*I appreciate the dedication, commitment, guidance, and aid of the following groups and individuals, without them I would not have been able to complete this Master's thesis. Thank you.*

UCSF Master of Science in Biomedical Imaging Program:

Students and Faculty

UCSF Lupo Lab:

Mentor - Dr. Janine Lupo, PhD

Focus Group for Machine Learning in Brain Imaging

Defense Committee:

Dr. Javier Villaneuva-Meyer, MD

Dr. Yan Li, PhD

Dr. Valentina Padoia, PhD

UCSF Center for Intelligent Imaging

Nvidia

HDFC Cancer Center and Cancer Imaging Resources

## Table of Contents

1 Introduction .....	1 - 5
2 Methods .....	6 - 12
2.1 Datasets .....	6, 7
2.2 Preprocessing and Augmentation .....	7, 8
2.3 Network Architectures and Hyperparameters .....	8 - 10
2.4 Experimental Set up .....	11, 12
3 Results .....	12 - 15
5 Discussion .....	15 - 18
4 Conclusions .....	18, 19
6 Figures and Tables .....	20 - 28
7 References .....	29 – 32

**List of Figures**

Figure 1 ..... 20

Figure 2 ..... 20

Figure 3 ..... 21

Figure 4 ..... 21

Figure 5 ..... 22

Figure 6 ..... 23

Figure 7 ..... 23

Figure 8 ..... 24

Figure 9 ..... 24

Figure 10 ..... 25

Figure 11 ..... 25

Figure 12 ..... 26

Figure 13 ..... 26



## List of Tables

Table 1 .....	27
Table 2 .....	27
Table 3 .....	28
Table 4 .....	28

## **Introduction:**

Central nervous system malignancies originating in glial cells are the most common form of brain cancer among adults<sup>[1]</sup>. Primary gliomas can cause devastating changes to brain function, quality of life, and lifespan. In order to guide diagnosis and treatment, they are separated into grades and subtypes according to the World Health Organization (WHO) Classification of Central Nervous System Tumors<sup>[2]</sup>. Low - grade tumors include gliomas of Grade I and Grade II, including astrocytoma or oligoastrocytomas. On average, Low-grade gliomas tend to have more well-defined boundaries, are slower growing, and have better long-term prognosis. Gliomas of Grade III and IV are grouped into the High - grade category. These include anaplastic astrocytoma and glioblastoma multiforme tumors. In general, High-grade gliomas may consist of undifferentiated cells, are rapidly progressing, highly malignant, and have worse long term outcomes, specifically for certain molecular subtypes such as Isocitrate dehydrogenase (IDH) wild type glioblastoma (GBM) <sup>[3]</sup>.

The global mortality rates for primary malignant brain tumors in women are 2.0/100,000, and 2.8/100,000 in men<sup>[4]</sup>. Although gliomas are relatively uncommon, survival can be as short as four months in the most aggressive form of GBM<sup>[1]</sup>. However, the prognosis of gliomas vary greatly, with five year median survival rates ranging from 6% to 73%<sup>[4,5]</sup>. This variability can be accounted for by patient age, grade, and tumor subtype. Histologic type and grade, age, extent of resection, tumor location, molecular and genetic characteristics have all been well correlated to outcomes and treatment response in both clinical trial and population registry data<sup>[1,6]</sup>. In addition to the molecular and histological analyses, volumetric spatial measurements from MRI are a widely used clinical standard for assessing tumor progression and treatment response<sup>[6,7]</sup>.

These anatomical measurements are made in accordance with the Response Assessment in Neuro-Oncology (RANO) criteria<sup>[3]</sup> Assessments are made by expert Neuro-oncologists and Neuro-radiologists largely based on standard clinical imaging protocols that include acquiring T1-weighted pre- and post-contrast injection of a gadolinium-based agent, T2-weighted, and T2 Fluid Attenuated Inversion Recovery (FLAIR) images<sup>[7]</sup>. The 2D product of maximum bidimensional diameters of contrast-

enhancing tumor and qualitative evaluation of T2 FLAIR hyperintensities, are the two main basis for RANO criteria<sup>[2,3]</sup>. Using these images to make manual estimates of tumor volume and RANO criteria are highly time consuming and subject to interobserver variability<sup>[6]</sup>. Automated approaches for tumor volume segmentation have been shown to save time, minimize inter-rater variability, correlate with patient outcome, and can be sensitive to more subtle changes over time<sup>[6,7]</sup>, with potential to provide a high level of utility to the clinic.

There are technical challenges associated with automated brain tumor lesion segmentation. Many arise because the lesions are differentiated by signal intensity changes relative to healthy tissue, and gradients between neighboring structures can be smoothed or obscured by partial-volume or bias field artifacts<sup>[8]</sup>. These challenges can cause even manual delineations to vary with respect to inter-observing experts<sup>[7]</sup>. Another challenge to the automated tumor segmentation task is that priors used to account for location or shape cannot be used due to spatial variations in location or structural changes that arise from mass effect displacement of healthy brain structure, or the high variation in tumor location that is not present in segmentation tasks for other organs<sup>[8]</sup>.

Despite these challenges, deep learning neural networks have been used in order to provide highly accurate and speedy segmentations for newly-diagnosed, treatment naïve gliomas <sup>[1,6,7,9-13]</sup>. The most prevalent methods in computer vision literature for automated tumor segmentation use encoder decoder architecture in an end-to-end approach to generate lesion masks<sup>[10]</sup>. These methods are largely inspired by the pioneering developments of fully convolution encoder-decoder architectures capable of segmentation in works like U-net<sup>[11]</sup> and V-net<sup>[12]</sup>. Many of the prevailing neural networks for brain tumor lesion segmentation have been trained using the Multimodal Brain Tumor Image Segmentation Benchmark (BraTS) data set<sup>[8]</sup>. This dataset has in the past consist primarily of, and currently only consists of anatomical images from newly diagnosed glioma patient scans before surgery. Lesions from newly diagnosed gliomas tend to have better boundary signal delineation from healthy tissue, lack the presence of cavities left by tumor resection and tissue deformities that result from adjuvant treatment, or disease progression. These variations in signal intensities are exacerbated in patients who have had

multiple follow ups with various treatments and are not represented in the BraTS dataset. This underrepresentation of unique features causes challenges of generalizability and translation to clinical use for patients with glioma post-treatment. Training convolutional neural networks for glioma segmentation has been done with a variety of architectures and training strategies. The model that won the BraTS 2018 brain tumor segmentation challenge followed the encoder-decoder structure, with an asymmetrically large encoder to extract deep image features, and decoder to reconstruct dense segmentation masks. They also used a variational autoencoder (VAE) branch to the network to reconstruct the input images jointly with segmentation in order to regularize the shared encoder<sup>[13]</sup>. Here this network will serve as a starting point to evaluate our methods of improving generalizability.

After treatment, signal intensity gradients near the edges of T2 hyperintense lesions can be weak. In order to provide networks with sufficient training examples to learn these precise features, data augmentation may be employed. Passing expert annotated lesion masks through a distance-based boundary selection threshold, morphological dilation, Laplacian kernel convolutions and addition to the original image, similar to high pass filter addition of the lesion boundary in the Fourier domain, can generate synthetic images with lesion boundaries enhanced. Similarly, synthetic images with tumor boundaries blurred, i.e. contrast between surrounding healthy tissue and lesion lowered, can be produced using a similar pipeline with a Gaussian blurring step replacing the Laplacian convolution step. Augmenting data in this way may provide a larger representation of images where tumor boundary signal gradients vary in training. This may aid the network generalize to these challenging post-treatment cases during testing.

Some recent works have explored the use of distance based or other types of anatomical feature weighting to learning in the form of loss function modification that can provide additional performance gains for segmentation<sup>[20,21]</sup>. One approach that addresses issues of low boundary signal gradient in tumor edges is the boundary aware method proposed by Hatamaizadeh et al. Their network is designed to account for organ boundary information, both by providing a special network edge branch and edge-aware loss terms in addition to a shape stream encoding branch that processes feature maps at the

boundary level [21]. They included a modification to their loss function that utilized a weighted loss term based on the predicted lesion edge and ground truth lesion edge. Using this, their network accounted for edge information and improved segmentation performance in an end-to-end capable training method[21]. Another approach to improving performance of glioma segmentation proposed an automated data augmentation method that uses generative adversarial networks (GANs) to learn a Coarse-to-fine Boundary-aware generation of synthetic images[22]. Their model similarly incorporates the use of boundary aware loss into training. They were able to generate additional training data for the segmentation task, improving dice scores by 36% in validation compared to networks that were trained on synthetic images generated by traditional GANs.

To address the challenge of generalizing to data with a larger representation of tissue deformities and resection cavities, we employed a loss term that incorporates information of unique pathological features to images of treated glioma in order to bias the network to this domain. A cavity-aware loss term may provide additional performance gains with respect to segmentation of gliomas after treatment. Similar to the method developed by Caliva et. al., distance mapping could be used to penalize the network by weighting distance relationships between cavity and tumor boundaries in a variety of ways. Distances of the lesion to the cavity may be penalized either inversely, or both proportional and inversely where pixels of the lesion that are farthest away and closest to the cavity will be weighted most highly. These weightings can be used in a cross entropy loss term to penalize the network more for incorrect predictions at these locations.

Other methods have addressed the consistency of network performance on glioma segmentation[23,24]. One, by Wang et al., addresses the challenge of predicting the likelihood of success for a given segmentation by developing an uncertainty estimation of a generated tumor lesion. They do this by utilizing cascaded networks in test-time augmentations to provide multiple prediction results of the same input image with different spatial transformations and intensity changes. The disagreement between these predictions provides an uncertainty estimation of the segmentation[24]. Including this

uncertainty information can provide guidance for manual correction or automated correction in places where the network was uncertain about its segmentation.

A common problem in biomedical imaging tasks is that when a predictive model approximating  $f_s(x)=P(y|x)$ , is trained to make predictions in a particular imaging source domain,  $D_s=\{X_s,P(X_s)\}$ , with feature space  $X_s$ , and marginal distributions  $P(X_s)$ , with ground truth labels  $(Y_s,X_s)$ ,  $y \in Y_s$  and  $x \in X_s$ , from source images,  $x_{si}$ , of the feature space  $X_s=\{x_{s1},x_{s2},\dots,x_{sn}\}$ , that model can experience a decrease in performance when applied to images of a new target domain,  $D_t=\{X_t,P(X_t)\}$ , where the feature space and marginal distributions,  $X_t \neq X_s$  and  $P(X_t) \neq P(X_s)$ , differ from the training source<sup>[14]</sup>. This occurs when the data used to train a network, the source domain, is not fully representative of the target imaging domain in which the model is desired to be used. The problem of adapting a predictive network to approximate an ideal model that either works well in both domains, or the target domain alone,  $f_{s,t}(x)=P(y|x)$  or  $f_t(x)=P(y|x)$ , also known as, domain adaptation, has been addressed before in both supervised and unsupervised methods<sup>[14-19]</sup>.

Here we are focused on the differences in the anatomical imaging feature differences of the disease states between the BraTS dataset, a more uniform dataset of newly diagnosed gliomas, and post-treatment glioma. There are three methods that we address here, of training a network such that it can perform well in the target domain of post-treatment glioma. The goal is to improve the generalizability of CNNs to post-treatment cases of glioma by: 1) incorporating sufficient training examples with features of the target domain i.e. training on a dataset reflective of a mixture of domains  $D_{s,t}=\{X_{s,t},P(X_{s,t})\}$ , s.t.  $X_{s,t}=\{x_{s1},x_{s2},\dots,x_{sn},x_{t1},x_{t2},\dots,x_{tm}\}$  into current models, 2) exploring the use of a modified loss function ( $L_{tot}=L(y,x)+L(y_t,x_t)$ ) that incorporates a penalization term  $L(y_t,x_t)$  based on the distance relationship between surgical cavities and the leading edge of the tumor into training, and 3) developing a tumor boundary augmentation technique to increase the training examples with greater variability of features relevant to the target domain.

## **Methods:**

### **-Datasets-**

#### *BraTS:*

This dataset is known as the Multimodal Brain Tumor Image Segmentation Benchmark (BraTS) data set<sup>[8]</sup>. The 2018 version of this dataset that we used includes a total of 285 pre-operative, multi-contrast MR scans collected from glioma patients annotated by expert neuro-radiologists. Multi-contrasts include T1, T1-contrast enhanced, T2, and T2 FLAIR images. Whole tumor (based on the T2-hyperintensity lesion), tumor core (or necrosis), and contrast-enhancing tumor lesions, have been provided as annotation labels for each of the volumes. 75 of the multi-contrast images were acquired from LGG and 210 from HGG. Images were already co-registered to the same anatomical template, interpolated to the same resolution (1 mm<sup>3</sup>) and skull stripped.

#### *UCSF:*

The UCSF data set consisted of 208 newly diagnosed patients with glioma and 217 patients with treated gliomas at the time of suspected recurrence. These datasets were all acquired on a 3T GE scanner at UCSF within 48 hours before a patient underwent surgical resection. The post-treatment patients all underwent prior surgical resection, radiation therapy, and chemotherapy and have lesions that contain a mixture of recurrent tumor and treatment induced injury. All patients provided informed consent for their images to be used in research.

Image volumes consisted of T2, T2 FLAIR, T1, and T1 post-contrast images. Expert neuro-radiologists guided the annotation of T2-hyperintensity, necrotic, and contrast enhancing lesions. Only the T2all lesions were used to compare the models. All volumes were cropped to [240,240,155] spatial dimensions in pre-processing. All scans were reformatted to an axial orientation.

In the post-treatment dataset, only initial scans with unique patient IDs were selected to prevent data leakage between training datasets and testing. 25 of these treated datasets also contain an annotated cavity region that was used for generating distance maps in the modified loss function during training.

Skull stripping was performed, and brain masks were generated in an additional image using FSL brain extraction toolkit<sup>[28]</sup>. Skull stripped masks were only applied to training data when models were compared against BraTS.

#### -Preprocessing and Augmentation-

##### *Part 1: Inclusion of post-treatment data*

Augmentation and preprocessing were not changed from the methods reported in the paper that described the BraTS18 winning model<sup>[13]</sup>. Images acquired at UCSF were cropped to [240,240,155]. These images were also skull stripped to match the BraTS data in the multi-channel experiments. In the single channel experiments using UCSF acquired, newly diagnosed and post-treatment glioma, the images were left without skull stripping to determine if the same performance gains can be seen without this preprocessing step that may not be available in routine clinical use.

##### *Part 2: Loss Function Modification*

Within the Dioscoridess framework, cropping to [224,224,116] with random center was not used. Instead of zero mean normalization to range [-1,1], images were normalized to the range [0,1] and divided by the 85th percentile to account for outliers. Cavity distance maps were generated prior to training and used as an additional label input into training in order to avoid extraneous computation during training and speed up training times. Two variations of this technique were used. First, the binary mask of the annotated cavity lesion is subtracted from 1 and the absolute value is taken. This is followed by a Euclidean distance transform to generate a distance map where pixels far away from the cavity have the highest scalar value. This map was then normalized to values between 0 and 1 and multiplied by the binary T2 lesion mask. 1 is added to mitigate any zero weighting. This generates a whole tumor lesion where pixels close to the cavity are weighted by low values and pixels far away from the cavity are weighted by higher values as shown in **Figure 2**. When this distance map is used in the calculation of the loss, it is referred to as inverse distance weighted.



The second technique uses this T2 distance weighted lesion and subtracts the values from 1, taken as the absolute value, and multiplies them by the original binary T2 lesion mask. This has the effect of generating a distance map where pixels of the lesion that are close to the cavity are weighted highly and pixels far away from the cavity are weighted less. This map is then normalized to the range between 0 and 1, subtracted from the other distance map, and taken as the absolute value. 1 is added to mitigate any zero weighting. This generates a T2 lesion where pixels both close to and far from the cavity are weighted most highly, and pixels in the center are weighted less, as shown in **Figure 3**. When this distance map is used in the calculation of the loss, it is referred to as combination distance weighted.

### *Part 3: Boundary Augmentation*

A preprocessing pipeline for boundary augmentation was developed but not incorporated into training at this stage. This pipeline includes using the expert annotated binary tumor lesion masks and performing a Euclidean distance transform. A boundary can be selected by thresholding values near the edges of the lesion. Then a morphological dilation followed by a Laplacian kernel convolution is used to generate a border lesion image where edges are sharpened. These pixels are then added back to the original image to generate synthetic images with lesion boundaries enhanced. Similarly, synthetic images with tumor boundaries blurred, i.e. contrast between surrounding healthy tissue and lesion lowered, can be produced using a similar pipeline with a Gaussian blurring step replacing the Laplacian of Gaussian convolution step. These edge blurring or sharpening augmentations can be performed as a preprocessing step, with randomly selected Gaussian parameters or edge thresholding, to create a large variation of boundary signal intensity gradients. A schematic outlining this method is shown in **Figure 4**.

-Network Architectures and Hyperparameters-

### *Part 1: Inclusion of post-treatment data*

The network was adapted from the 2018 BraTS challenge winner and pulled from the Nvidia GPU Cloud catalog. Training was performed using Nvidia's Clara v2 framework. The original network is a modified V-net with a variational branch that encodes the original input image and uses an L2

regularization loss penalty to enforce regularity in the output segmentation masks. This variational auto encoder component was not used due to the fact that similar initial testing accuracies were being achieved with and without, and due to the concern that additional parameters would slow training times or limit available GPU compatibility. The encoding blocks follow the structure of a Seg-Res-Net<sup>[27]</sup>, with 4 encoding blocks, each containing two (3,3,3) convolutions with stride 2, Group normalization, and ReLU activation. 16 initial convolution filters were used, with an increase in feature dimension by a factor of two through each convolutional layer. The decoding path included similar 4 decoding blocks to the encoder, where skip connections are used at each upsampling step. The upconvolutions reduce the number of features by a factor of 2 with kernels of (1,1,1) and doubling the spatial dimension by using 3D bilinear upsampling. The network was adapted to use 16 initial filters instead of 32 to fit into available GPU capacity. The network was used in the original configuration to take input of multi-contrast image data. Similarly, the output included the whole tumor, enhancing tumor, and tumor core. Dropout probability used was 0.2. Regularization weight decay was 1e-5. A schematic of the original network architecture developed by Nvidia that won the BraTS18 challenge <sup>[13]</sup> and modified to fit the above parameters, is shown in **Figure 1**.

The main differences between our network and the BraTS18 winner were that we used automatic mixed precision, removed the auto-encoding branch, and decreased the initial filters from 32 to 16. Ensembling the networks was additionally shown to increase segmentation performance; however, this technique was not used. This configuration may not be the optimal architecture for achieving high performance, however the goal of these experiments was to generate reproducible training while varying the training data and evaluating segmentation performance changes in the post-treatment domain, with the aim of generating insights about performance gains that can be applied to more optimized model architecture and hyperparameters to increase performance.

The second set of experiments modified the network described above to generate an output of just one segmentation mask, the T2all lesion (denoted whole tumor in the BraTS dataset), based on only one input, the T2 FLAIR images.

Networks were trained with an Adam optimizer with Dice loss. Models trained for 300 epochs, with one network stopping early (266 epochs) due to connectivity issues. A batch size of 1 was used to allow the multichannel data to fit into the GPU's memory capacity and kept consistent for the single channel experiments. For the multi-channel experiments, a learning rate of  $1e-4$  was used, while the single-channel experiments employed a learning rate of  $3e-4$ . These networks were trained on Tesla V100-32GB GPU within the UCSF Radiology scientific computing servers, using Nvidia's Clara-v2 training framework.

### *Part 2: Loss Function Modification*

Dioscorides framework was used to train models using various loss functions with Adam optimizer and learning rates of  $5e-5$  and  $1e-4$ . This framework was developed by the Podoia laboratory with the Center for Intelligent Imaging (Ci<sup>2</sup>) at UCSF, is similarly written in TensorFlow, and was used due to its flexibility to accommodate additional image inputs with customizable loss functions. The architecture for these experiments used a smaller V-net with 4 levels, instead of 7, compared to the models used in training with the Nvidia Clara framework. These layers used convolutions per level of 1,2,4,1, with a dropout probability of .05. The architecture used here was not the most optimal configuration for achieving high performance but was selected to generate reproducible and deterministic training for comparing performance among various loss functions.

Once the distance maps were generated as described above, a cross entropy term was added to the Dice loss term, to formulate the overall loss. This method is largely based on the distance-based penalty term developed by Caliva et. al<sup>[20]</sup>. This cross entropy is computed between the pixels of the predicted lesion and the ground truth, where pixels of the ground truth are weighted by their distance relationship to the surgical cavities. The formulations of these Loss terms are as follows:  $L_{tot}=\lambda_1L_{Dice}+\lambda_2L_{cavity}$ , where  $L_{cavity}=w(y_i)*-\sum_i(y_i)\log(y_i')$ , s.t,  $w(y_i) = 1 + \text{dist}_{cavity}(y_i)$ , where  $y_i$  is the ground truth lesion, and  $y_i'$  is the predicted lesion, and  $\text{dist}_{cavity}(y_i)$  is some function of distance between the cavity and the tumor. This cross entropy term was weighted at  $\lambda_2=0.1$ , and Dice at  $\lambda_1=1$ . This was compared against different learning

rates to identical models using a Dice loss term alone. These networks were trained using a Titan X-16GB GPU within the UCSF Radiology scientific computing server.

-Experimental Set up-

*Part 1: Inclusion of post-treatment data*

In the multi-channel experiments, 3 identical networks were trained to compare the effect of including post-treatment glioma into training on BraTS data. Network 1 had a Train/Val split of 227/57 images from the BraTS data. Network 2 was trained on 61/16 images from post-treatment data. Network 3 was trained on 288/73 total images from both UCSF acquired post-treatment glioma and the BraTS dataset. All images used for training and inference were skull-stripped and cropped to [240,240,155] to match the format of the BraTS dataset. All three models were then tested on a set of 25 independent patients with post-treatment glioma separated from the original dataset prior to training and chosen at random from the post-treatment dataset. The testing results are shown in **Table 1**.

Next, five models were trained using only T2 FLAIR image contrast, to generate only the T2 lesion. These models used data input that had not been skull stripped to avoid a non-trivial pre-processing step that may not be available in clinical use. These models additionally used equal total amounts of training data, something that may have confounded the above experiments. These experiments were focused as well, on a domain comparison between newly diagnosed and post-treatment glioma acquired at UCSF. This provided an additional domain comparison to comparing post-treatment patients to BraTS, in that little is changed between groups of UCSF data aside from the disease state. These experiments may then serve as a more directed domain comparison between segmentation of newly diagnosed and post-treatment gliomas. Models were trained with differing ratios of post-treatment and newly diagnosed glioma patients from scans acquired at UCSF. 200 total patients were used to train these models with varying ratios of post-treatment and newly diagnosed glioma, all with training validation splits of 80%/20%. Ratios of patient disease state inclusion in training, with testing dice score on 19 patients with post-treatment glioma are shown in **Table 2**.

Finally, four additional networks were trained starting with 153 post-treatment glioma patients, with 50 newly diagnosed glioma patients added into the training set for each successive model. These models were validated on 39 post-treatment patients and tested on a set of 19 separate patients with post-treatment glioma. The aim of these experiments was to determine if additional images of newly diagnosed glioma would negatively impact segmentation performance on the post-treatment glioma patients. Ratios of patient disease state inclusion in training, with testing Dice scores on 19 patients with post-treatment glioma are shown in **Table 3**. All networks were trained on a Tesla V100-32GB GPU within the UCSF Radiology scientific computing server.

#### *Part 2: Loss Function Modification*

For the loss modification experiments, a small subset of the post-treatment data with expert cavity lesion annotations was used to train. 25 patients were used to train, 25 separate patients were used to validate, and 25 patients were used to test. All patients had post-treatment glioma. To determine if the distance mappings are affecting convergence at any stage in training, different learning rates were used. The models were trained with learning rates of  $1e-5$  and  $5e-5$  and compared to models trained with Dice loss alone. Models were selected by their top validation accuracy before 8000 iterations for testing, then resulting predictions were compared.

### **Results:**

#### *Part 1: Inclusion of post-treatment data*

In the first set of experiments, models trained on BraTS data, post-treatment glioma, and a mix of both, were compared. The BraTS trained model performed the lowest with a mean testing Dice score of 0.721 and standard deviation of 0.216. The model trained on post-treatment data with about  $\frac{1}{4}$  of the training data compared to the BraTS model, had a mean Dice score of 0.810 and a standard deviation of 0.107. The model trained on a mix of both performed the highest with a mean Dice score of 0.830 and a standard deviation of 0.102. Since the UCSF data had necrotic lesion and contrast enhancing lesion annotated differently from the BraTS tumor core and enhancing tumor, we were unable to change the

Clara framework calculation of those lesions correctly within the timeframe. Because of this, these experiments are focused on the singular T2 lesion predictions. Upon observation of the predicted T2 (Whole Tumor) lesion masks, it appears as though the BraTS model falsely segments hyperintense pixels near sulci, as well as pixels within distorted ventricles and large cavities near the lesion, some examples are shown in **Figures 9 and 10**.

In the second set of experiments, models were trained using only T2 FLAIR images as input to predict the whole tumor lesion. 200 patients were used to train these models with varying ratios of post-treatment and newly diagnosed glioma, each with Train/Val splits of 80%/20%. A trend emerged that illustrated higher testing performance for models where a larger proportion of post-treatment data were mixed into training. Dice scores of the models and their ratio of post-treatment to newly diagnosed glioma are shown in **Table 2**. The distribution of the Dice scores in the testing set was also shown to be tighter following this trend. This may be indicative that the features present in cases that were causing the model to perform very poorly, are being learned through training inclusion. However, at about 30-35% inclusion, performance gains plateau. This effect is illustrated in **Figure 11 (left)**. Examples from testing on 19 patients with post-treatment glioma are shown in **Figure 12 and 13** that illustrate increased performance for models trained on post-treatment data.

Finally, training was done starting with images from 153 post-treatment glioma patients and increasing the amount of newly diagnosed patients added into training by 50 in each successive run until 200 newly diagnosed patients were added into training. Each model was validated on 39 patients with post-treatment glioma and tested on a separate set of 19 patients with post-treatment glioma. Here, the testing Dice scores remain steady around .82, indicating that addition of training data from newly diagnosed glioma did not negatively impact the segmentation performance of post-treatment glioma. The mean testing Dice score for the model that stopped at 266/300 epochs was slightly lower than the rest at .79. The mean testing Dice scores for these experiments are shown in **Table 3** and visualized in **Figure 11 (right)**.

### *Part 2: Loss Function Modification*

Models trained and validated on 25/25 images from the post-treatment glioma domain with standard Dice loss, inverse distance weighted cross entropy loss plus Dice, and combination distance weighted cross entropy plus Dice, where trained with the same architecture at different learning rates and tested on 25 patients with post-treatment glioma. The models trained with a distance weighted cross entropy term were weighted with  $L_{\text{tot}} = \lambda_1 L_{\text{Dice}} + \lambda_2 L_{\text{cavity}}$ ,  $\lambda_1 = 1.0$  and  $\lambda_2 = 0.1$ . The mean testing Dice scores and standard deviations for the highest validating models at 8000 iterations, are shown in **Table 3**. All testing Dice scores were significantly lower than the state-of-the-art models, with an average Dice score of 0.6572. The model trained with combination distance weighted cross entropy plus Dice outperformed the model trained with standard Dice loss with a learning rate of  $1e-4$  by 0.0304. The highest performing model in testing was the model trained with a standard Dice loss at a learning rate of  $5e-5$ . Although this model had a higher testing Dice score than the best performing model trained with the distance weighted cross entropy term by .055, upon examination of the lesion segmentations, it appears that in situations with irregular tumor lesion shapes with respect to post-treatment cavities, and lesions containing small disconnected components, the models trained with distance weightings were able to segment these portions of the tumor more favorably than their standard Dice trained counterparts. Some examples are shown in **Figures 6, 7, and 8**.

### *Part 3: Boundary Augmentation*

Synthetic images were generated using the boundary lesion sharpening and smoothing techniques described above. The T2 lesion was dilated and eroded by disk shaped morphological structuring elements of sizes 5 and 3 and the original T2 lesion was subtracted. This generated a mask where pixels of surrounding tissue to the lesion were selected. The log intensities of the pixels selected by this mask were compared to the log intensities of the pixels selected by the original T2 lesion. These were plotted in histograms and compared for the non-augmented image and synthetic images resulting from both the edge smoothing and edge sharpening operations. The histograms, along with pixel selections are shown in **Figure 5**. The degree of overlap in intensities between the pixels surrounding the lesion and the lesion

itself serves as a measure of contrast at the boundary of the lesion. Overlap between lesion and boundary tissue was shown to decrease following the boundary enhancing operation, and increase following the boundary blurring operation.

## **Discussion:**

### *Part 1: Inclusion of post-treatment data*

Our results indicate that including a threshold of training examples from patients with post-treatment glioma will ultimately improve segmentation performance in the target post-treatment domain. We saw that with 200 training volumes, greater than 30% inclusion into training from post-treatment glioma showed the highest performance gains, with plateaus at increasing levels of inclusion. This increase in performance is shown in **Figure 11 (left)**. At this amount of training examples, 30% constitutes 60 volumes from post-treatment glioma in order to generalize. An absolute minimum threshold is unclear; however, this may provide useful information when training a network that is desired to perform segmentation on brain tumors with specific disease states that may be prevalent in the clinical setting. We also noted that once this threshold inclusion is met, and the network is performing well on post-treatment glioma, addition of images from newly diagnosed glioma did not negatively impact segmentation performance on post-treatment glioma. This effect is illustrated in **Figure 11 (right)**.

Although we would like to compare segmentation accuracies of all three lesions that current models are capable of segmenting, these experiments have demonstrated that including a certain threshold of examples from the target clinical disease state of post-treatment glioma, into training may improve generalizability to post-treatment patients. Additionally, optimal architecture and hyperparameters for these experiments were not used. These implications of these results however may provide an insight to achieve performance gains when training a segmentation model for clinical use on post-treatment glioma patients and may be applied to higher performing network architectures and training configurations. These experiments should be repeated using more optimized architecture and hyperparameters in an aim to improve the segmentation performance of post-treatment glioma to the current standard of BraTS



testing accuracies, with some models achieving Dice scores of greater than .90 for the whole tumor lesion.

When comparing the models trained on BraTS, newly diagnosed glioma, and post-treatment glioma, we only evaluated generalizability in a task-based assessment by evaluating testing Dice scores on post-treatment glioma. An alternative method that may provide another approach to quantifying domain adaptation to the post-treatment disease state would be to explore the use of a discriminator network. Discriminators have been used in the past to adapt a network to a given domain by adversarial training against a discriminator that classifies the features of an image passed through the encoding branch of a segmentation network<sup>[14]</sup>. If the classification has a high accuracy, it indicates the features of a given image that are being extracted belong to a distinct domain. In this way, the network will have learned to extract features that are specific to a certain domain. Penalizing the network for correct classifications during training had the effect of allowing the network to extract features that are not distinct to either domain, thus illustrating generalizability to both the target and source domains <sup>[14]</sup>. In an alternate use of this method, a discriminator could be used to classify the features of an image passed through the encoding branch of the of our network. If the classification accuracy is low, this may be indicative that the network is learning to extract features that are relevant to segmentation in both domains, and thus has generalized.

#### *Part 2: Loss function modification*

The results of the loss experiments are preliminary. The models trained here performed with overall testing scores significantly lower than the state-of-the-art models. Despite this fact, in select situations, models trained with a loss term that incorporates the distance relationship between the post-treatment cavities and T2all lesion, were showing an ability to segment more favorably at boundaries of the lesion near and far away from the cavities and with respect to small disconnected components of the lesion, compared to their non-cavity distance penalized counterparts. Whether this is indicative that using this term is allowing the networks to converge faster towards the same minimum in the loss landscape of post-treatment glioma as their unweighted counterparts or allowing convergence to a more optimal global

minimum is unclear. More experiments are required to elucidate this relationship. It should also be noted that alternate experimental set-ups have yet to be tested with variation to loss weightings, iterations, learning rates, other hyperparameters, and architectures. An extensive hyperparameter search for an ideal training configuration for using these loss terms has not yet been done. It may be the case that the distance weighted loss terms may optimize performance when compared to the unweighted terms with alternate hyperparameter configurations or weightings. It may yield more informative results to compare models with the optimal training scheme for each loss function independently. Additionally, the architecture of these models was significantly changed from the models used in the first set of experiments, by reducing the number of convolutional layers in order to fit into smaller GPU with more availability. With a larger number of parameters, these networks may experience higher testing accuracies, and may yield a fairer comparison of training with a cavity distance weighted loss term, perhaps with overall testing Dice scores closer to the current standard of performance.

Additional experiments are needed to determine the exact effect of using a penalization that incorporates post-treatment cavities into training. For the experiments done here, extensive hyperparameter search needs to be done, with many weighting schemes for both variations of the cross entropy loss terms. An automated hyperparameter search can be done for the same networks, with and without using the cavity distance weighted loss terms. Then the optimal training configurations for using these terms can be compared against one another. This may provide a clearer comparison between training with cavity distance maps and without, due to the fact that these different loss functions may have different optimal training configurations. Additionally, using a different metric that places more emphasis on boundary regions of the lesion or disconnected components to evaluate segmentation performance, such as boundary Dice, or by evaluating on a case by case basis, may provide more informative evaluations of the potential gains provided by training with a cavity distance weighted penalization.

Once the effect of training with cavity distance penalization is determined, the most optimal configuration of those loss functions can then be used in additional experiments, to determine their value with respect to transfer learning. A network trained on a large available dataset that is not reflective of the

features unique to post-treatment glioma, such as BraTS, can then be fine-tuned by training on a small subset of data from the post-treatment domain with a loss function that is modified to include features of this domain. This network would be compared to training both on the BraTS dataset mixed with the small subset of post-treatment glioma, and to training on BraTS or newly diagnosed glioma alone. Additionally, weights of the network may be initialized using this loss function by training on a small subset of the post-treatment patients, then transferring the weights into training on the BraTS dataset. This model could be compared to training on BraTS alone or training on the mix of BraTS and post-treatment data using a standard Dice loss.

### *Part 3: Boundary Augmentation*

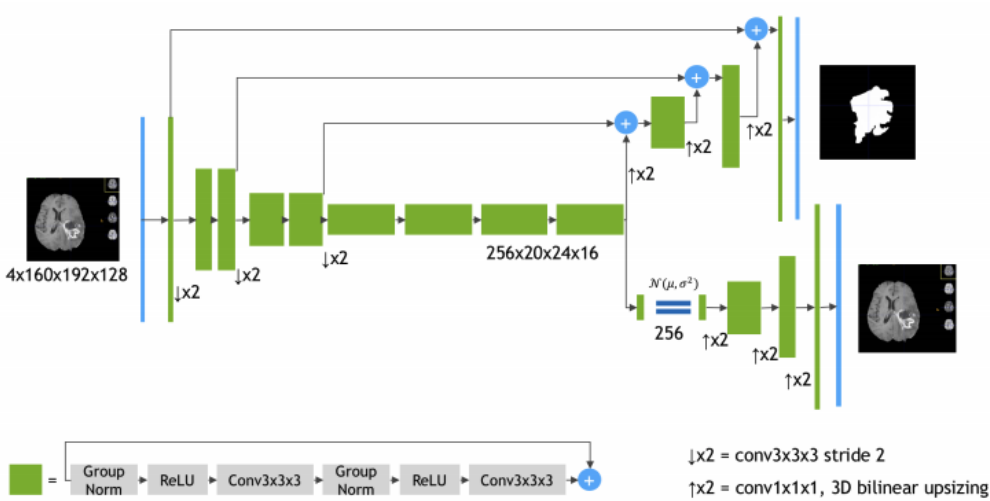
In addition to these methods, a preprocessing pipeline for boundary augmentation been implemented, but not yet incorporated into training. This additional step could improve generalizability to the domain of post-treatment glioma. Although this has previously been done to generate synthetic images by using GANs<sup>[22]</sup>, this requires training a network, where a computationally less complex solution, or one that does not require training an adversarial network may be available.

### **Conclusions:**

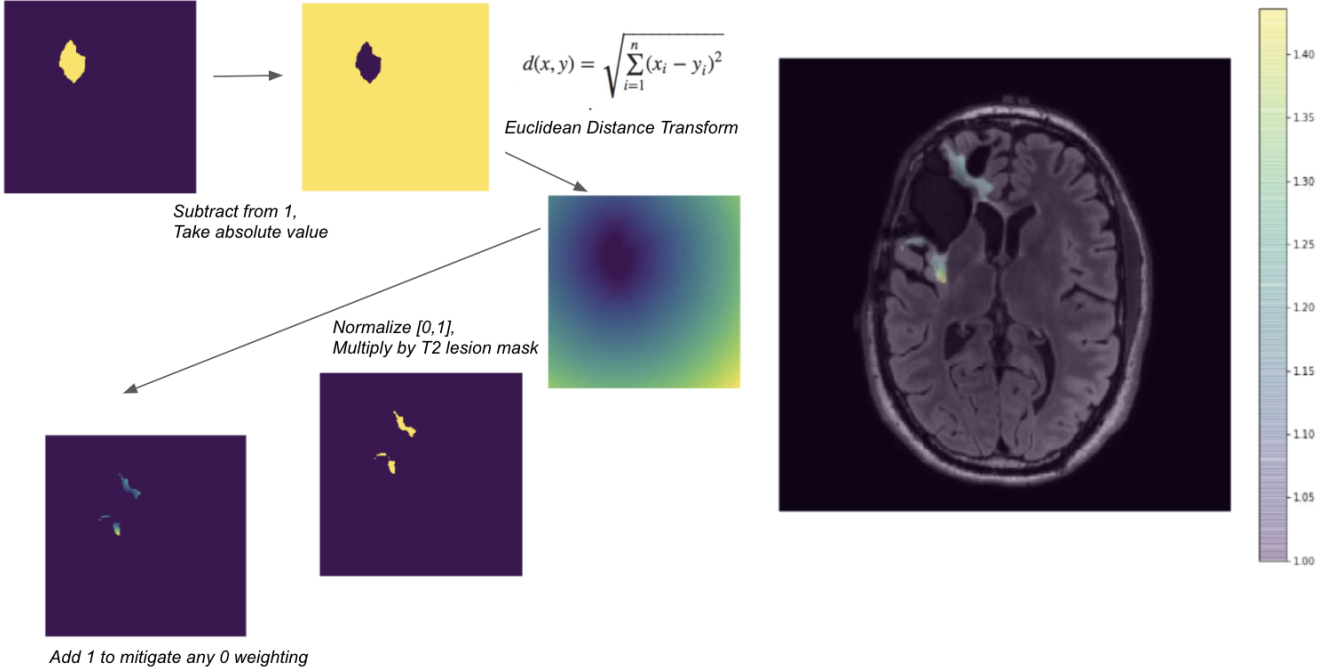
From incorporating images of patients with post-treatment glioma into training, and utilizing a loss term that takes into account the location of the cavity and leading tumor edge into training, we were able to evaluate two methods that can provide useful information towards training a segmentation network that is capable of generalizing to post-treatment glioma. Specifically, we learned that with 200 total training examples, after including a threshold of 30% or greater patients with treated gliomas, segmentation performance of the post-treatment T2 lesion improved and plateaued. We also found that after this inclusion threshold is met, additional images from newly diagnosed glioma did not negatively impact segmentation performance on post-treatment patients. We saw that preliminary models trained with cavity distance penalizations showed qualitatively more desirable segmentations of tumor lesions with respect to cavities and small disconnected components in selected examples. To more clearly

understand the effect of using this loss modification, training with alternate hyperparameters, larger cohorts, and evaluating with more focused performance assessments, will be explored. Lastly, we have developed a pre-processing augmentation method to vary the boundary signal delineation of the T2 lesion, aiming to resemble changes in treated glioma. This augmentation technique, along with including data from post-treatment glioma into training and modifying the loss function to incorporate cavity position relative to the T2 lesion, may be useful methods to improve the generalizability of CNN brain tumor segmentation to treated gliomas.

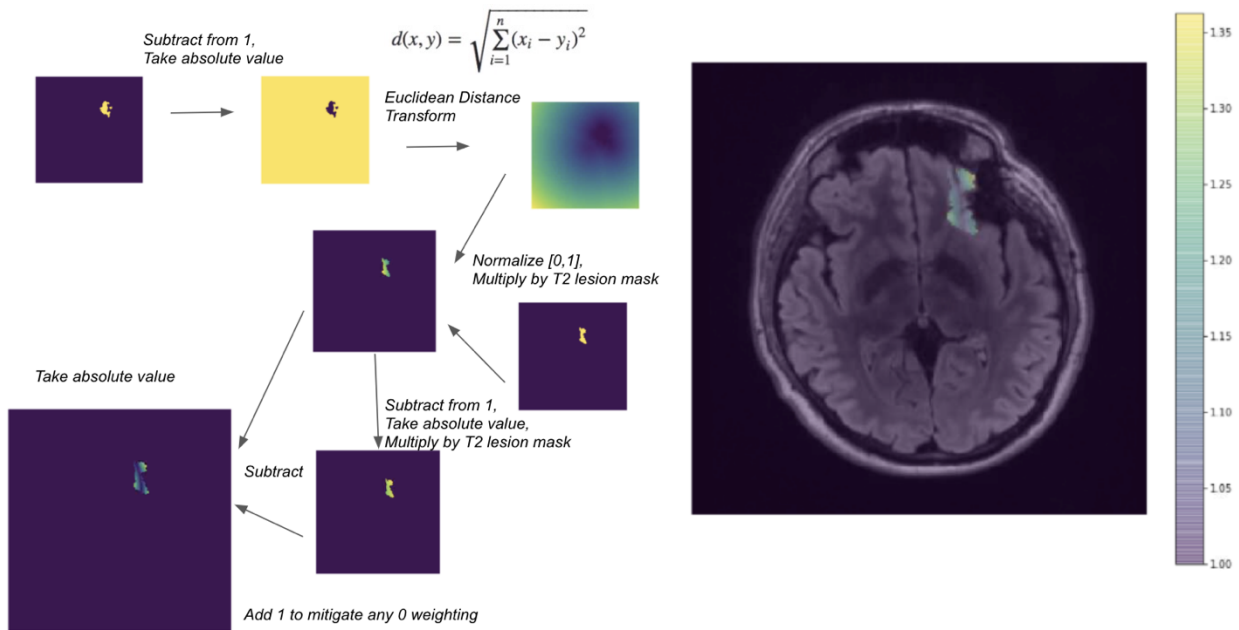
Figures:



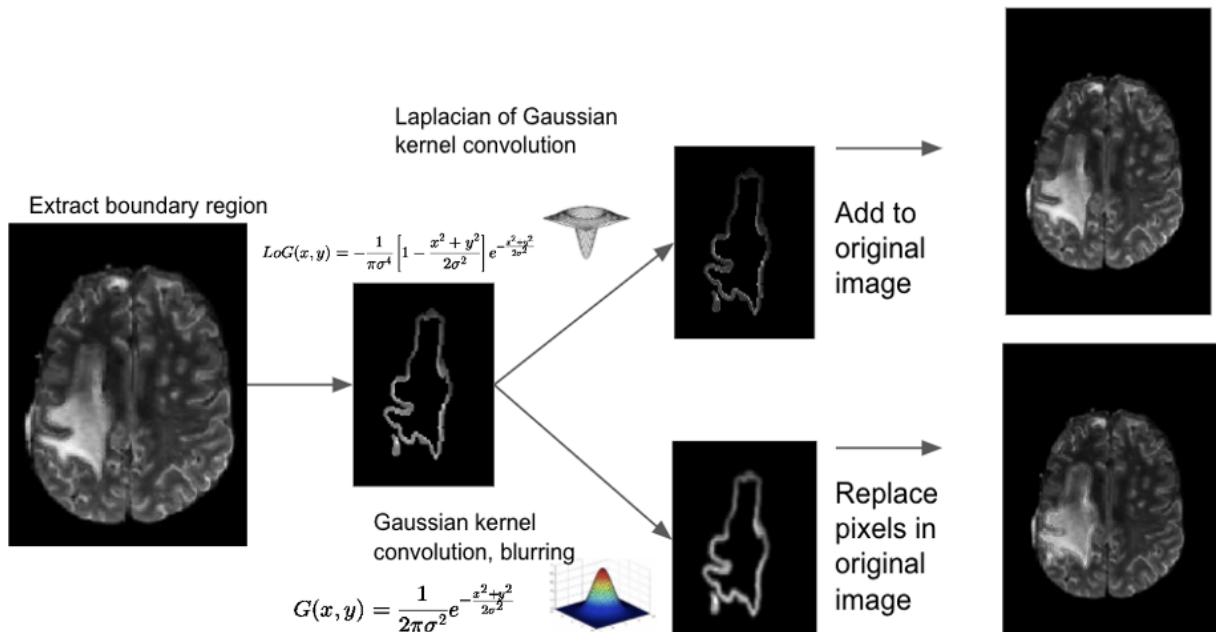
**Figure 1:** Architectural Schematic of the variational auto-encoder taken from the paper on ‘3D MRI brain tumor segmentation using autoencoder regularization’. This is the structure of the model that was modified for the experiments where Nvidia’s Clara v-2 framework was used for training.



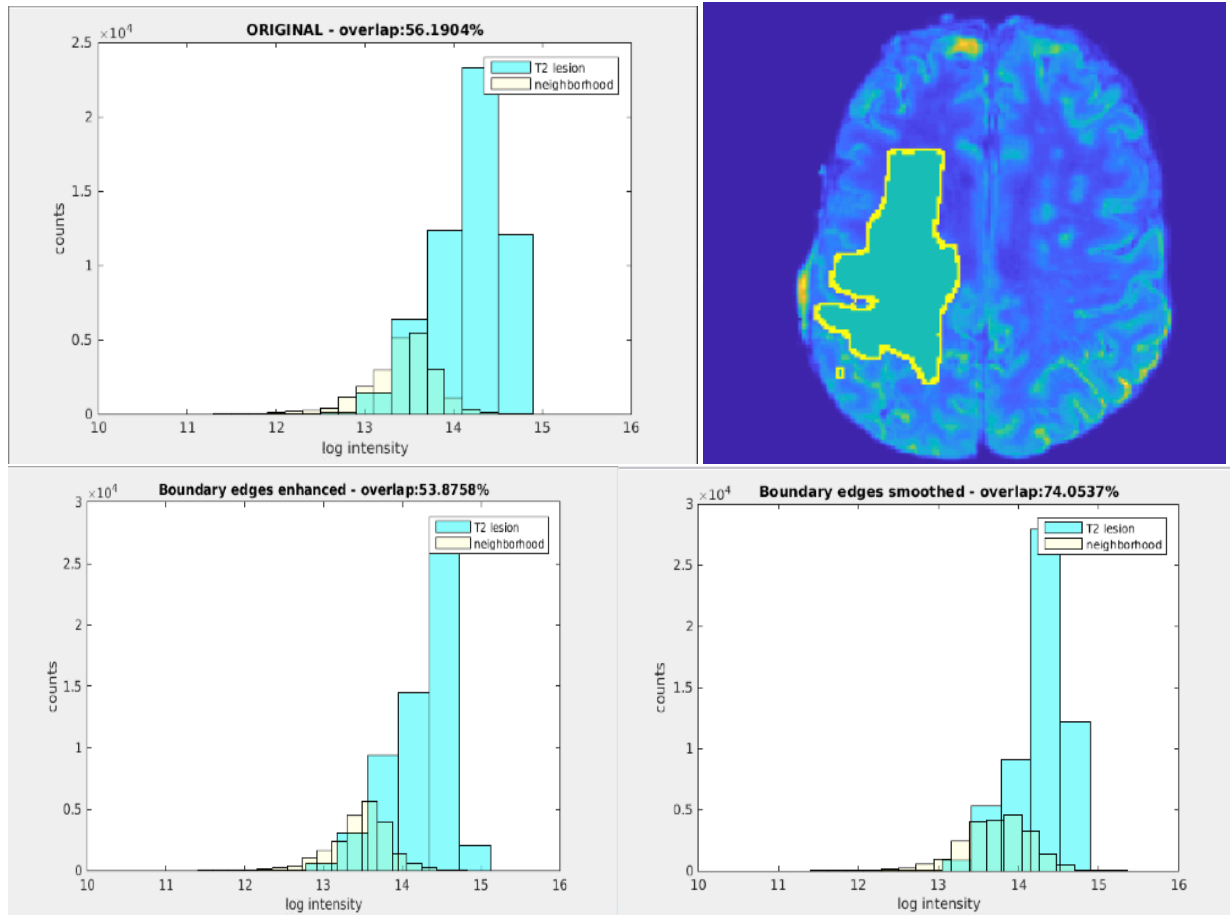
**Figure 2:** Preprocessing pipeline for distance map generation used in implementing the inverse cavity distance based loss term. Pixels of the T2 lesion, weighted inversely by distance to the resection cavity, are weighted most highly in pixels furthest away from the cavity and lowest near to the cavity.



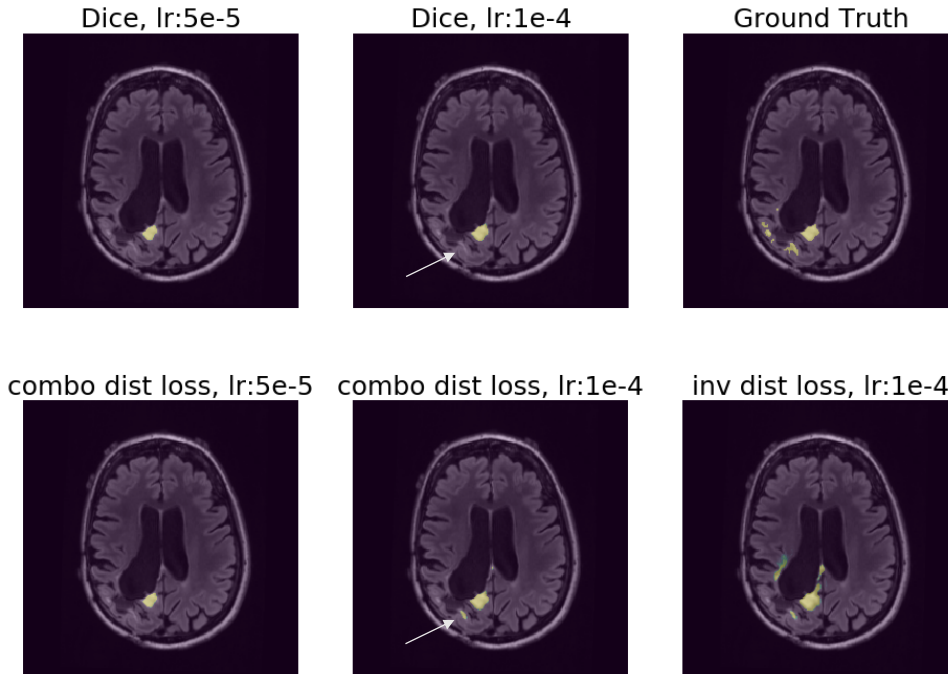
**Figure 3:** Distance map implementation of the combination cavity distance based loss term. The inverse distance weighted T2all lesion is inverted and subtracted from the inverse distance weighted T2all lesion. The absolute value is taken. 1 is added to mitigate zero weighting. Pixels of the T2all lesion both near and far away from the cavity are weighted most highly.



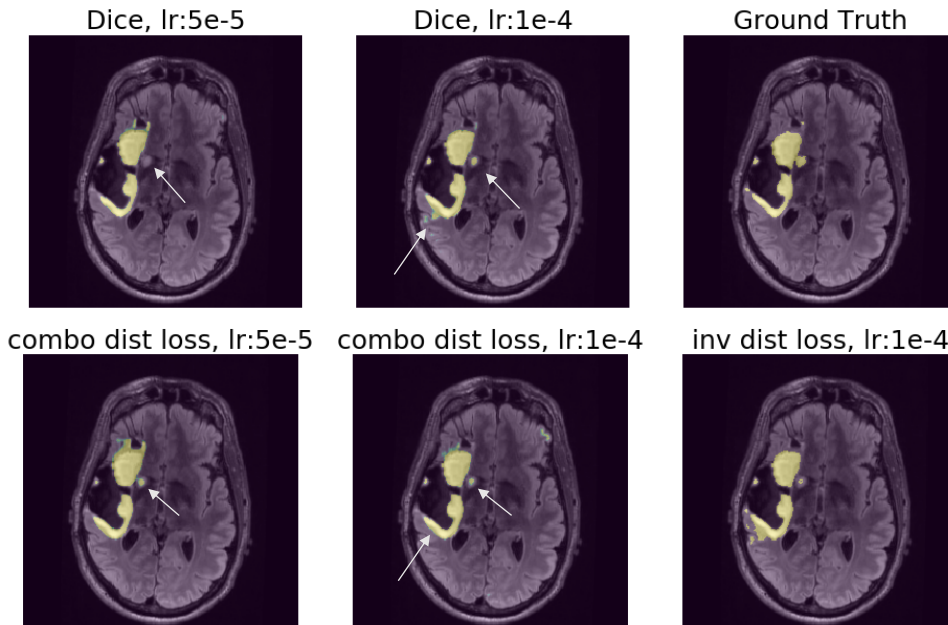
**Figure 4:** Boundary augmentation preprocessing pipeline schematic. Boundary region extraction threshold, Laplacian of Gaussian parameters, and Gaussian blurring parameters can be selected in preprocessing to control the degree of edge sharpening or smoothing or chosen randomly to generate augmented images



**Figure 5: (Top-right)** Non-augmented T2 FLAIR image where healthy tissue surrounding the tumor is selected in yellow, hyperintense tumor is selected in turquoise. **(Top-left and Bottom)** Histograms of the log intensities of healthy tissue (yellow) surrounding the T2 hyperintense lesion (turquoise) on the original T2 FLAIR image and following boundary augmentation. Boundary sharpening operations are shown to decrease the degree of overlap between healthy and tumor intensities compared to the non-augmented image. This indicates increased contrast between tumor and surrounding healthy tissue. Boundary smoothing operation is shown to increase the degree of overlap, indicating decreased contrast between surrounding healthy tissue and tumor.

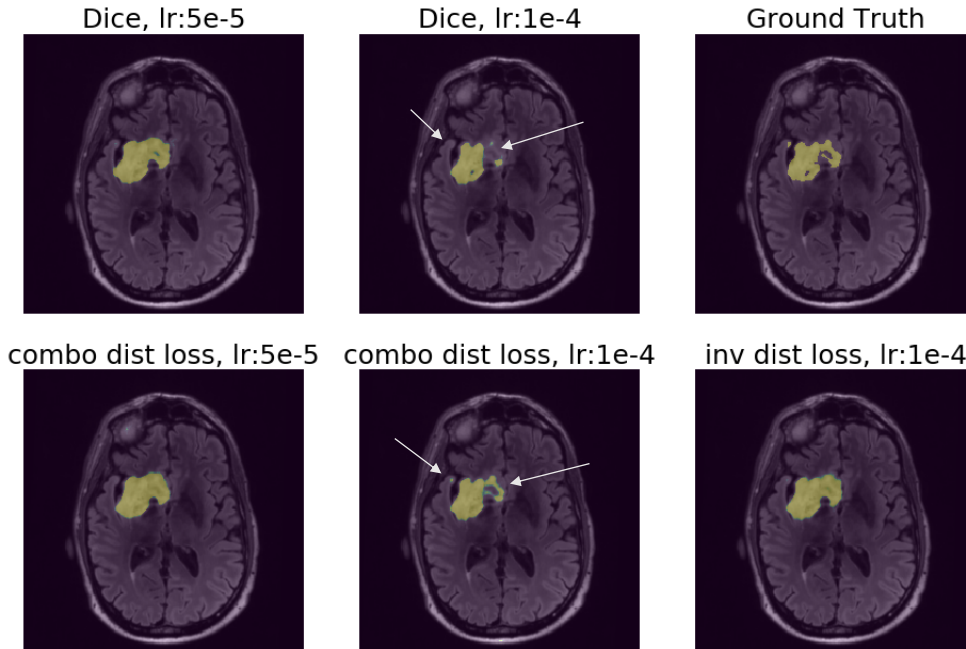


**Figure 6:** Example of testing models trained with various loss functions and learning rates, on 25 patients with post-treatment glioma. Yellow overlay are the outputs of the sigmoid activation, not thresholded. Yellow values represent predicted probabilities closer to 1, while green values represent values closer to 0. Note that the combination distance weighted cross entropy plus Dice loss was able to segment a small detached segment of the lesion below the left ventricle, compared to standard Dice alone at the same learning rate of  $1e-4$ .

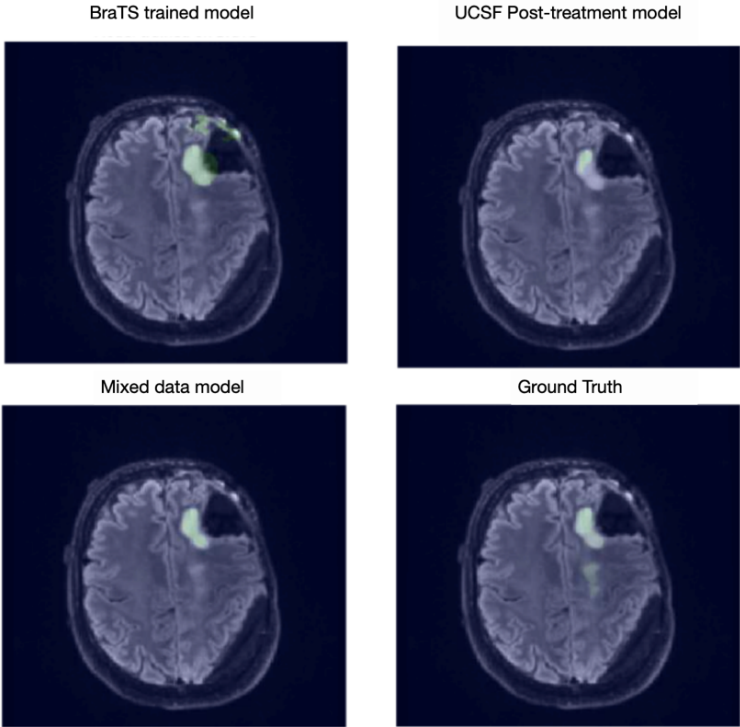


**Figure 7:** Example of testing models trained with various loss functions and learning rates, on 25 patients with post-treatment glioma. Yellow overlay are the outputs of the sigmoid activation, not thresholded. Yellow values represent predicted probabilities closer to 1, while green values represent values closer to 0. The combination distance weighted cross entropy plus Dice loss was able to segment a small detached segment of the lesion to the right of the cavity, compared to Dice at  $5e-5$ . Additionally, at a learning rate of  $1e-4$ , the model trained on Dice alone over segmented the portion of the lesion at the bottom left of the cavity, compared to combination distance loss at the same learning rate.

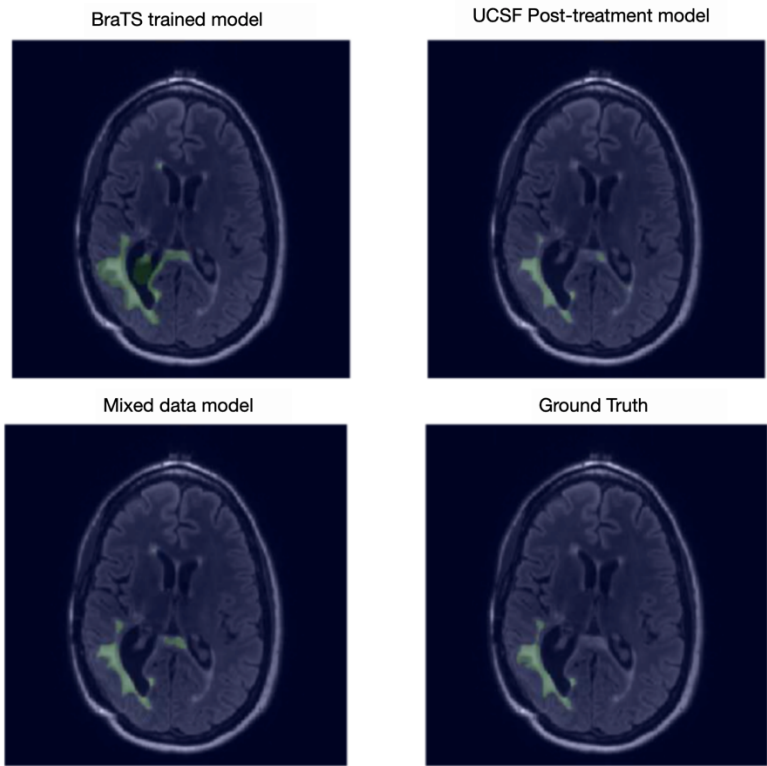




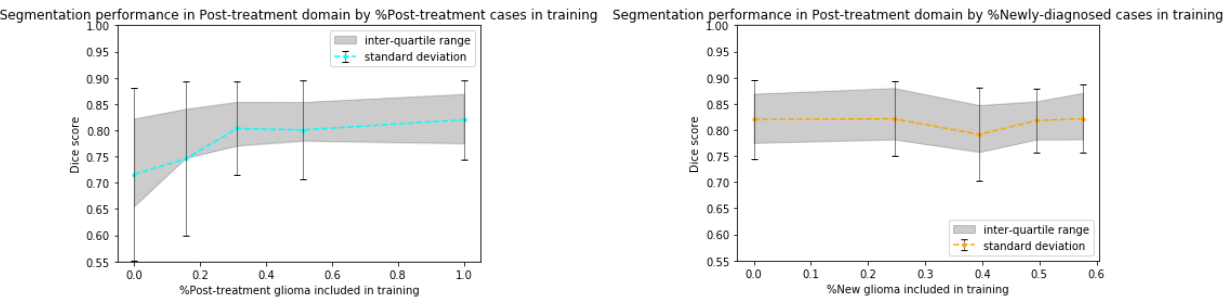
**Figure 8:** Example of testing models trained with various loss functions and learning rates, on 25 patients with post-treatment glioma. Yellow overlay are the outputs of the sigmoid activation, not thresholded. Yellow values represent predicted probabilities closer to 1, while green values represent values closer to 0. Note that the combination distance weighted cross entropy plus Dice loss was able to segment a weakly intense branch to the right of the lesion as well as a small disconnected region of the lesion near the top left of the cavity, compared to Dice alone at learning rate of  $1e-4$ .



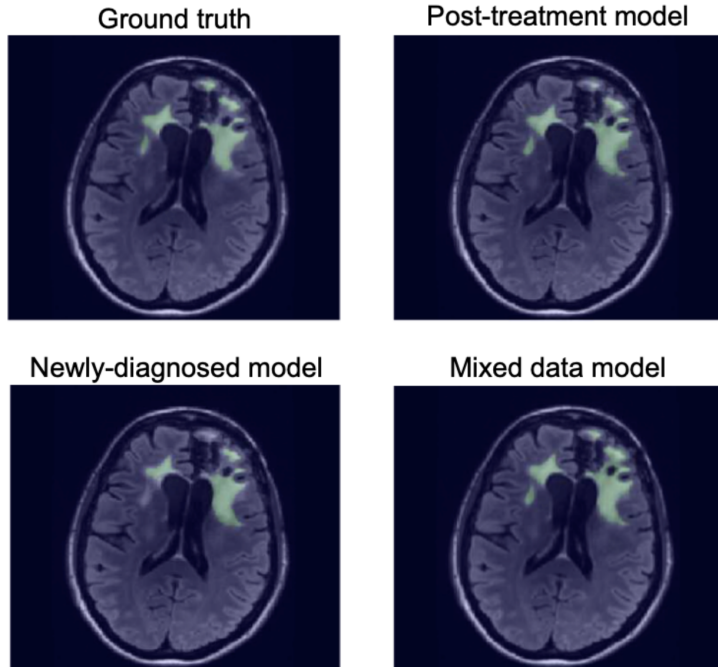
**Figure 9:** Example testing on a patient with post-treatment glioma for the three multi-channel models. BraTS model is seen to over segment near sulci and into the large surgical cavity compared to the models trained on post-treatment data from UCSF.



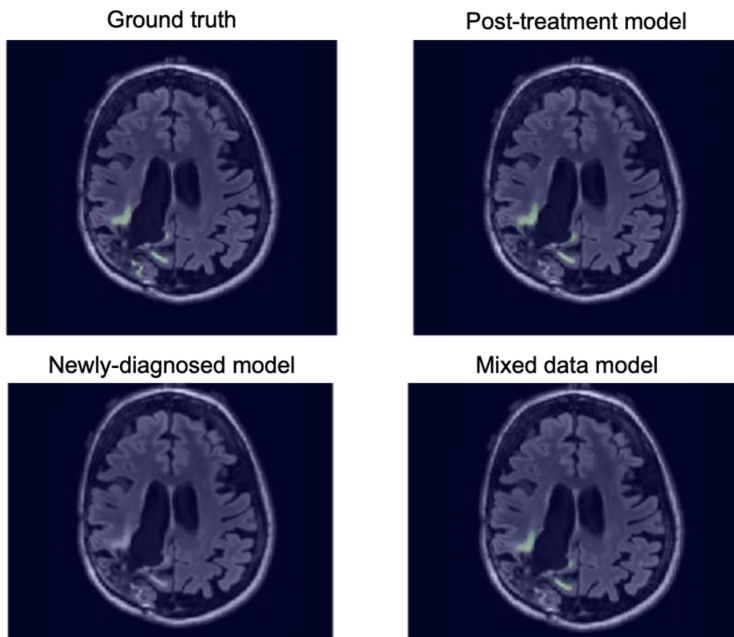
**Figure 10:** Example testing on a patient with post-treatment glioma for the three multi-channel models. The BraTS model over segments into distorted left ventricle compared to the models trained on post-treatment data from UCSF.



**Figure 11: (Left)** Visualization of performance gain from including post-treatment glioma into training. Testing mean Dice scores, interquartile range, and standard deviation of 19 cases of post-treatment glioma are shown. **(Right)** Visualization of no negative performance impact from including newly diagnosed glioma into training, once the threshold for post-treatment patients has been met. The third data point was from a model that stopped training early due to connectivity issues. This model completed 266 epochs of desired 300. Testing mean Dice scores, interquartile range, and standard deviation are shown.



**Figure 12:** Example of testing single-channel models trained on either newly diagnosed patients, post-treatment glioma, or a mixed data at 51% inclusion in training. This example shows under segmentation in the prediction of the newly diagnosed model near the edges of the hyperintense lesion that have weak signal intensity gradients. The mixed model and post-treatment model performed similarly in this case.



**Figure 13:** Example of testing models trained on either newly diagnosed patients, post-treatment glioma, or a mixed data at 51% inclusion in training. This example shows a missed segmentation in the prediction of the newly-diagnosed model near the edges of a distorted left ventricle. The mixed model and post-treatment model performed similarly in this case.

## Tables:

**Table 1:** Testing Dice scores for models trained on either BraTS data, post-treatment glioma acquired at UCSF, or a mix of both. Each model was trained on various amounts of training data with a Train/Val split of 80%/20%. Models were trained on 4 channel inputs of T1, T1 post-gadolinium contrast, T2 FLAIR, and T2 image contrast. Only T2 lesion Dice scores are reported. Models were tested on 25 patients with post-treatment glioma.

Testing Dice Scores	Mean	Standard deviation
BraTS – 227/57	0.721	0.216
UCSF post-treatment glioma – 61/16	0.810	0.107
Mixed data – 288/73	0.830	0.102

**Table 2:** Testing Dice scores for models trained on ratios of post-treatment and newly diagnosed glioma. Each model was trained on 200 images with a Train/Val split of 80%/20%. Models were tested on 19 patients with post-treatment glioma.

Testing Dice Scores	Mean	Standard deviation
Newly diagnosed glioma Only	0.716	0.165
15.625% post-treatment glioma	0.746	0.147
31.25% post-treatment glioma	0.803	0.089
51.25% post-treatment glioma	0.801	0.094
post-treatment glioma only	0.820	0.075

**Table 3:** Testing Dice scores for models trained on ratios of post-treatment and newly diagnosed glioma. Each model was trained by adding in 50 volumes acquired from newly diagnosed glioma at each successive training. The initial, post-treatment glioma only model, started with 153 training images and 39 validation images, both from post-treatment glioma. Models were tested on 19 patients with post-treatment glioma. The model at 39.53% newly diagnosed glioma included into training was stopped at 266 epochs instead of 300 due to connectivity issues.

Testing Dice Scores	Mean	Standard deviation
post-treatment glioma only	0.820	0.075
24.63% Newly diagnosed glioma	0.821	0.071
39.53% Newly diagnosed glioma*	0.791	0.089
49.50% Newly diagnosed glioma	0.818	0.061
57.62% Newly diagnosed glioma	0.821	0.066

**Table 4:** Testing Dice scores for models trained on 25 post-treatment glioma. Models were trained using various loss functions and learning rates. Each model was selected for testing based on highest validation accuracy from validation with 25 patients with post-treatment glioma before 8000 iterations. Models were tested on 25 separate patients with post-treatment glioma. Mean and standard deviations of the testing Dice scores for each model are reported below.

Testing Dice Scores	Mean	Standard deviation
Dice loss, Learning rate: 1e-4	0.6247	0.2111
Dice loss, Learning rate: 5e-5	0.7078	0.1456
Combination Cavity Distance Weighted Cross Entropy plus Dice loss, Learning rate: 1e-4	0.6551	0.1987
Combination Cavity Distance Weighted Cross Entropy plus Dice loss, Learning rate: 5e-5	0.6601	0.1686
Inverse Cavity Distance Weighted Cross Entropy plus Dice loss, Learning rate: 1e-4	0.6383	0.1856

## References

- (1) Louis, D. N.; Perry, A.; Reifenberger, G.; von Deimling, A.; Figarella-Branger, D.; Cavenee, W. K.; Ohgaki, H.; Wiestler, O. D.; Kleihues, P.; Ellison, D. W. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: A Summary. *Acta Neuropathol. (Berl.)* 2016, *131* (6), 803–820. <https://doi.org/10.1007/s00401-016-1545-1>.
- (2) Huang, R. Y.; Rahman, R.; Ballman, K. V.; Felten, S. J.; Anderson, S. K.; Ellingson, B. M.; Nayak, L.; Lee, E. Q.; Abrey, L. E.; Galanis, E.; Reardon, D. A.; Pope, W. B.; Cloughesy, T. F.; Wen, P. Y. The Impact of T2/FLAIR Evaluation per RANO Criteria on Response Assessment of Recurrent Glioblastoma Patients Treated with Bevacizumab. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* 2016, *22* (3), 575–581. <https://doi.org/10.1158/1078-0432.CCR-14-3040>.
- (3) Wen, P. Y.; Macdonald, D. R.; Reardon, D. A.; Cloughesy, T. F.; Sorensen, A. G.; Galanis, E.; Degroot, J.; Wick, W.; Gilbert, M. R.; Lassman, A. B.; Tsien, C.; Mikkelsen, T.; Wong, E. T.; Chamberlain, M. C.; Stupp, R.; Lamborn, K. R.; Vogelbaum, M. A.; van den Bent, M. J.; Chang, S. M. Updated Response Assessment Criteria for High-Grade Gliomas: Response Assessment in Neuro-Oncology Working Group. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* 2010, *28* (11), 1963–1972. <https://doi.org/10.1200/JCO.2009.26.3541>.
- (4) Survival Rates for Selected Adult Brain and Spinal Cord Tumors <https://www.cancer.org/cancer/brain-spinal-cord-tumors-adults/detection-diagnosis-staging/survival-rates.html> (accessed May 28, 2020).
- (5) Bondy, M. L.; Scheurer, M. E.; Malmer, B.; Barnholtz-Sloan, J. S.; Davis, F. G.; Il'yasova, D.; Kruchko, C.; McCarthy, B. J.; Rajaraman, P.; Schwartzbaum, J. A.; Sadetzki, S.; Schlehofer, B.; Tihan, T.; Wiemels, J. L.; Wrensch, M.; Buffler, P. A. Brain Tumor Epidemiology: Consensus from the Brain Tumor Epidemiology Consortium (BTEC). *Cancer* 2008, *113* (7 Suppl), 1953–1968. <https://doi.org/10.1002/cncr.23741>.
- (6) Kickingereder, P.; Isensee, F.; Tursunova, I.; Petersen, J.; Neuberger, U.; Bonekamp, D.; Brugnara, G.; Schell, M.; Kessler, T.; Foltyn, M.; Harting, I.; Sahm, F.; Prager, M.; Nowosielski, M.; Wick, A.; Nolden,

- M.; Radbruch, A.; Debus, J.; Schlemmer, H.-P.; Heiland, S.; Platten, M.; Deimling, A. von; Bent, M. J. van den; Gorlia, T.; Wick, W.; Bendszus, M.; Maier-Hein, K. H. Automated Quantitative Tumour Response Assessment of MRI in Neuro-Oncology with Artificial Neural Networks: A Multicentre, Retrospective Study. *Lancet Oncol.* 2019, 20 (5), 728–740. [https://doi.org/10.1016/S1470-2045\(19\)30098-1](https://doi.org/10.1016/S1470-2045(19)30098-1).
- (7) Chang, K.; Beers, A. L.; Bai, H. X.; Brown, J. M.; Ly, K. I.; Li, X.; Senders, J. T.; Kavouridis, V. K.; Boaro, A.; Su, C.; Bi, W. L.; Rapalino, O.; Liao, W.; Shen, Q.; Zhou, H.; Xiao, B.; Wang, Y.; Zhang, P. J.; Pinho, M. C.; Wen, P. Y.; Batchelor, T. T.; Boxerman, J. L.; Arnaout, O.; Rosen, B. R.; Gerstner, E. R.; Yang, L.; Huang, R. Y.; Kalpathy-Cramer, J. Automatic Assessment of Glioma Burden: A Deep Learning Algorithm for Fully Automated Volumetric and Bidimensional Measurement. *Neuro-Oncol.* 2019, 21 (11), 1412–1422. <https://doi.org/10.1093/neuonc/noz106>.
- (8) Menze, B. H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; Lanczi, L.; Gerstner, E.; Weber, M.-A.; Arbel, T.; Avants, B. B.; Ayache, N.; Buendia, P.; Collins, D. L.; Cordier, N.; Corso, J. J.; Criminisi, A.; Das, T.; Delingette, H.; Demiralp, Ç.; Durst, C. R.; Dojat, M.; Doyle, S.; Festa, J.; Forbes, F.; Geremia, E.; Glocker, B.; Golland, P.; Guo, X.; Hamamci, A.; Iftekharuddin, K. M.; Jena, R.; John, N. M.; Konukoglu, E.; Lashkari, D.; Mariz, J. A.; Meier, R.; Pereira, S.; Precup, D.; Price, S. J.; Raviv, T. R.; Reza, S. M. S.; Ryan, M.; Sarikaya, D.; Schwartz, L.; Shin, H.-C.; Shotton, J.; Silva, C. A.; Sousa, N.; Subbanna, N. K.; Szekely, G.; Taylor, T. J.; Thomas, O. M.; Tustison, N. J.; Unal, G.; Vasseur, F.; Wintermark, M.; Ye, D. H.; Zhao, L.; Zhao, B.; Zikic, D.; Prastawa, M.; Reyes, M.; Van Leemput, K. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans. Med. Imaging* 2015, 34 (10), 1993–2024. <https://doi.org/10.1109/TMI.2014.2377694>.
- (9) 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. *Med. Image Anal.* 2017, 35, 18–31. <https://doi.org/10.1016/j.media.2016.05.004>.

- (10) Reddy, C.; Gopinath, K.; Lombaert, H. Brain Tumor Segmentation Using Topological Loss in Convolutional Networks. 4.
- (11) Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *ArXiv150504597 Cs* 2015.
- (12) Milletari, F.; Navab, N.; Ahmadi, S.-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *ArXiv160604797 Cs* 2016.
- (13) Myronenko, A. 3D MRI Brain Tumor Segmentation Using Autoencoder Regularization. *ArXiv181011654 Cs Q-Bio* 2018.
- (14) Kamnitsas, K.; Baumgartner, C.; Ledig, C.; Newcombe, V. F. J.; Simpson, J. P.; Kane, A. D.; Menon, D. K.; Nori, A.; Criminisi, A.; Rueckert, D.; Glocker, B. Unsupervised Domain Adaptation in Brain Lesion Segmentation with Adversarial Networks. *ArXiv161208894 Cs* 2016.
- (15) Perone, C. S.; Ballester, P.; Barros, R. C.; Cohen-Adad, J. Unsupervised Domain Adaptation for Medical Imaging Segmentation with Self-Ensembling. *ArXiv181106042 Cs* 2019.
- (16) Zhang, Y.; Wei, Y.; Zhao, P.; Niu, S.; Wu, Q.; Tan, M.; Huang, J. Collaborative Unsupervised Domain Adaptation for Medical Image Diagnosis. *ArXiv191107293 Cs Stat* 2019.
- (17) Ghafoorian, M.; Mehrtash, A.; Kapur, T.; Karssemeijer, N.; Marchiori, E.; Pesteie, M.; Guttmann, C. R. G.; de Leeuw, F.-E.; Tempany, C. M.; van Ginneken, B.; Fedorov, A.; Abolmaesumi, P.; Platel, B.; Wells III, W. M. Transfer Learning for Domain Adaptation in MRI: Application in Brain Lesion Segmentation. *ArXiv170207841 Cs* 2017, *10435*, 516–524. [https://doi.org/10.1007/978-3-319-66179-7\\_59](https://doi.org/10.1007/978-3-319-66179-7_59).
- (18) Valindria, V. V.; Lavdas, I.; Bai, W.; Kamnitsas, K.; Aboagye, E. O.; Rockall, A. G.; Rueckert, D.; Glocker, B. Domain Adaptation for MRI Organ Segmentation Using Reverse Classification Accuracy. 9.
- (19) Kouw, W. M.; Loog, M. An Introduction to Domain Adaptation and Transfer Learning. *ArXiv181211806 Cs Stat* 2019.
- (20) Caliva, F.; Iriondo, C.; Martinez, A. M.; Majumdar, S.; Pedoia, V. Distance Map Loss Penalty Term for Semantic Segmentation. 5.



- (21) Hatamizadeh, A.; Terzopoulos, D.; Myronenko, A. End-to-End Boundary Aware Networks for Medical Image Segmentation. 8.
- (22) Mok, T. C. W.; Chung, A. C. S. Learning Data Augmentation for Brain Tumor Segmentation with Coarse-to-Fine Generative Adversarial Networks. *ArXiv180511291 Cs* 2019, *11383*, 70–80.  
[https://doi.org/10.1007/978-3-030-11723-8\\_7](https://doi.org/10.1007/978-3-030-11723-8_7).
- (23) Jungo, A.; Meier, R.; Ermis, E.; Herrmann, E.; Reyes, M. Uncertainty-Driven Sanity Check: Application to Postoperative Brain Tumor Cavity Segmentation. *ArXiv180603106 Cs* 2018.
- (24) Wang, G.; Li, W.; Ourselin, S.; Vercauteren, T. Automatic Brain Tumor Segmentation Based on Cascaded Convolutional Neural Networks With Uncertainty Estimation. *Front. Comput. Neurosci.* 2019, *13*. <https://doi.org/10.3389/fncom.2019.00056>.
- (25) Wacker, J.; Ladeira, M.; Nascimento, J. E. V. Transfer Learning for Brain Tumor Segmentation. *ArXiv191212452 Cs Eess* 2019.
- (26) Iglovikov, V.; Shvets, A. TerausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation. *ArXiv180105746 Cs* 2018.
- (27) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *ArXiv151203385 Cs* 2015.
- (28) Jenkinson, M.; Pechaud, M.; Smith, S. BET2 - MR-Based Estimation of Brain, Skull and Scalp Surfaces. 1.

## Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:

*Jacob Ellison*

BDFDD0E6EC69494...

Author Signature

9/7/2020

Date