Towards improved medical image segmentation using deep learning

by

Nabila Abraham

Bachelor of Engineering, Biomedical Engineering,

Ryerson University, 2017

A thesis

presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Master of Applied Science

in the Program of

Electrical and Computer Engineering

Toronto, Ontario, Canada, 2019

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A THESIS

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my dissertation may be made electronically available to the public.

Towards improved medical image segmentation using deep learning

Master of Applied Science 2019

Nabila Abraham

Electrical and Computer Engineering

Ryerson University

Abstract

Convolutional neural networks have been asserted to be fast and precise frameworks with great potential in image segmentation. Within the medical domain, image segmentation is a pre-cursor to several applications including surgical simulations, treatment planning and patient prognosis. In this thesis, we attempt to solve two major limitations of current segmentation practices: 1) dealing with unbalanced classes and 2) dealing with multiple modalities. In medical imaging, unbalanced classes present as the regions of interest that are typically significantly smaller in volume than the background class or other classes. We propose an improvement to the current gold standard cost function to boost the focus of the network to the smaller classes. Another problem within medical imaging is the variation in both anatomy and pathology across patients. Utilizing multiple imaging modalities provides complementary, segmentation-specific information and is commonly employed by radiologists when contouring data. We propose a image fusion strategy

for multi-modal data that uses the variation in modality specific features to guide the task specific learning. Together, our contributions propose a framework to maximize the representational power of the dataset using models with less complexity and higher generalizability. Our contributions outperform baseline models for multi-class segmentation and are modular enough to be scaled up to deeper networks. We demonstrate the effectiveness of the proposed cost function and multimodal framework, both individually and together, on benchmark datasets including the Breast Ultrasound Dataset B (BUS) [1], the International Skin Imaging Collaboration (ISIC 2018) [2], [3] and the Brain Tumor Segmentation Challenge (BraTs 2018) [4]. In all experiments, the proposed methods match or outperform the baseline methods while employing simpler networks.

Acknowledgements

I want to express my greatest thanks to Dr. Naimul Khan for his unwavering support and encouragement over the last two years. His confidence in my abilities has inspired me to continue on an academic path and I am truly thankful for his guidance and his friendship.

I am particularly fortunate to have been a part of the Ryerson Multimedia Lab (RML) where I have met the brightest minds that have all had an impact on me over the last two years. Thank you to all RML members for the discussions, the research feedback, the code review and a memorable graduate experience.

Dedication

To my loving parents,
my supportive sisters,
my semi-encouraging partner,
and
my hardworking 1080 Ti.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Problem context

Medical image analysis plays a substantial role in determining patient prognosis and diagnosis. From a computer vision (CV) perspective, image segmentation is the process of dividing the individual pixels of an image into a set of groups that have similar properties [5]. In the medical domain, the common property signifies that all pixels belong to a specific anatomical structure. Image segmentation is arguably the most important part of the image processing pipeline as it has applications in functional mapping, surgical simulations, mass and tumor detections, clinical studies and treatment planning [6]. In contrast to natural images, medical imaging modalities have varied acquisition protocols and contrast injections which results in images that highlight different anatomical structures. For example, Magnetic Resonance Imaging (MRI) provides detailed soft tissue definition and is commonly used to diagnose brain abnormalities. Computed Tomography (CT) scans provide high resolution images in a short exposure time and are used heavily in imaging bony structures [7]. Positron Emission Tomography (PET) images have the property of high sensitivity due to the molecular imaging technique, but output lower resolution images and are used to map functional processes [7]. Clinicians rely on these varied imaging schemes to delineate tissue abnormalities. However, due to large variations in pathology, imaging scanners, inter-observer variability and labor intensive

manual contouring, clinicians have recently begun to benefit from computer-assisted interventions.

Automatic and semi-automatic image segmentation practices rely on feature descriptors based on shape, texture, spatial arrangement and image statistics [8]. These conventional approaches are handcrafted based on prior knowledge and are tailored for specific imaging modalities. Therefore, intelligent feature selection lies at the heart of such computer-assisted approaches. Recently, deep learning approaches have had unprecedented success in medical image segmentation due to their ability to *learn* task-specific features without any human intervention. The burden of feature engineering is now absorbed into an optimization problem allowing for automated approaches with improved generalization.

Most of the dominant approaches in image segmentation in the medical domain stem from ideas proposed for natural images in the CV society. A common trend is to build deeper models with more layers and parameters. This approach is highly effective when the number of training samples is large such that the model has the capacity to extract highly complex features from a diversified sample set. However, for medical applications, we are limited to a small sample size due to privacy restrictions associated with open-source data and cost associated with annotation. Therefore, one of the main challenges arises from the ability to build deep models without suffering from over-fitting. Class imbalance is another dataset issue that arises when the regions of interest (ROI) class is significantly underrepresented in comparison to other classes. Within small lesion segmentation for example, models are biased to the background class and produce predictions with low sensitivity.

Another challenge within medical segmentation is that the variation in pixel space is much smaller when compared to natural images that depict distinct objects. As a result, deep models for medical applications can benefit from image fusion practices which involve combining complementary imaging modalities of the same anatomy. It has been empirically proven that multi-modal segmentation performance when compared to using a single modality [9] is superior. In this thesis, we propose architectural and learning based constraints to combat these commonly faced challenges in medical based segmentation.

1.2 Contributions

This thesis explores two prevalent problems related to medical image segmentation. ROI in medical images typically occupy a small fraction of the image or volume space. The imbalance between the ROI class and the background class leads to instability in training and can produce segmentation maps with low sensitivity. We propose a new loss function to mitigate the effects of class imbalance and empirically show it is superior to current methods in terms of balanced precision and recall curves.

Secondly, we tackle the multi-modal data fusion problem. Large variations in pathology make it difficult for one modality to delineate boundaries between healthy and non-healthy tissue. We propose a latent model for feature fusion of multiple heterogeneous imaging modalities. In contrast to conventional deep learning techniques, we show our approach is less complex in terms of learnable parameters and is able to achieve competitive performance. Together, our contributions propose a framework that aims to maximize the representational power of the small dataset for a given task and outperforms baseline models with increased model complexity.

1.3 Structure of Thesis

The rest of this thesis is structures as follows: Chapter 2 reviews recent literature on conventional methods to improve class imbalance and recent multi-modal strategies for image segmentation. Chapter 3 describes in detail our proposed Focal Tversky loss function with an attention U-net variant. In Chapter 4, we detail our multi-modal fusion strategy. In both studies, we present implementation details, ablation studies and experimental results on benchmark datasets. We conclude our work and discuss future research direction in Chapter 5.

Chapter 2

Related works

In this section, we review the literature on deep learning based approaches to medical image segmentation. We highlight common strategies to deal with class imbalance and popular learning-based pipelines for multi-modal data fusion.

2.1 Deep learning

Deep learning based applications have been popularized by the success of AlexNet, a Convolutional Neural Network (CNN) which won first place in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012 [10]. CNNs are a specialized type of neural network that contain multiple layers of stacked convolution operations [11]. Filter weights for each convolutional layer are learned by backpropagating the error computed by a cost function. Gradient-based optimization methods minimize the cost function which guides the model parameters to their most optimal setting. The strength behind deep networks lies in their ability to extract rich hierarchical features. With progressive convolutions, the effective spatial field of view of each filter (also known as the receptive field), increases with more layers allowing for deeper models to extract more complex, non-linear combinations of low level features.

2.1.1 Semantic Segmentation using deep learning

AlexNet was the first of many CNNs proposed for image classification. Current state-of-the-art CNNs boast deeper layers such as the Visual Geometry Group (VGG) network [12], or contain branched layers for improved gradient flow as proposed in Residual Networks (ResNet) [13]. Progress on whole-image classification extended into structured output tasks such as semantic segmentation. The Fully Convolutional Network (FCN) was the first work to train CNNs in an end-to-end manner for pixelwise prediction using supervised learning [14]. FCNs reuse the convolutional feature pipeline from classification networks to *encode* the image space into a latent feature space resulting in coarse feature maps. A symmetric convolutional pipeline is used to *decode* the latent features into pixelwise prediction maps through upsampling operations. The convolutional nature of the encoder and decoder blocks allow the model to extract semantic features and localize them with respect to the original input image. The FCN model has been widely adopted in the CV society, however, it requires a large training set size to predict segmentation maps with high localization accuracy. Since this is not within reach for biomedical tasks, Ronneberger et al proposed U-Net, the current state of the art in medical image segmentation tasks [15]. U-Nets build upon FCN by incorporating skip connections between corresponding encoder and decoder blocks, as depicted in Figure 2.1. By combining high resolution feature maps from the encoding path with upsampled decoded representations, successive convolutions are able to assemble more precise segmentation predictions.

Several efforts to advance medically focused image segmentation try to encode or decode features by using newer and more powerful architectures. The popular ResNets architecture introduces a residual identity branch of the input feature that skips one or more layers [16]. Gradients can flow through the short-cut connections making it easier to train deeper models with residual connections. In the medical community, several authors adopt ResNet based encoders to improve feature extraction in the U-Net segmentation pipeline [17], [18]. Dense connections were also proposed to improve gradient flow by concatenating several feature maps of the same scale. Instead of going deeper, DenseNets exploit the potential of the network through feature reuse. Similarly, many medical works incorporate dense connections in their segmentation pipeline [19], [20]. Since U-Net is still the gold standard for medical segmentation problems, it is common

to adopt typical CV strategies into the encoder, decoder or skip connection blocks.

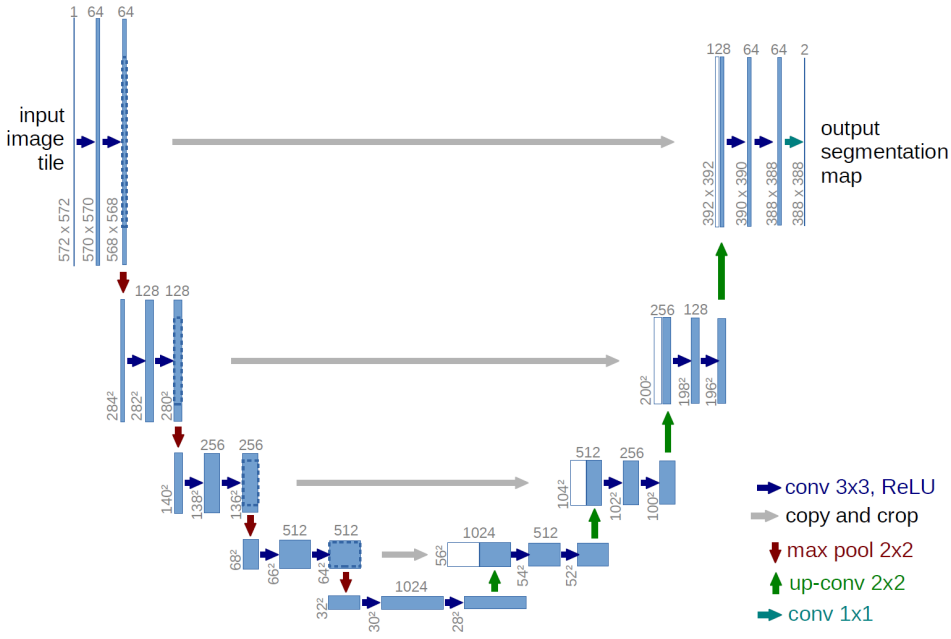


Figure 2.1: The U-Net architecture proposed by Ronneberger et al [15]. Each blue box corresponds to a multi-channel feature map. The number of channels present at each layer is denoted on top of the box. The spatial dimensions are provided at the lower left edge of each box. White boxes represent features that have been mirrored and propagated to the decoding layers.

2.2 Class Imbalance

Within semantic segmentation in the medical community, class imbalance is a recurring issue. Class imbalance arises when the ROI is under-represented in the dataset. In medical imaging applications, this phenomenon is especially problematic as the unbalanced training leads to predictions with low sensitivity. From a clinicians perspective, a model with high precision but low recall to a certain pathology may not be useful. The extensive review by Buda et al identify two major categories of methods for addressing class imbalance [21]. The first is a data-level method that alters the dataset by changing class distributions. The other category covers algorithmic level methods that keep the training set unchanged while adjusting training or inference algorithms.

2.2.1 Data-level methods

Oversampling is one of the most commonly used data-level methods to tackle imbalanced datasets [21]. The most basic version uses random minority sampling which randomly oversamples from minority classes. In the first FCN for segmentation, Long et al utilize a patch-based approach and oversample the ROI class by a hyperparameter factor [14]. Standard efforts in oversampling involve extensive data augmentation through affine, geometric and photometric trasnformations [21], and very recently, Generative Adversarial Networks (GANs) [22], [23]. Oversampling has been shown to be effective but can lead to overfitting [24]. Conversely, Valverde et al randomly undersample the larger, negative class in their data corpus by effectively ignoring certain images [25]. A major disadvantage of this approach is that it discards a portion of available data. To overcome this shortcoming, the training set can be intelligently pruned for redundant or weak data points. For example, one-sided selection identifies redundant examples close to the boundary between classes and removes them from the training set [26].

2.2.2 Algorithmic Improvements

Cost functions

While sampling based methods are commonly used, these approaches tend to change the *apriori* distribution of the classes and can result in biased, over-segmentations [21]. Cost sensitive learning is an algorithmic effort to tackle class imbalance and is commonly approached using one of two popular loss functions. The **Cross-entropy** loss is commonly used in the CV society for multi-class segmentation. It compares the model's prediction p_{ic} for class c, over pixel-space i with the true data distribution g_{ic}, depicted in Equation 2.1.

$$CE = -\sum_{c}\sum_{i} g_{ic}\log(p_{ic}) \tag{2.1}$$

This loss is effective as it penalizes the model in an exponential fashion if predicted class probabilities approach zero. The downside of cross-entropy is that it assumes that all

classes are weighted equally, which is detrimental in the medical domain as the ROI class is typically significantly smaller than the background class. A weighted variant of cross-entropy Weighted Cross-Entropy (WCE) is notably used by Ronneberger et al [15] in a medical image setting and is depicted in Equation 2.2.

$$WCE = -\sum_{c}\sum_{i} w_c g_{ic} \log(p_{ic}) \tag{2.2}$$

where the class weights w_c are pre-computed using the inverse class frequency [15] or more commonly treated as a hyperparameter to be set through cross-validation [27].

Reweighting the cross-entropy loss is adopted in several works such as in brain lesions [28], [29] and brain tumor segmentation [30], but it becomes difficult to calibrate the class weights and is, therefore, very application-specific. To combat class-reweighting, Milletari et al proposed a loss function based on the Dice Similarity Coefficient (DSC), parametrized in Equation 2.3 [31].

$$DSC = \frac{2 * TP}{2 * TP + FP + FN} \tag{2.3}$$

where the DSC is computed over all the predicted classes c. The Dice score is effective because it is a harmonic mean of precision and recall and, therefore, weights false positive (FP) and false negative (FN) predictions, equally. The Dice score is reshaped into the Dice loss (DL) by minimizing the complement of DSC summed over all classes c, as denoted in Equation 2.4 [1]:

$$DL = \frac{1}{C}\sum_{c} 1 - DSC_c \tag{2.4}$$

The Dice loss function is heavily adapted in the literature for both the binary and multi-class segmentation as it takes into account the spatial arrangeoment of pixels in a holistic

[1]The Dice loss is typically calculated as $1 - DSC$. However, in practice, it is common to simply minimize $-DSC$. Moreover, because the DSC is a normalized metric, the DL is typically scaled by the number of classes $\frac{1}{C}$. However in practice, it is common to simply minimize the $-DSC$ *without* class normalization as this produces a larger error signal for gradient updates. In this case, the maximum DSC is equivalent to the number of classes. In this thesis, we present normalized metrics for consistency with the literature while our code repositories reflect the conventional code practices.

manner. However, because medical images are still plagued by many small ROIs when compared to the background class, the DL tends to produce segmentation maps with high precision but poor recall due to large false negative predictions. Recently, Hashemi et al have proposed to use the generalized form of the Dice loss to improve the balance between precision and recall [32]. The Tversky Index (TI), also known as the F-score is depicted in Equation 2.5:

$$TI = \frac{2TP}{2TP + \alpha FP + \beta FN} \tag{2.5}$$

where α and β are hyperparameters used to control the contribution of FPs and FNs, respectively. Large values of β penalize the network more for false negative predictions boost overall model recall. In practice, α and β are complements of each other to preserve the normalized DSC score.

Following the strategy by Milletari et al, Hashemi and coauthors fashion the TI into the Tversky Loss (TL) function by minimizing its complement, as formulated in Equation 2.6.

$$TL = \frac{1}{C} \sum_{c} 1 - TI_c \tag{2.6}$$

The authors report a 3% boost in DSC score using the $TL(\beta = 0.7, \alpha = 0.2)$ against the standard Dice loss.

Focal Loss A more recent variant of the cross-entropy loss is the focal loss proposed by Lin et al which attempts to mitigate intra-class imbalance [33]. The authors reason that *easy* examples dominate the gradient while the gradient contributions from *hard* examples are small and get averaged out. The focal loss therefore incorporates a modulating parameter γ to down-weight easy examples and allow the network to focus more examples it would ordinarily overlook. This form of the cross-entropy builds upon the weighted-cross entropy as it aims to tackle class-imbalance *between* an ROI class for the object

detection task. The full form depicted in Equation 2.7:

$$FL = \sum_{c} \sum_{i} -w_c g_{ic} (1 - p_{ic})^{\gamma} \log(p_{ic}) \tag{2.7}$$

where w_c is the class weight, $(1 - p_{ic})^{\gamma}$ is the modulation factor where $\gamma >= 0$.

Limitations of current losses

The Dice loss and cross-entropy are widely utilized for segmentation in the medical community. The Dice loss provides improved spatial overlap for segmentation but small misclassifications can lead to a large decrease in DSC accuracy. The cross-entropy is robust to outliers however it treats all classes equally which requires substantial parameter tuning. Several works propose hybrid versions of both the DL and variants of the cross-entropy loss functions to get the benefits of both for class imbalance [27], [34]. However, with an unbounded cross-entropy and normalized Dice loss, elaborate cost function tuning is recquired to help the model converge. More recently, a hybrid focal Dice loss has been proposed by Wang et al in [35]. The authors acheive a 1% boost in overall DSC accuracy but lower, and sometimes equivalent, precision and recall than the conventional DL. We propose a generalized loss to address most limitations within segmentation loss functions in the medical community. Our cost function takes advantage of the spatial consistency of the DL while addressing the issue of class imbalance and predictions with poor sensitivity. Our loss function is an extension of the Tversky loss, motivated by [33] and is presented in Section 3.1.1.

Attention based architectures

Improving a network's focus on underrepresented classes can also be tackled through architectural modifications. Existing applications rely heavily on multi-stage cascaded CNNs such as the Dense-Net based FCN [36]. However, this approach leads to excessive and redundant use of computational resources and model parameters; for instance, similar low-level features are repeatedly extracted from all layers within the cascade increasing the model's memory requirement.

Attention within deep learning can be defined as a feature pruning tactic to propagate important features and suppress non-discriminative activations. Hard attention, for example through iterative region proposal and cropping [37] is not a differentiable operation and cannot be trained end-to-end. Contrarily, soft attention is probabilistic and can be learned through standard back-propagation. A notable attention mechanism termed the Squeeze and Excitation Networks (SEN), uses a channel-based attention scheme to learn scaling coefficients that highlight important features [38]. The SEN architecture, depicted in Figure 2.2 was a top performer in the ILSVRC 2017 Challenge, highlighting that architectural modifications rather than simply increasing layer depth can improve a model's predictive power. Residual squeeze blocks is used for organs-at-risks CT segmentation with improved performance [34].

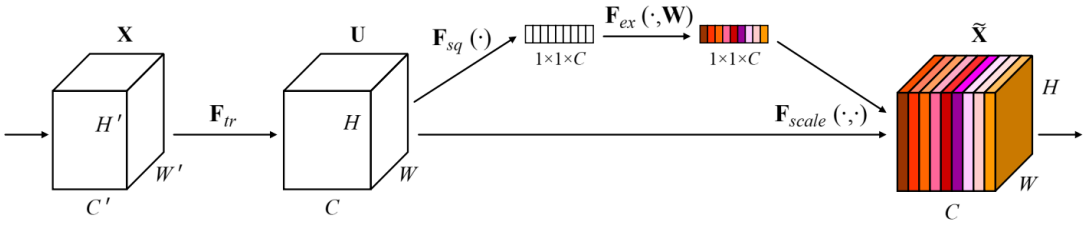


Figure 2.2: The notable Squeeze and Excite (SE) block used in Squeeze and Excite Networks [38]. Each feature block X undergoes a global transformation U using a convolutional operator F_{tr}. Global features are squeezed by the F_{sq} operation into a 1-dimensional tensor. Channel inter-dependencies are then modeled fully connected layers which are excited by non-linear activations, F_{ex}. The resulting F_{scale} coefficients weight subsequent channels using global information and, thereby, eases the learning process and enhances the representational power of the network.

.

The SE-block factors out spatial dependency by global average pooling to learn a channel specific descriptor. Another aspect of architectural attention is improving local spatial representations. Grid based attention uses information from coarser scales to guide feature propagation from more semantic level scales[39]. Oktay and collaborators use the first grid-based attention gating scheme for medical image segmentation [40]. They combine coarse spatial features with refined semantic level features and apply ReLU and sigmoid non-linearities to obtain a scalar attention matrix. Similarily, the work in [41] define a hypercolumn as a pixel-descriptor of all previous CNN activations above that

pixel. This representation guides fine-grained feature extraction for all layers below the pixel. In [42], authors combine a spatial based attention with the global SE block to improve feature recalibration. Quantitative results on CT segmentation show that spatial attention alone shows a larger improvement when compared to channel attention alone, while the combination acheives the best results [42].

Architectural based attention is a promising approach to improve feature representations when dealing with imbalanced classes. These tactics typically do not increase model complexity by much when compared to simply increasing layer depth. For instance, incorporating SE blocks in a network increases relative learnable parameters by roughly 4%. Moreover, pruning strategies can be tailored based on the severity of class imbalance in the dataset. In this thesis, we combat class imbalance architecturally by using grid based attention. Moreover, we supplement the model with additional multi-scale input features by using an image pyramid. Details on the model architecture are discussed in Section 3.1.2.

2.3 Multimodal fusion

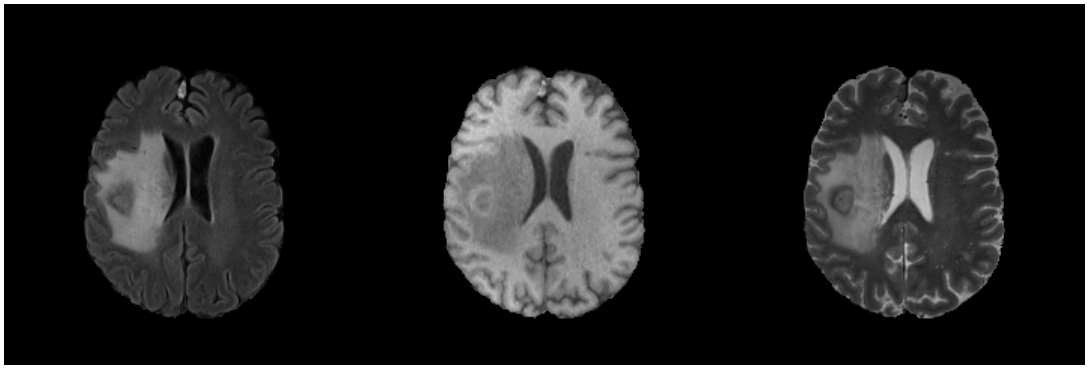
Multimodal fusion is the process of integrating multiple unimodal representations into one compact joint representation [43]. The philosophy behind using multimodal data stems from the fact that multiple sources of information can be exploited to make a better decision than simply using a single resource. In medical imaging, multiple modalities refer to images acquired from multiple scanners such as CT or MRI which produce different intensity responses to different tissue structures.[2] Moreover, within each modality, parameters such as exposure time and constrast can be adjusted to focus on specific anatomy and create new modalities. For example, MRI is widely used in neuroimaging studies to obtain a variety of complementary scans to assist radiologists in contouring. T1-weighted MR scans and Fluid-Attenuated In-version Recovery (FLAIR) sequences can both delineate basic brain anatomy but FLAIR can also detect white matter abnormalities associated with a variety of neuropathological conditions. Similarly, cerebral edema can be contoured from T2-weighted MRI but FLAIR is used to cross-check the extension and discriminate against healthy ventricular structures [4]. Figure 2.3 depicts the complementary information present within each MR modality and its utility in applications such as tumor delineation.

Automated methods to fuse information from multiple modalities is therefore of primary interest due to its ability to maximize the statistical power of each individual biomarker. In the context of deep learning based fusion, the key architectural challenge lies in *how* to fuse information and *where* in the pipeline to incorporate a fusion block. There are three such fusion pipelines that are commonly used in deep learning and are thus adopted by the medical community for the segmentation problem.

2.3.1 Early fusion

Early fusion methods create a joint representation of input features from multiple modalities, often by concatenation in the input image block [43], as depicted in Figure 2.4(a). The model architecture typically follows a simple encoder-decoder design and implicitly

[2] In the medical literature, fusion between different imaging modalities, such as between CT and MRI, is typically referred to as cross-modal sinces images do not contain pixel-to-pixel correspondances.



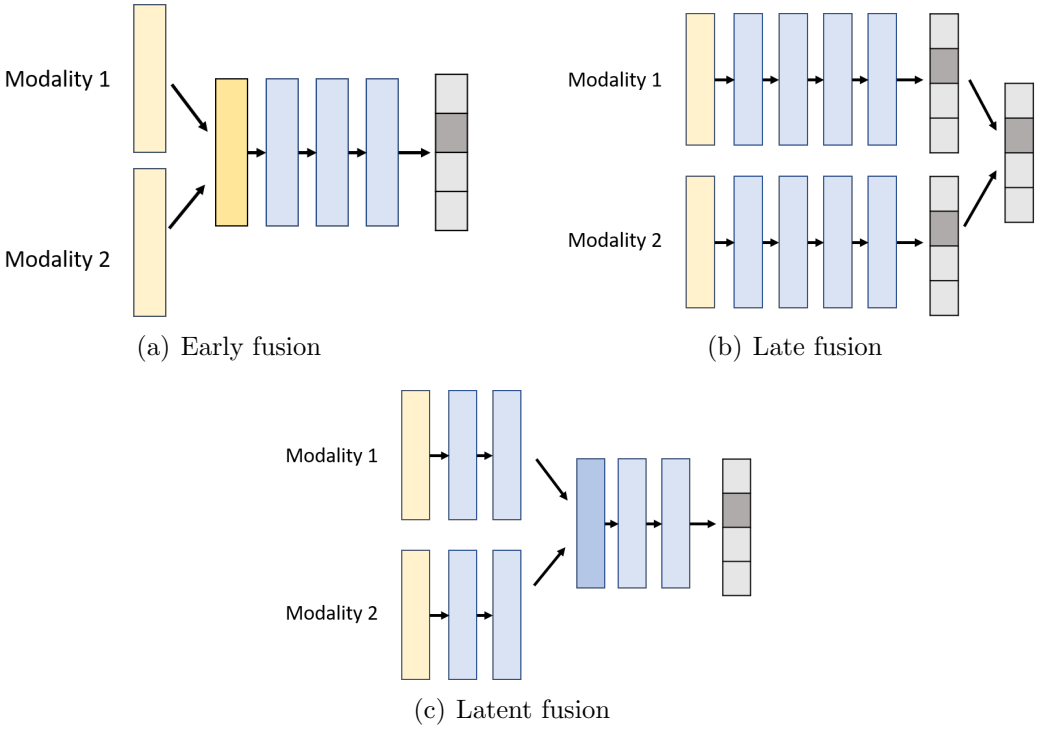
(a) FLAIR (b) T1 (c) T2

Figure 2.3: Mutli-parametric MR sequences of a patient with glioblastoma taken from the Brain Tumor Segmentation Challenge (BraTs 2018) dataset [4]. Each MR modality highlights different types of intra-tumoral structures such as the tumor core visible in T1 and the cerebral edema visible in both FLAIR and T2 sequences.

assumes a single network can capture the joint, task-specific semantics from the low-level image space. Moreover, since early fusion methods propagate information from the first layers, these fusion models are often susceptible to noise that affects a single modality. Despite these challenges, most works on multimodal segmentation use early fusion to aggregate joint representations under the assumption that deep networks can learn both modality invariant and noise invariant features [44], [45], [46].

2.3.2 Late fusion

Late fusion methods utilize different models for each modality and fuse predictions at the decision level by means of averaging, voting or a learned model [47]. This allows for flexibility in the case of missing modalities and is more robust to noise as each modality is processed independently. However, late-fusion operates at the decision level which means it is unable to model intra-modality interactions at the feature level.



(a) Early fusion

(b) Late fusion

(c) Latent fusion

Figure 2.4: Locality of the fusion operator is an important architectural decision. Typical feature fusion can occur at early in the pipeline, as depicted in (a) or at the decision level, depicted in (b). Latent fusion models, illustrated in (c), map each modality to a latent feature space where fusion operations such as addition or averaging are used to create a joint feature space.

2.3.3 Latent fusion

Recent efforts in multi-modal fusion operate in the latent space and employ a CNN to learn modality specific features. These latent features are combined into a single, joint representation and transformed into class segmentation maps using a learnt decoder. Latent space fusion is commonly carried out by concatenation of representative features as proposed in [48]. This method essentially allows the network to learn for itself the best combination of modality-specific features. Despite its effectiveness when used with a large enough parameter space, concatenation is an inefficient operation as it increases the dimensionality of the feature space leading to increased model complexity. Moreover, this approach does not scale well when there are multiple modalities involved. Since the latent space is essentially an embedding of the image space, arithmetic operations such

as addition or averaging are well-defined and have semantic meaning [49].

To this end, Van et al propose average fusion of N latent representations based on the argument that each modality specific network should output similar representations. The fused representation Z_{fused} is depicted below as:

$$Z^f_{fused} = \frac{1}{N}\sum_N (Z^f_1, ..., Z^f_N) \tag{2.8}$$

where Z^f_N represents the f feature plane for modality N.

In a highly complex space, the modality specific features should embody *highly correlated* representations. However, the differences between MR modalities depicted in Figure 2.3 will be reflected in their latent space projections. Therefore, the average representation may not account for the outlier features which contribute rich information to the segmentation task.

In [50], Chartsias and coauthors propose to use the maximum activation of each modality plane as the fused representation, as depicted in Equation 2.9.

$$Z^f_{fused} = max(Z^f_1, ..., Z^f_N) \tag{2.9}$$

Selecting the pixel-wise maximum activation between latent responses guarantees that outliers are not suppressed by averaging out feature planes and can have an effect on the class decision. The challenge now is that the maximum activation may be reflective of an outlier feature or a noisy response. This allows for outlier features to be propagated forward which may deter the network from its segmentation task.

A more robust latent fusion model, termed Hetero-modal Image Segmentation (HeMIS) fusion was formulated by Havaei et al in [49]. The fused representation is the concatenation of the first and second moments of each modality's feature planes:

$$Z^f_{fused} = concat(Z^f_{mean}, Z^f_{var}) \tag{2.10}$$

$$where: Z^f_{mean} = \frac{1}{N}\sum_{N}(Z^f_n, ..., Z^f_N) \tag{2.11}$$

$$Z^f_{var} = \frac{1}{N-1}\sum_{n\in N}(Z^f_n - Z^f_{mean})^2 \tag{2.12}$$

The base assumption with HeMIS fusion is each modality is independent of each other and so feature-wise variances can be computed without modelling conditional covariances. Using deep networks and even so with HeMIS-like fusion, the network still cannot explicitly rely on more informative outputs because it can potentially learn to ignore the covariance features and focus on decoding the average representations [50].

2.3.4 Improvements to Multimodal latent fusion

The optimal fusion pipeline will be robust to missing modalities and find an effective way to combine modality specific information into a joint feature representation. Latent fusion models are able to create a modality invariant latent space and apply an operation or a sequence of non-linearities to create a joint feature. One limitation is the location of the fusion block. In the most latent layer, features are highly abstract representations of the image space but they have reduced locality. Creating fused representations at multiple scales would benefit the network's ability to both learn and use multi-scale joint representations. The very recently proposd MMFNet adopts this scheme [51]. Moreover, within the modality invariant space, dominant modalities have the ability to take over and force the task-specific decoder to learn how to operate with its features as opposed to the joint representation.

We propose to build upon HeMIS fusion by importance weighting the mean and variance of the modality specific features. Moreover, we incorporate our fusion block at every scale to use multi-scale modality specific information. Our model architecture is detailed in section 4.1.1.

Chapter 3

Class Imbalance

3.1 Technical Approach

3.1.1 Focal Tversky loss

The cross-entropy and Dice loss work well in balanced dataset scenarios while the Tversky loss can be tuned to perform well in unbalanced cases. However, all loss functions are plagued by gross variations within the ROI class itself. To this end, we propose the Focal Tversky loss (FTL) parametrized by γ, for control between *easy* and *hard* training examples. In [33], the focal parameter exponentiates the cross-entropy loss to focus on hard classes detected with lower probability for the object detection task. Motivated by their contribution, we exponentiate the TI to focus on hard examples within the ROI class and balance precision and recall better. In this context, the FTL improves both the inter-class balance, caused by imbalance between the background and foreground classes and the intra-class imbalance caused by varying sizes of the foreground, ROI class.

The focalized Tversky Index (FTI) is defined as follows:

$$FTI = (\frac{2TP}{2TP + \alpha FN + \beta FP})^{1/\gamma} \quad = TI^{1/\gamma} \tag{3.1}$$

Therefore, training with the FTL is defined as the minimization of the of the focalized Tversky Index, summed over all the classes c:

$$FTL = \frac{1}{C}\sum_{c}(1 - TI_c)^{1/\gamma} \tag{3.2}$$

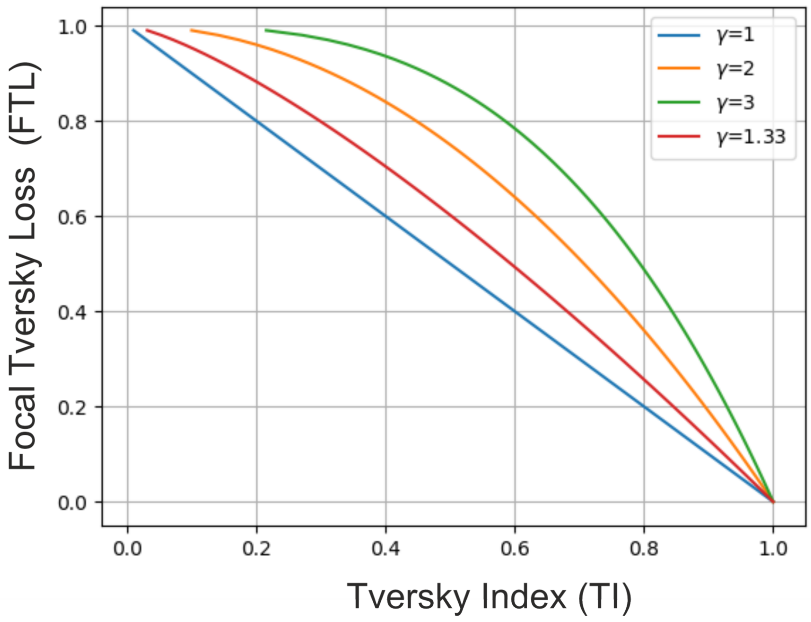


Figure 3.1: The focal Tversky loss non-linearly focuses training on hard examples ($TI < 0.5$). With increasing values of γ,the FTL increases the network's focus on examples classified poorly.

The behavior of the FTL as a function of the Tversky Index is depicted in Figure 3.1. With increasing values of γ, the network is non-linearly forced to focus more on examples that are misclassified with low TI scores (ie $TI < 0.5$). This phenomenon is depicted in the curvature of each FTL plot in Figure 3.1. For example, a pixel classified with $TI = 0.7$ will result in a $TL = 0.3$ while the $FTL(\gamma = 1.33) = 0.405$. The FTL proves to be beneficial to small ROIs as depicted in ablation studies in Section 3.3 however its exponential behavior can sometimes lead to instability. To combat this, we propose to use the FTL with deep supervision where multiple loss functions can average out the parabolic effects.

The proposed FTL is a generalized version of the commonly used Dice loss. When $\gamma = 1$, the FTL simplifies to the TL and when $\alpha = \beta = 0.5$, the TL simplifies to the DL. In this manner, our focal Tversky loss function is able to take advantage of structure that is a characteristic of overlap-based loss functions while at the same time, penalize poor predictions without the need to tune class weighting parameters. In all our works, $\beta = 1 - \alpha$ therefore, the proposed FTL only requires two hyperparameters to be tuned,α and γ. From our empirical evaluations, $\alpha \in [0.6 - 0.7]$ and $\gamma \in [1.1 - 1.4]$ produce results with improved precision and recall curves than when using the conventional Dice loss.

3.1.2 Network Architecture

We utilize an Attention U-Net with the addition of an input image pyramid inspired by conventional image pyramids. The network architecture is depicted in Figure 3.3.

Attention gates

Attention gates (AG) produce attention coefficients $\alpha_i \in [0, 1]$ at each pixel i. These coefficients scale input feature maps x_i^l, at layer l, to output semantically relevant features, $\hat{x_i}^l$, as depicted in Figure 3.2. A gating signal, g, is used for each pixel i to determine focus regions. It is collected from a coarser scale than the input query signal, x_i^l to compute intermediate activation maps:

$$q_{attn}^l = \psi^T(\sigma_1(W_x^T x_i^l + W_g^T g_i + b_g)) + b_\psi \tag{3.3}$$

where the linear attention coefficients, q_{attn}^l, are computed by the element-wise sum and 1x1 linear transformations, parameterized by W_x, b_x, W_g and b_g. The intermediate maps are transformed by ReLU and sigmoid non-linearities applied as σ_1 and σ_2, respectively:

$$\alpha_i^l = \sigma_2(q_{attn}^l(x_i^l, g_i)) \tag{3.4}$$

The attention coefficients α_i scale the low level query signal x_i^l by an element-wise product and retain only relevant activations. These pruned features are then concatenated with

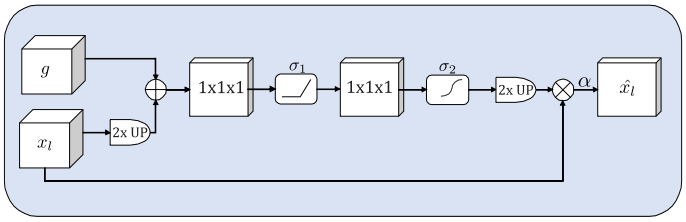


Figure 3.2: Schematic of additive attention gate (AG) adapted from [40]. Input features x_l are scaled with attention coefficients α_i to propagate relevant features to the decoding layer output $\hat{x}_l$. The coarser gating signal g provides contextual information while spatial regions from the input x_l provide locality information. Feature map resampling is computed by bilinear interpolation.

upsampled output maps at each scale in the expansive stage. The lowest-level feature maps, i.e. the first skip connections, are not used in the gating function as they do not represent input data in a high dimensional space [40]. A 1x1x1 convolution and sigmoid activation is applied on each output map in the expansive stage.

Input pyramid

Some class details are more easily accessible at different scales. Motivated by the success of image pyramids [52] and recent interpretations such as PSP-Net [53], we inject the encoder layers with a down-scaled input image before each of the max-pooling layers. Each input image is sub-sampled using average pooling and fed through two Conv-ReLU-BatchNorm2d blocks. We combine the input image features through concatenation with the feature representations from previous layers. This method enforces spatial priors into each convolutional layer combined with cascaded feature maps to maximize the representation of each image. The image pyramid offers a convenient, multi-resolution set of features that prove to be useful when the ROI pixel space is severely imbalanced.

3.1.3 Training strategy

The hyperbolic nature of the FTL results in instability during training as the model approaches convergence. This phenonmenon is visbile in the focal Dice work [35] where training is depicted to be very noisy and spurious. To combat this, we employ deep

supervision at multiple intermediate layers to average out the contributions from the FTL. Moreover, we train the last layer with the TL to retain true gradients to the segmentation task. In addition to stable training, deep supervision improves the segmentation accuracy for datasets where small ROI features can get lost in cascading convolutions and helps to ensure that attention unit has the ability to influence the responses to a large range of image foreground content.

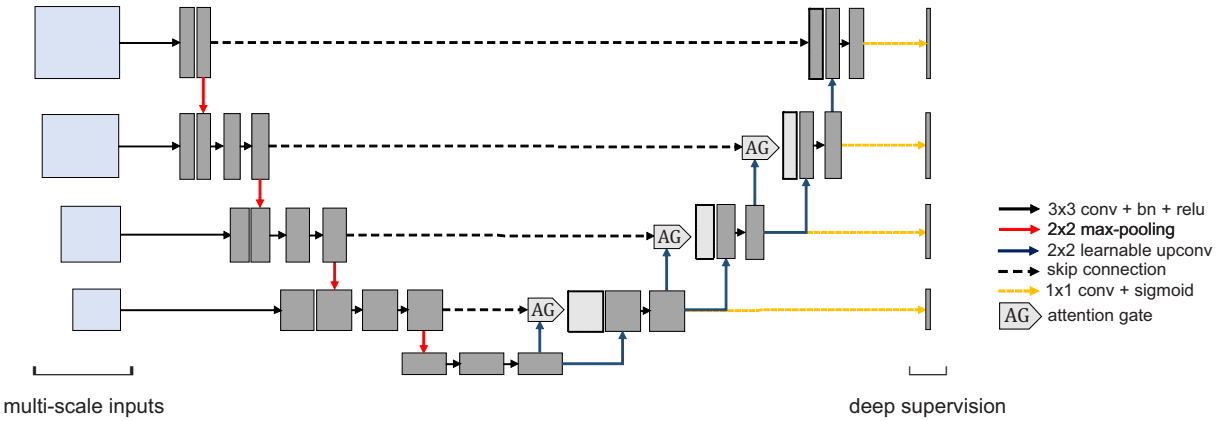


Figure 3.3: Proposed Attention U-Net architecture with the addition of an input image pyramid. The model is deeply supervised with the FTL at every layer except the final layer which is trained with the TL.

3.2 Experiments

3.2.1 Datasets

We validate the FTL on two datasets where the ROI class is significantly smaller than the background class and achieve large performance gains.

We experiment with the Breast Ultrasound Dataset B (BUS) open-sourced in [1]. This dataset consists of 163 ultrasound images of breast lesions from different women. The average image size is 760 x 570 pixels where each of the images presented one or more lesions. Example lesions are depicted in Figure 3.4 with their corresponding ground truth segmentation masks. It is evident from Figure 3.4 that the variability in segmentation contours presents a challenge for generalized learning. Moreover, the speckle noise from the ultrasound acquisition process makes it hard to delineate boundaries and focus on the ROI class. For our experiments, this dataset is resampled to 128 x 128 pixels with a 75-25 train-test split.

To extend our proposed method to larger datasets, we extract training data from the Skin Lesion Analysis Towards Melanoma Detection Challenge collected by International Skin Imaging Collaboration (ISIC 2018), [2], [3]. This dataset consists of 2,594 RGB images of skin lesions with an average image size of 2166 x 3188 pixels. Examples from the ISIC 2018 are presented in Figure 3.5. Contrary to the BUS dataset, the RGB color space allows for improved contrast between background and foreground. However, the variability between ground truth annotations presents a challenge as some images depict sharp boundaries while others have been contoured with less detail. For our experiments, the ISIC 2018 dataset is resampled to 192 x 256 pixels with 75-25 train-test split.

To present a fair evaluation of our multi-scaled attention U-Net supervised with the focal Tversky loss, we do not augment our datasets or incorporate any transfer learning. We study 7 cases of variations within U-Net and the Tversky loss function while comparing to the baseline U-Net trained with Dice loss. Ablation test results are recorded in Section 3.3 where each experiment is averaged 5 times with a random test fold each time. We present results for Dice scores, precision and recall.

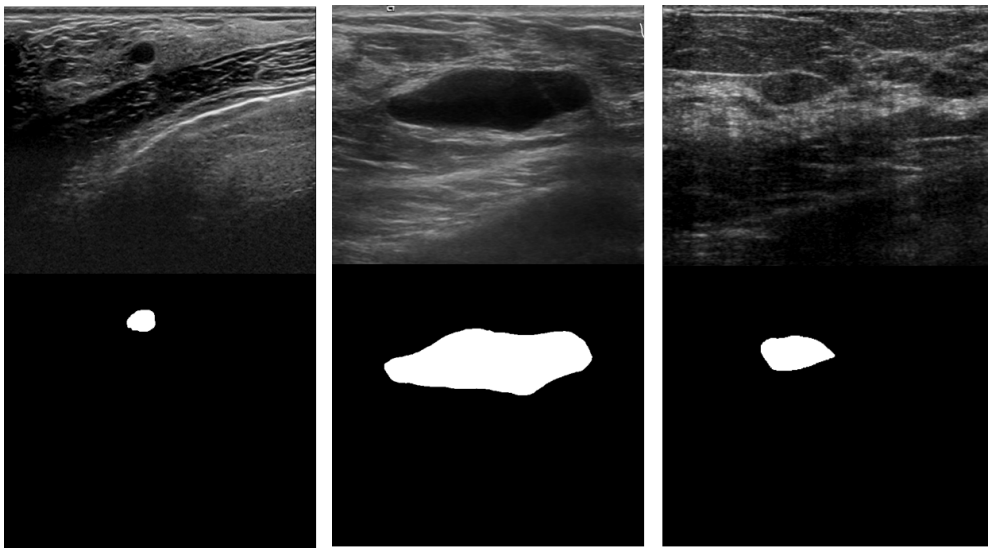


Figure 3.4: The BUS dataset B examples with corresponding ground truth. The lesion class in the BUS dataset B occupies on an average 5.43% ± 4.84% pixels when compared to the background. Moreover, the dataset contains speckle noise which makes it hard to delineate the lesion contours from the background class and image noise. The dataset is also very small, containing only 163 images in total.

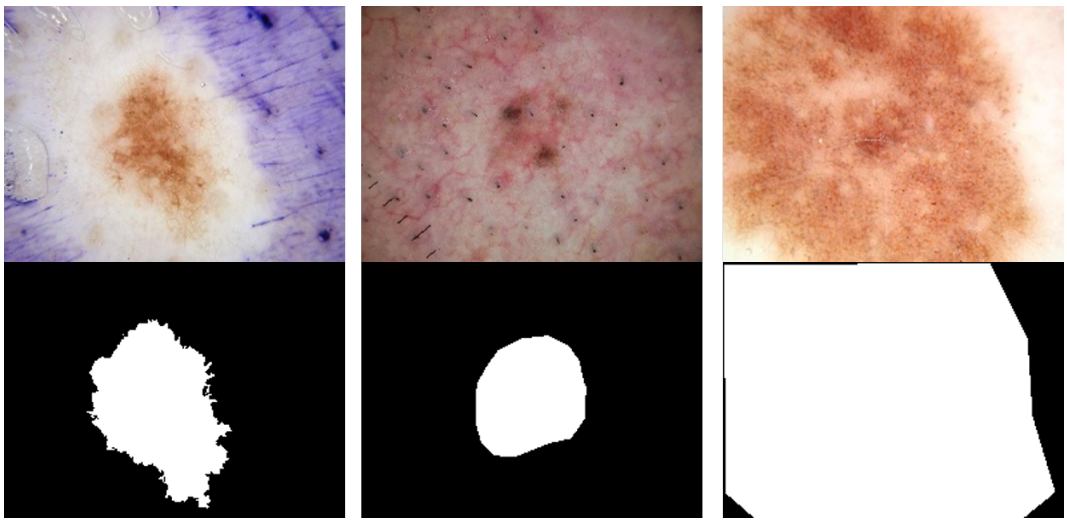


Figure 3.5: The ISIC 2018 dataset and corresponding ground truth. The ISIC 2018 dataset contains 2,594 high resolution RGB images. The skin lesions vary largely in size, roughly occupying 21.4% ± 20.3% of every image in the dataset. The variability between ground truth annotations presents a challenge as some images depict sharp boundaries (first column) while others have been contoured with less detail (second and third column)

.

3.3 Results

Table 3.1: Performance on BUS 2017 Dataset B with 40 test images

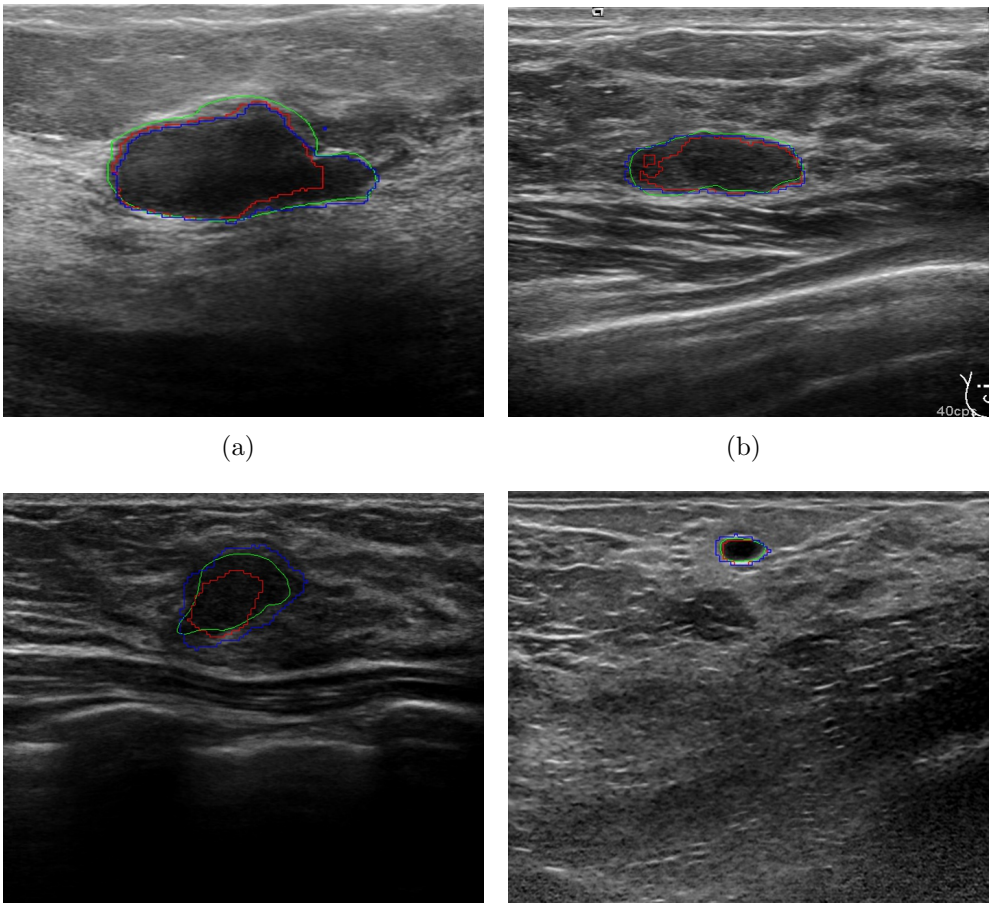
Model	DSC	Precision	Recall
U-Net + DL	0.547 ± 0.04	0.653 ± 0.171	0.658 ± 0.146
U-Net + TL	0.657 ± 0.02	0.732 ± 0.072	0.723 ± 0.074
U-Net + FTL	0.669 ± 0.033	0.775 ± 0.047	0.715 ± 0.057
Attn U-Net + DL	0.615 ± 0.020	0.675 ± 0.042	0.658 ± 0.049
Attn U-Net + Multi-Input + DL	0.716 ± 0.041	0.759 ± 0.092	0.751 ± 0.046
Attn U-Net + Multi-Input + TL	0.751 ± 0.042	0.802 ± 0.073	0.768 ± 0.056
Attn U-Net + Multi-Input + FTL	**0.804 ± 0.024**	**0.829 ± 0.027**	**0.817 ± 0.022**

Table 3.2: Performance on ISIC 2018 dataset with 649 test images

Model	DSC	Precision	Recall
U-Net + DL	0.820 ± 0.013	0.849 ± 0.038	0.867 ± 0.048
U-Net + TL	0.838 ± 0.026	0.822 ± 0.051	0.917 ± 0.033
U-Net + FTL	0.829 ± 0.027	0.797 ± 0.040	**0.926 ± 0.012**
Attn U-Net + DL	0.806 ± 0.033	0.874 ± 0.080	0.827 ± 0.055
Attn U-Net + Multi-Input + DL	0.827 ± 0.055	**0.896 ± 0.019**	0.829 ± 0.076
Attn U-Net + Multi-Input + TL	0.841 ± 0.012	0.823 ± 0.038	0.912 ± 0.026
Attn U-Net + Multi-Input + FTL	**0.856 ± 0.007**	0.858 ± 0.020	0.897 ± 0.014

Table 3.1 shows that the baseline U-Net trained with the Dice loss function has the worst performance. The large standard deviation in the precision and recall scores suggest the learning is not stable. In contrast, U-Net models trained with TL and FTL show increased DSC and more balanced precision-recall scores which occurs due to weighting α higher in the loss function than β. We observe incorporating attention in U-Net trained with DL depicts lower Dice scores than the baseline, probably due to the intra-lesion variation. Injecting an input pyramid into the model improves the DSC significantly suggesting features of small lesions are easily lost when class imbalance is high. Training the attention model with FTL combines the benefits of improved feature selection with focused training to outperform all other methods. The proposed architecture (last row) is able to segment lesions with a Dice score of 0.804 on training with a small subset of 100 images.

Contrary to the BUS scores, ISIC 2018 results in Table 3.2 show the baseline U-Net trained with DL performs well due to the large training sample size, variation in lesion structures and distinct features present in the RGB images. Training U-Net with TL and FTL, we observe an improved DSC score. However, when the Tversky index is high for misclassified examples, the focal exponent γ suppresses the contribution to the error signal and since α is weighted higher than β, the model converges to the highest reported recall at 0.926, but lowest precision. To address this issue, when training the proposed attention model, we supervise the last layer with TL so that a true error signal will still propagate back when the model is close to convergence. As a result, our improved attention U-Net model with FTL (last row) obtains slightly lower but overall better balanced recall and precision and consequently, the best DSC score. We outperform the baseline by 3.6% with a low spread of 0.7%.



(a) (b)

(c) (d)

Figure 3.6: Example segmentation contours from the BUS dataset. Green contours reflect the ground truth annotation and red contours reflect the standard U-Net model trained with the Dice loss. The blue contours depicts our proposed model, improved attention U-Net trained with FTL. In (a) and (b), the improvement in our method is evident as the contour matches the ground truth boundary much closer. However, as the FTL encourages more false positive predictions when compared to false negatives, we sometimes observe over-segmentation as depicted in (c) and (d). In both these example instances, the DSC is higher than the red contour but at the cost of specificity. Since the ratio of false positives to false negatives is not very high, we are still able to achieve a good balance, as depicted in the last row of Table 3.1.

Chapter 4

Multimodal Fusion

4.1 Technical Approach

4.1.1 Fusion Architecture

Early fusion and late fusion are classical approaches to the multi-modal problem. Recent fusion architectures aim to model inter-modality features using latent space models. In our work, we exploit the inter-modality dynamics by creating fusion blocks at every convolutional scale.

We follow a similar approach to most latent fusion models and use a modality specific encoder and one common decoder for the segmentation task. Each encoder has four convolutional blocks which each encompass a Conv2d-BatchNorm-ReLU operation and are followed by a 2x2 max-pooling. We use the notation $C_k^{(s)}$ to describe the feature map at each scale s prior to the max-pool operation. The modality invariant decoder uses four transpose convolution layers followed by a final 1x1 conv layer to upsample joint representations. Using the same encoder structure for each modality ensures that at every scale, each modality's features have the same receptive field and so operations such as addition and averaging have semantic meaning. Subsequently, we compute the HeMIS feature representation at each intermediate scale as opposed to just in the latent space.

Let $C_{k,l}^{(s)}$ represent the convolutional feature map at an intermediate scale s, for modality $k \in [0, 1, ..., K]$, at layer l. At every scale, the joint multimodal representation C_l^s is computed using HeMIS mean and variance features:

$$E_l[C_l^{(s)}] = \frac{1}{K} \sum_{k \in K} C_{k,l}^{(s)} \tag{4.1}$$

$$Var_l[C_l^{(s)}] = \frac{1}{K-1} \sum_{k \in K} (C_{k,l}^{(s)} - E_l[C_l^{(s)}])^2 \tag{4.2}$$

Higher level image moments can be calculated using non-linear combinations of the the mean and variance. However, this requires a complex decoder or a prior on the feature space to encourage the network to learn useful representations. To mitigate the need for increased layer complexity, we use a Moment Gated Fusion (MGF) feature recalibration block with HeMIS features to aid the network at selecting useful representations.

Moment Gated Fusion block

As previously discussed in [50] and [54], the average multimodal feature can capture significant structural information. The variance features on the other hand are a representation of disagreement between modality specific features. We argue that they are still important attributes as they account for possible fringe cases that might be missed if only the average feature was used. We are motivated by the fact that different image modalities have varying intensity responses to certain tissues, as depicted in Figure 2.3 and will on occasion deviate from the mean multimodal representation. Therefore, we propose to recalibrate the variance features using an SE block in the channel dimension. A squeeze and excite operation will allow the network to learn relevance to the outlier features and learn to ignore or pay more attention to them based on the error from the segmentation task.

A naive approach would be to apply the SE blocks to all HeMIS features. However, by design, the SE block factors out any spatial dependency by using global average pooling to learn a channel specific descriptor. Instead, we introduce a hybrid channel-wise and spatial-wise re-weighting block termed the MGF. Image moments computed from the

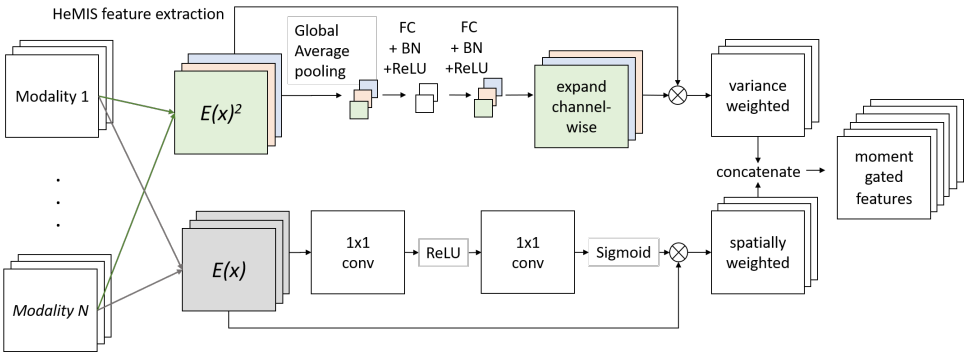


Figure 4.1: Overview of our proposed MGF block. The upper portion of the MGF pipeline excites variance ($E^2(x)$) features in order to capture possible outlier representations that are important for segmentation. The lower portion guides the mean ($E(x)$) feature extraction by following a similar spatial squeeze operation. By applying non-linear activations independently to feature moments, we discourage the network from being biased to only the mean features and therefore can account for outlier features in the fused representations.

HeMIS abstraction layer are gated using non-linear excitation operations. Depicted in Figure 4.1, the variance features are independently gated using the squeeze and excitation pipeline. The non-linear activations include a global average pooling to reduce maps into a single descriptor which is then squeezed and excited using a multilayer perceptron and ReLU activations. The mean features are spatially excited using a similar workflow as the channel-wise SE blocks. As depicted in Figure 4.1, this workflow includes a reduction using a 1x1 Conv2d block, followed by a ReLU, 1x1 Conv2d and a Sigmoid activation operation. Independently processing both spatial and channel-wise activations will encourage the network to learn relevance to structural features but also relevance to outlier features captured in the variance maps.

Utilizing the mean multimodal features to gate subsequent representations allows our network to scale well when compared to the commonly used concatenation based fusion. In Figure 4.2, we present our proposed architecture which scales to N modalities as the modality fusion occurs at the intermediate abstraction layer.

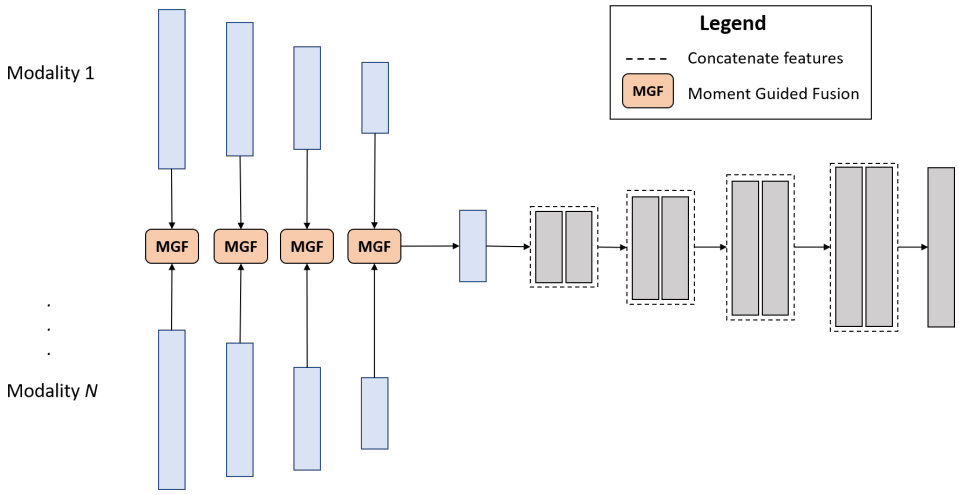


Figure 4.2: Overview of our proposed Multimodal Fusion Network. Each blue encoder block represents a Conv2d-BatchNorm-ReLU operation while each gray decoder block represents a transpose Conv2dTranspose-BatchNorm-ReLU operation. The last gray block utilizes a Sigmoid activation across every channel. MGF blocks are used to extract and prune multimodal feature representations based on HeMIS features abstraction.

4.1.2 Loss function

We utilize our generalized, multi-class FTL to supervise the multomodal attention network, reiterated in Equation 4.3 below:

$$FTL = \frac{1}{C}\sum_{c}(1 - TI_c)^{1/\gamma} \tag{4.3}$$

The hyperbolic nature of the FTL encourages the network early on to penalize false negative predictions. However, when close to convergence, the network is sometimes unstable. We improve the performance of FTL by incorporating a decay schedule for the exponent $1/gamma$. In our work, we decay $1/gamma$ in steps of 0.05 until $1/gamma = 1$. We empirically find that the network is close to convergence around 25 epochs and accordingly decay gamma step-wise every 5 epochs.

4.2 Experiments

We validate our multimodal fusion network on the BraTs 2018 dataset [4]. This dataset consists of 220 patients with co-registered T1, T1c, T2 and FLAIR MRI scans. Each patient has a volume of 240x240x155 however we discard the first and last 15 blank slices and operate our 2D model at the slice level. The BraTs 2018 ground truth consist of three classes: non-enhancing/necrotic tissue, peritumoral edema and enhancing structures [4]. A combination of each of these classes forms a sub-category for the evaluation board. Enhancing Tumor (ET) denotes the enhancing class structures, Tumor Core (TC) constitutes the nonenhancing structures and edema classes and the Whole Tumor (WT) refers too all classes in the annotation. We adopt the strategy in [46] and train the network to predict sub-regions ET, TC and WT.

We create a train, validation and test dataset based on a 70-20-10 split. We train all fusion models with Stochastic Gradient Descent (SGD); local fusion models used learning rate of 0.0001 while global fusion models used a larger learning rate of 0.001 as convergence was an issue with smaller learning steps. To allow for a fair evaluation between fusion practices, each model trains for a maximum of 300 epochs with an early stopping policy that ensures the validation loss cannot stagnate past a patience of 10 epochs. Each model was trained three times and average DSC, precision and recall scores are reported in Section 4.3.

4.3 Results

Below we outline results from common latent fusion strategies. The networks were trained with the DL.

We observe that concatenation of features results in the highest DSC score when compared to using the mean or variance features as the multimodal representation. However concatenation inefficiently scales the latent space by number of modalities and greatly increases the model complexity. Moreover, we observe applying the squeeze block did not have any benefit on the model's predictive power probably due to the simplistic decoder. Compared to HeMIS fusion and the squeeze HeMIS versions, our proposed MGF block in

Table 4.1: Performance of latent space fusion on BraTs 2018 dataset of 22 patients

Latent Models	DSC	Precision	Recall	Params
concat	**0.8301 ± 0.0055**	0.8927 ± 0.0128	0.8123 ± 0.0128	2.130 M
concat + SE blocks	0.8250 ± 0.0060	0.8869 ± 0.0073	0.8146 ± 0.0123	2.163 M
mean	0.8167 ± 0.0082	0.8630 ± 0.0147	**0.8178 ± 0.0021**	1.312 M
variance	0.8002 ± 0.0060	0.8500 ± 0.0158	0.8067 ± 0.0091	1.312 M
HeMIS	0.8105 ± 0.0038	0.8709 ± 0.0110	0.8005 ± 0.0122	1.475 M
HeMIS + SE blocks	0.8100 ± 0.0104	0.8569 ± 0.0196	0.8127 ± 0.0103	1.478 M
MGF	0.8251 ± 0.0022	**0.8928 ± 0.0120**	0.8020 ± 0.0111	1.475 M

the latent space produces competitive results with similar parameter complexity. When compared to concatenation latent model, our proposed MGF block only results in a 0.5% reduction in DSC score while maintaining competitive precision, but reduces the number of parameters to 1.475M from 2.13M (last column of Table 4.1). This implies that the MGF block can achieve similar performance utilizing a much more efficient network.

Since the BraTs 2018 dataset contains MRI volumes that are co-registered, all feature maps have the same receptive field. Therefore, we extend our latent MGF model to fuse modality specific features at every feature scale. We compare our network with the conventional U-Net that is most commonly employed for BraTs 2018 segmentation. Since U-Net with early fusion essentially learns non-linear combinations of all modalities from all scales, we term it a global fusion model. From Table 4.2, we observe our proposed multimodal fusion architecture is able to achieve competitive results when compared to its U-Net counter part with roughly 50% fewer model parameters. Using the FTL, the proposed results (last row of Table 4.2) depicts balanced precision and recall scores. Through cross validation, the best FTL parameters were found to be: $alpha = 0.7$, $beta = 0.3$, $^1/_\gamma = 0.7$.

Table 4.2: Performance of global fusion on BraTs 2018 dataset of 22 patients

Global Models	DSC	Precision	Recall	Params
U-Net + DL	0.8349 ± 0.0016	**0.9010 ± 0.0129**	0.8242 ± 0.0055	9.335 M
MGF Net + FTL	**0.8408 ± 0.0068**	0.8529 ± 0.0271	**0.859 ± 0.0154**	4.449 M

Chapter 5

Conclusion

5.1 Thesis Summary

In this thesis, we have tackled the class imbalance problem for lesion segmentation and the multimodal fusion problem in glioma segmentation.

Our first work presented a novel improvement to the gold standard Dice loss function for medical image segmentation. By exponentiating the Tversky index, we achieve a loss function that focuses more on mis-classified examples. Through experimentation, we justify the use of a parabolic loss function by incorporating a decay schedule for $1/gamma$ such that a network is initialized with our proposed focal Tversky loss and eventually decays to the Tversky index. In our earlier work, we were also able to mitigate any parabolic instabilities by incorporating deep supervision which had an averaging effect on any large gradient swings. We improve upon lesion segmentation by incorporating a conventional image pyramid to force the network to retain its attention to small image features where certain ROIs are more visible. Across two datasets with significantly small ROIs and large variance in annotations, we achieve 25.7% and 3.6% performance boost over the conventional U-Net, respectively.

Our second work explores commonly employed latent space fusion models for multimodal segmentation. We build upon HeMIS feature extraction by selectively pruning the mean and variance features independently. We propose the Multimodal Fusion block to weight

variance features in the channel dimension and utilize a spatial version of SE blocks to weight structural features. We train our proposed model with the multi-class version of the FTL and observe 1% in improvement over the conventional U-Net with early fusion with 50% fewer parameters for theBraTs 2018 segmentation problem.

5.2 Future Work

The most direct extension of both our works is to expand model architectures to incorporate space information either with 3D models or temporal structures such as convolutional RNNs. From the literature on the BraTs 2018 challenges, multimodal segmentation with co-registered images is very commonly solved using early fusion. Our explorations in 2D demonstrate that feature pruning can attain similar if not better results with almost half the parameters. Future research can focus on pruning temporal features independently from spatial and channel-wise features. This has the potential to mimic a radiologists action's when contouring data as every anatomical plane is typically analyzed independently. Moreover, combining attention pipelines with image fusion practices might encourage joint, task specific modality mixing at various scales which has potential to improve segmentation performance.

On a more general note, an interesting direction would be to incorporate generative modelling into image segmentation pipelines. In both the unimodal and multimodal setting, one has access to several information resources (scale, texture, modality-specific etc.) which one could use to construct a generative latent space. Sampling from a common manifold would improve a model's generalizability to varying ROI shapes and locations. Such a network would have a large impact on the adoption of deep networks in clinical use because generative models such as the Variational Autoencoder (VAE) for example are explained by their factors of variation. The authors in [46] had success in using the VAE to encourage a generalized encoder for segmentation, however the latent representations were not strongly encouraged to follow the Gaussian prior and were also not decoded for segmentation. This suggests a stronger image prior is needed for latent variable models, especially in the medical realm as the variations in ROI structure and locality are quite large. A good starting point might be to adopt feature pruning via our

proposed attention or recalibration strategies to reduce noisy features from contributing to a latent manifold.

References

[1] M. H. Yap, G. Pons, J. Martí, S. Ganau, M. Sentís, R. Zwiggelaar, A. K. Davison, and R. Martí, "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE journal of biomedical and health informatics*, vol. 22, no. 4, pp. 1218–1226, 2018.

[2] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*. IEEE, 2018, pp. 168–172.

[3] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions," *arXiv preprint arXiv:1803.10417*, 2018.

[4] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[5] E. Smistad, T. L. Falch, M. Bozorgi, A. C. Elster, and F. Lindseth, "Medical image segmentation on gpus–a comprehensive review," *Medical image analysis*, vol. 20, no. 1, pp. 1–18, 2015.

[6] A. Norouzi, M. S. M. Rahim, A. Altameem, T. Saba, A. E. Rad, A. Rehman, and M. Uddin, "Medical image segmentation methods, algorithms, and applications,"

IETE Technical Review, vol. 31, no. 3, pp. 199–213, 2014.

[7] J. Du, W. Li, K. Lu, and B. Xiao, "An overview of multi-modal medical image fusion," *Neurocomputing*, vol. 215, pp. 3–20, 2016.

[8] S. Rizzo, F. Botta, S. Raimondi, D. Origgi, C. Fanciullo, A. G. Morganti, and M. Bellomi, "Radiomics: the facts and the challenges of image analysis," *European radiology experimental*, vol. 2, no. 1, p. 36, 2018.

[9] Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li, "Medical image segmentation based on multi-modal convolutional neural network: Study on image fusion schemes," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 903–907.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu,

"Ce-net: Context encoder network for 2d medical image segmentation," *IEEE transactions on medical imaging*, 2019.

[18] S. Apostolopoulos, S. De Zanet, C. Ciller, S. Wolf, and R. Sznitman, "Pathological oct retinal layer segmentation using branch residual u-shape networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 294–301.

[19] E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson, and D. C. Barratt, "Automatic multi-organ segmentation on abdominal ct with dense v-networks," *IEEE transactions on medical imaging*, vol. 37, no. 8, pp. 1822–1834, 2018.

[20] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. Ayed, "Hyperdense-net: A hyper-densely connected cnn for multi-modal image segmentation," *IEEE transactions on medical imaging*, 2018.

[21] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural networks : the official journal of the International Neural Network Society*, vol. 106, pp. 249–259, 2018.

[22] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based data augmentation for improved liver lesion classification," 2018.

[23] A. Madani, M. Moradi, A. Karargyris, and T. Syeda-Mahmood, "Chest x-ray generation and data augmentation for cardiovascular abnormality classification," in *Medical Imaging 2018: Image Processing*, vol. 10574. International Society for Optics and Photonics, 2018, p. 105741M.

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[25] S. Valverde, M. Cabezas, E. Roura, S. González-Villà, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira, A. Oliver, and X. Lladó, "Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach," *NeuroImage*, vol. 155, pp. 159–168, 2017.

[26] M. Kubat, S. Matwin *et al.*, "Addressing the curse of imbalanced training sets:

one-sided selection," in *Icml*, vol. 97. Nashville, USA, 1997, pp. 179–186.

[27] K. C. Wong, M. Moradi, H. Tang, and T. Syeda-Mahmood, "3d segmentation with exponential logarithmic loss for highly unbalanced object sizes," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 612–619.

[28] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," *Medical image analysis*, vol. 36, pp. 61–78, 2017.

[29] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Auto-context convolutional neural network (auto-net) for brain extraction in magnetic resonance imaging," *IEEE transactions on medical imaging*, vol. 36, no. 11, pp. 2319–2330, 2017.

[30] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in mri images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1240–1251, 2016.

[31] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE, 2016, pp. 565–571.

[32] S. R. Hashemi, S. S. M. Salehi, D. Erdogmus, S. P. Prabhu, S. K. Warfield, and A. Gholipour, "Asymmetric similarity loss function to balance precision and recall in highly unbalanced deep medical image segmentation," 2018.

[33] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[34] W. Zhu, Y. Huang, H. Tang, Z. Qian, N. Du, W. Fan, and X. Xie, "Anatomynet: Deep 3d squeeze-and-excitation u-nets for fast and fully automated whole-volume anatomical segmentation," *arXiv preprint arXiv:1808.05238*, 2018.

[35] P. Wang and A. C. S. Chung, "Focal dice loss and image dilation for brain tumor segmentation," in *DLMIA/ML-CDS@MICCAI*, 2018.

[36] M. Khened, V. A. Kollerathu, and G. Krishnamurthi, "Fully convolutional multi-

scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers," *Medical image analysis*, vol. 51, pp. 21–45, 2019.

[37] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

[38] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[39] M. A. Islam, M. Rochan, S. Naha, N. D. Bruce, and Y. Wang, "Gated feedback refinement network for coarse-to-fine dense semantic image labeling," *arXiv preprint arXiv:1806.11266*, 2018.

[40] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[41] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 447–456.

[42] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel squeeze & excitationin fully convolutional networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 421–429.

[43] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," *arXiv preprint arXiv:1806.00064*, 2018.

[44] H. Jiang and Y. Guo, "Multi-class multimodal semantic segmentation with an improved 3d fully convolutional networks," *Neurocomputing*, 2019.

[45] L. Chen, Y. Wu, A. M. DSouza, A. Z. Abidin, A. Wismüller, and C. Xu, "Mri tumor segmentation with densely connected 3d cnn," in *Medical Imaging 2018: Image Processing*, vol. 10574. International Society for Optics and Photonics, 2018, p. 105741F.

[46] A. Myronenko, "3d mri brain tumor segmentation using autoencoder regularization,"

in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 311–320.

[47] K. Liu, Y. Li, N. Xu, and P. Natarajan, "Learn to combine modalities in multimodal deep learning," *arXiv preprint arXiv:1805.11730*, 2018.

[48] X. Liang, P. Hu, L. Zhang, J. Sun, and G. Yin, "Mcfnet: Multi-layer concatenation fusion network for medical images fusion," *IEEE Sensors Journal*, 2019.

[49] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio, "Hemis: Hetero-modal image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 469–477.

[50] A. Chartsias, T. Joyce, M. V. Giuffrida, and S. A. Tsaftaris, "Multimodal mr synthesis via modality-invariant latent representation," *IEEE transactions on medical imaging*, vol. 37, no. 3, pp. 803–814, 2018.

[51] H. Chen, Y. Qi, Y. Yin, T. Li, G. Gong, and L. Wang, "Mmfnet: A multi-modality mri fusion network for segmentation of nasopharyngeal carcinoma," *arXiv preprint arXiv:1812.10033*, 2018.

[52] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden, "Pyramid methods in image processing," *RCA engineer*, vol. 29, no. 6, pp. 33–41, 1984.

[53] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[54] G. van Tulder and M. de Bruijne, "Learning cross-modality representations from multi-modal images," *IEEE transactions on medical imaging*, vol. 38, no. 2, pp. 638–648, 2018.

Acronyms

AG Attention gates. 20

BraTs 2018 Brain Tumor Segmentation Challenge. 14, 32, 33, 35

BUS Breast Ultrasound Dataset B. 23, 24, 27

CNN Convolutional Neural Network. 4, 5, 10, 15

CT Computed Tomography. 1, 11–13

CV computer vision. 1, 2, 5–7

DL Dice loss. 8–10, 20, 25, 26, 32

DSC Dice Similarity Coefficient. 8–10, 25–27, 32

ET Enhancing Tumor. 32

FCN Fully Convolutional Network. 5, 7, 10

FLAIR Fluid-Attenuated In-version Recovery. 13, 32

FN false negative. 8, 9

FP false positive. 8, 9

FTI focalized Tversky Index. 18

FTL Focal Tversky loss. 18–23, 25–27, 31, 33, 35

GANs Generative Adversarial Networks. 7

HeMIS Hetero-modal Image Segmentation. 16, 17, 28–34

ILSVRC ImageNet Large Scale Visual Recognition Challenge. 4, 11

ISIC 2018 International Skin Imaging Collaboration. 23, 24, 26

MGF Moment Gated Fusion. 29–33

MRI Magnetic Resonance Imaging. 1, 13, 32, 33

PET Positron Emission Tomography. 1

ResNet Residual Networks. 5

ROI regions of interest. 2, 3, 6–9, 18, 19, 21–23, 34, 35

SE Squeeze and Excite. 11, 12, 29, 30, 33, 35

SEN Squeeze and Excitation Networks. 11

SGD Stochastic Gradient Descent. 32

TC Tumor Core. 32

TI Tversky Index. 9, 18, 19

TL Tversky Loss. 9, 20, 22, 26

VAE Variational Autoencoder. 35

VGG Visual Geometry Group. 5

WCE Weighted Cross-Entropy. 8

WT Whole Tumor. 32