

Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation

Michael Yeung, *E-mail*: `mjyy2@cam.ac.uk`

Department of Radiology, University of Cambridge, Cambridge CB2 0QQ, United Kingdom

School of Clinical Medicine, University of Cambridge, Cambridge CB2 0SP, United Kingdom

Evis Sala, *E-mail*: `es220@medschl.cam.ac.uk`

Department of Radiology, University of Cambridge, Cambridge CB2 0QQ, UK

Cancer Research UK Cambridge Centre, University of Cambridge, Cambridge CB2 0RE, UK

Carola-Bibiane Schönlieb, *E-mail*: `cbs31@cam.ac.uk`

Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 0WA, United Kingdom

Leonardo Rundo¹, *E-mail*: `lr495@cam.ac.uk`

Department of Radiology, University of Cambridge, Cambridge CB2 0QQ, United Kingdom

Cancer Research UK Cambridge Centre, University of Cambridge, Cambridge CB2 0RE, United Kingdom

¹Corresponding author. *Phone*: +44 (0) 1223 256255.

Address: Department of Radiology, University of Cambridge, Box 218, Cambridge Biomedical Campus, Cambridge CB2 0QQ, United Kingdom.

Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation

Michael Yeung^{b,a}, Evis Sala^{a,c}, Carola-Bibiane Schönlieb^d, Leonardo Rundo^{a,c,*}

^a*Department of Radiology, University of Cambridge,
Cambridge CB2 0QQ, United Kingdom*

^b*School of Clinical Medicine, University of Cambridge,
Cambridge CB2 0SP, United Kingdom*

^c*Cancer Research UK Cambridge Centre, University of Cambridge,
Cambridge CB2 0RE, United Kingdom*

^d*Department of Applied Mathematics and Theoretical Physics, University of Cambridge,
Cambridge CB3 0WA, United Kingdom*

Abstract

Automatic segmentation methods are an important advancement in medical image analysis. Machine learning techniques, and deep neural networks in particular, are the state-of-the-art for most medical image segmentation tasks. Issues with class imbalance pose a significant challenge in medical datasets, with lesions often occupying a considerably smaller volume relative to the background. Loss functions used in the training of deep learning algorithms differ in their robustness to class imbalance, with direct consequences for model convergence. The most commonly used loss functions for segmentation are based on either the cross entropy loss, Dice loss or a combination

*Corresponding author

Email addresses: `mjyy2@cam.ac.uk` (Michael Yeung), `es220@medsch1.cam.ac.uk` (Evis Sala), `cbs31@cam.ac.uk` (Carola-Bibiane Schönlieb), `lr495@cam.ac.uk` (Leonardo Rundo)

of the two. We propose a Unified Focal loss, a new framework that generalises Dice and cross entropy-based losses for handling class imbalance. We evaluate our proposed loss function on three highly class imbalanced, publicly available medical imaging datasets: Breast Ultrasound 2017 (BUS2017), Brain Tumour Segmentation 2020 (BraTS20) and Kidney Tumour Segmentation 2019 (KiTS19). We compare our loss function performance against six Dice or cross entropy-based loss functions, and demonstrate that our proposed loss function is robust to class imbalance, outperforming the other loss functions across datasets. Finally, we use the Unified Focal loss together with deep supervision to achieve state-of-the-art results without modification of the original U-Net architecture, with a mean Dice similarity coefficient (DSC)=0.948 on BUS2017, enhancing tumour region DSC=0.800 on BraTS20 and kidney tumour DSC=0.758 on KiTS19. This highlights the importance of carefully selecting a suitable loss function prior to the use of more complex architectures.

Keywords: Loss function, Class imbalance, Machine learning, Medical image segmentation, Computed Tomography, Magnetic Resonance Imaging, Ultrasound

1. Introduction

Image segmentation involves partitioning an image into meaningful regions, based on the regional pixel characteristics, from which objects of interest are identified (Pal & Pal, 1993). This is a fundamental task in computer vision and has been applied widely in face recognition, autonomous driving, as well as medical image processing. In particular, automatic segmentation

methods are an important advancement in medical image analysis, capable of demarcating structures across a range of imaging modalities including ultrasound (US), computed tomography (CT) and magnetic resonance imaging (MRI).

Classical approaches for image segmentation include direct region detection methods (such as the split-and-merge and region growing algorithms (Rundo et al., 2016)), graph-based methods (Chen & Pan, 2018), active contour and level set models (Khadidos et al., 2017). Later approaches have focused on applying and adapting traditional machine learning techniques (Rundo et al., 2020b), such as support vector machines (SVMs) (Wang & Summers, 2012), unsupervised clustering (Ren et al., 2019) and atlas-based segmentation (Wachinger & Golland, 2014). In recent years, however, significant progress has been achieved using deep learning (Ker et al., 2018; Rueckert & Schnabel, 2019).

The most well-known architecture in image segmentation, the U-Net (Ronneberger et al., 2015), is a modification of the convolutional neural network (CNN) architecture into an encoder-decoder network, similar to SegNet (Badrinarayanan et al., 2017), which enables end-to-end feature extraction and pixel classification. Since its inception, many variants based on the U-Net architecture have been proposed (Liu et al., 2020; Rundo et al., 2019a)—including the 3D U-Net (Çiçek et al., 2016), Attention U-Net (Schlemper et al., 2019) and V-Net (Milletari et al., 2016)—as well as integrated into conditional Generative Adversarial Networks (Armanious et al., 2020; Kessler et al., 2020).

To train deep neural networks, the optimisation of the loss function en-

ables model parameters to be updated by backpropagation accordingly. The cross entropy loss is typically the most widely used loss function in classification problems (Liu et al., 2020) and is applied in the U-Net (Ronneberger et al., 2015), 3D U-Net (Çiçek et al., 2016) and SegNet (Badrinarayanan et al., 2017). In contrast, Attention U-Net (Schlemper et al., 2019) and V-Net (Milletari et al., 2016) leverage the Dice loss function, which is based on the most commonly used metric for evaluating segmentation performance, and therefore represents a form of direct loss minimisation. Broadly, loss functions used in image segmentation may be classified into distribution-based losses (such as the cross entropy loss), region-based losses (such as Dice loss), boundary-based losses (such as the boundary loss) (Kervadec et al., 2019), and more recently compound losses. Compound losses combine multiple, independent loss functions, such as the Combo loss, which is the sum of Dice and cross entropy loss (Taghanaki et al., 2019).

A dominant issue in medical image segmentation is handling class imbalance, which refers to an unequal distribution of foreground and background elements. For example, automatic organ segmentation often involves organ sizes that are an order of magnitude smaller than the scan itself, resulting in a skewed distribution favouring background elements (Roth et al., 2015). This issue is even more prevalent in oncology, where tumour sizes are themselves often significantly smaller than the associated organ of origin. In these class imbalanced circumstances, careful selection of the loss function is crucial, with the Dice loss generally better suited than the cross entropy loss function. Taghanaki et al. (2019) distinguish between input and output imbalance, the former as aforementioned, and the latter referring to classification biases aris-

ing during inference. These include false positives and false negatives, which respectively describe background pixels incorrectly classified as foreground objects, and foreground objects incorrectly classified as background. Both are particularly important in the context of medical image segmentation; in the case of image-guided interventions, false positives may result in a larger radiation field or excessive surgical margins, and conversely false negatives may lead to inadequate radiation delivery or incomplete surgical resection. Therefore, it is important to design a loss function that can be optimised to handle both input and output imbalances.

Despite its significance, careful selection of the loss function is not widespread practice, and often suboptimal loss functions are chosen with performance repercussions. To inform loss function choice, it is important to perform large-scale comparisons of numerous loss functions. Seven loss functions were compared on the CVC-EndoSceneStill (gastrointestinal polyp segmentation) dataset, with the best performance seen with region-based losses and conversely the worst performance with the cross entropy loss (Sánchez-Peralta et al., 2020). Similarly, a comparison of fifteen loss functions using the NBFS Skull-stripping dataset (Eskildsen et al., 2012) (brain CT segmentation), which also introduces the log-cosh Dice loss, concluded that Focal Tversky loss and Tversky loss, both region-based losses, are generally optimal (Jadon, 2020). This is further supported by the most comprehensive loss function comparison to the date, with twenty loss functions compared across four datasets (liver, liver tumour, pancreas and multi-organ segmentation), which observed the best performance with Dice-based losses, where the most consistent performance was observed with the DiceTopK and DiceFocal loss

(Ma et al., 2021). It is apparent from these studies that region-based or compound losses are associated with consistently better performance than distribution-based losses. Less clear, however, is which of the region-based or compound losses to choose, with no agreement among the aforementioned. One major confounding factor is the degree of class imbalance in the datasets, with low class imbalance seen in the NBFS Skull-stripping dataset, moderate class imbalance in the CVC-EndoSceneStill dataset, and a combination of both low and high class imbalanced datasets present in (Ma et al., 2021).

Among medical imaging datasets, those involving tumour segmentation are associated with high degrees of class imbalance. Manual tumour delineation is both time-consuming and operator-dependent. Automatic methods of tumour delineation aim to address these issues, and public datasets, such as the Breast Ultrasound 2017 (BUS2017) dataset for breast tumours (Yap et al., 2017), Kidney Tumour Segmentation 19 (KiTS19) dataset for kidney tumours (Heller et al., 2019) and Brain Tumour Segmentation 2020 (BraTS20) for brain tumours (Menze et al., 2014), have accelerated progress towards this goal. In fact, there has been recent developments for translating the BraTS20 dataset into clinical and scientific practice (Kofler et al., 2020).

Current state-of-the-art models for the BUS2017 dataset incorporate attention gates, which may provide benefits in class imbalanced situations by using contextual information from the gating signal to refine skip connections, highlighting the regions of interest. In addition to attention gates, the RDAU-NET combines residual units and dilated convolutions to enhance information transfer and increase the receptive field, respectively, and was trained using the Dice loss (Zhuang et al., 2019). The multi-input Atten-

tion U-Net combines attention gates with deep supervision, and introduces the Focal Tversky loss, a region-based loss function designed to handle class imbalance (Abraham & Khan, 2019).

For the BraTS20 dataset, a popular approach is to use a multi-scale architecture where different receptive field sizes allow for the independent processing of both local and global contextual information (Havaei et al., 2017; Kamnitsas et al., 2017). Kamnitsas et al. (2017) used a two-phase training process involving initial upsampling of under-represented classes, followed by a second-stage where the output layer is retrained on a more representative sample. Similarly, Havaei et al. (2017) used a sampling rule to impose equal probability of foreground or background pixels at the centre of a patch, and used the cross entropy loss for optimisation.

For the KiTS19 dataset, the current state-of-the-art is the “no-new-Net” (nnU-Net) (Isensee et al., 2021, 2018), an automatically configurable deep learning-based segmentation method involving the ensemble of 2D, 3D and cascaded 3D U-Nets. This framework was optimised using the Dice and cross entropy loss. Recently, an ensemble-based method obtained comparable results to nnU-Net, and involved initial independent processing of kidney organ and kidney tumour segmentation by 2D U-Nets trained using the Dice loss, followed by suppression of false positive predictions of the kidney tumour segmentation using the network trained for kidney organ segmentation (Fatemeh et al., 2020). When the dataset size is small, results from an active learning-based method using CNN-corrected labeling, also trained using the Dice loss, showed a higher segmentation accuracy over nnU-Net (Kim et al., 2020).

It is apparent that for all three datasets, class imbalance is largely handled by altering either the training or input data sampling process, and rarely with adapting the loss function, except for example with the introduction of the Focal Tversky loss (Abraham & Khan, 2019). Even state-of-the-art solutions typically use unmodified versions of either the Dice loss, cross entropy loss or a combination of the two. However, popular methods—such as upsampling the underrepresented class—are inherently associated with an increase in false positive predictions, and more complicated, often multi-stage training processes require more computational resources.

In this paper, we propose the following contributions:

- (a) We summarise and extend the knowledge provided by previous studies that compare loss functions to address the context of high class imbalance, by using three highly class imbalanced datasets, covering both binary and multi-class classification, and across multiple imaging modalities.
- (b) We propose a new loss function, the Unified Focal loss, that generalises Dice-based and cross entropy-based loss functions for handling class imbalanced datasets.
- (c) Our proposed loss function improves segmentation quality over six other related loss functions across multiple classes and datasets, is associated with a better recall-precision balance, and is robust to class imbalance.
- (d) We achieve state-of-the-art performance using the Unified Focal loss without modification of the original U-Net architecture, highlighting

the importance of choosing the loss function with class imbalanced datasets.

The manuscript is organised as follows. Section 2 provides a summary of the loss functions used, including the proposed Unified Focal loss. Section 3 describes the chosen medical imaging datasets and defines the segmentation evaluation metrics used. Section 4 presents and discusses the experimental results. Finally, Section 5 provides conclusive remarks and future directions.

2. Background

The loss function defines the optimisation problem, and directly affects model convergence during training. This paper focuses on semantic segmentation, a sub-field of image segmentation where pixel-level classification is performed directly, in contrast to instance segmentation where an additional object detection stage is required. We describe seven loss functions that belong to either distribution-based, region-based or compound losses based of a combination of the two. A graphical overview of loss functions in these categories, and how all are derivable from the Unified Focal loss, is provided in Fig. 1. First, the distribution-based functions are introduced, followed by region-based loss functions, and finally concluding with compound loss functions.

2.1. *Cross entropy loss*

The cross entropy loss is one of the most widely used loss functions in deep learning. With origins in information theory, cross entropy measures the difference between two probability distributions for a given random variable

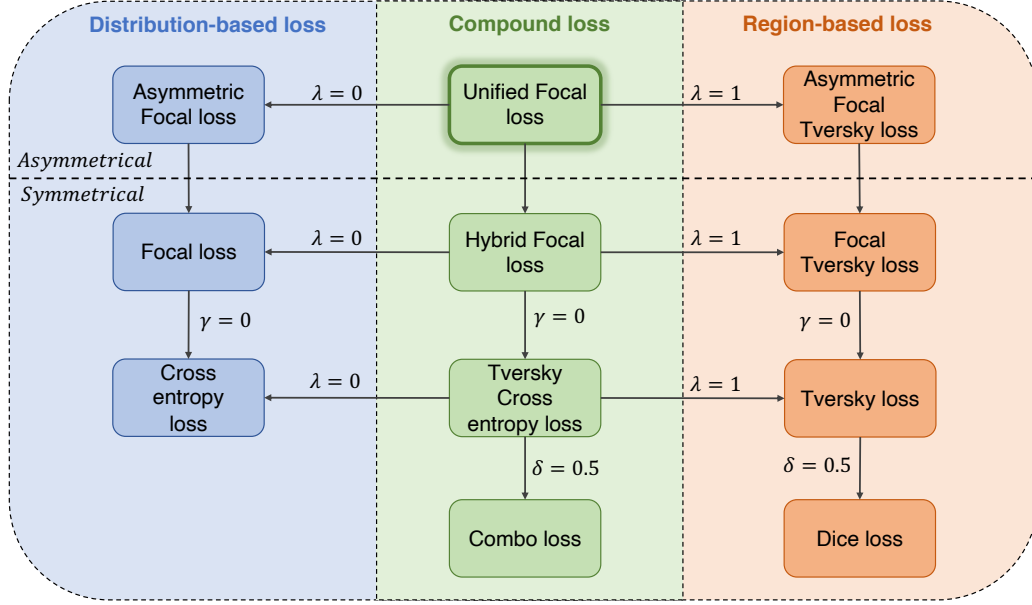


Figure 1: Our proposed framework unifying various distribution-based, region-based and compound loss functions. The arrows and associated hyperparameter values indicate the required hyperparameter value to set for the preceding loss function in order to recover the resulting loss function

or set of events. As a loss function, it is superficially equivalent to the negative log likelihood loss and, for binary classification, the binary cross entropy loss (\mathcal{L}_{BCE}) is defined as the following:

$$\mathcal{L}_{\text{BCE}}(\mathbf{y}, \hat{\mathbf{y}}) = -(\mathbf{y} \log(\hat{\mathbf{y}}) + (\mathbf{1} - \mathbf{y}) \log(\mathbf{1} - \hat{\mathbf{y}})). \quad (1)$$

Here, $\mathbf{y}, \hat{\mathbf{y}} \in \{0, 1\}^N$, where $\hat{\mathbf{y}}$ refers to the predicted value and \mathbf{y} refers to the ground truth label. This can be extended to multi-class problems, and

the categorical cross entropy loss (\mathcal{L}_{CCE}) is computed as:

$$\mathcal{L}_{\text{CCE}}(y, p) = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{i,c} \cdot \log(p_{i,c}), \quad (2)$$

where $y_{i,c}$ uses a one-hot encoding scheme of ground truth labels, $p_{i,c}$ is a matrix of predicted values for each class, and where indices c and i iterate over all classes and pixels, respectively. Cross entropy loss is based on minimising pixel-wise error, where in class imbalanced situations, leads to over-representation of larger objects in the loss, resulting in poorer quality segmentation of smaller objects.

2.2. Focal loss

The Focal loss is a variant of the binary cross entropy loss that addresses the issue of class imbalance faced by the standard cross entropy loss by down-weighting the contribution of easy examples enabling learning of harder examples (Lin et al., 2017). To derive the Focal loss function, we first simplify the loss in Eq. (1) as:

$$\text{CE}(p, y) = \begin{cases} -\log(p), & \text{if } y = 1 \\ -\log(1 - p), & \text{if } y = 0 \end{cases}. \quad (3)$$

Next, we define the probability of predicting the ground truth class, p_t , as:

$$p_t = \begin{cases} p, & \text{if } y = 1 \\ 1 - p, & \text{if } y = 0 \end{cases}. \quad (4)$$

The binary cross entropy loss (\mathcal{L}_{BCE}) can therefore be rewritten as:

$$\mathcal{L}_{\text{BCE}(p,y)} = \text{CE}(p_t) = -\log(p_t). \quad (5)$$

The Focal loss (\mathcal{L}_{F}) adds a modulating factor to the binary cross entropy loss:

$$\mathcal{L}_{\text{F}(p_t)} = \alpha (1 - p_t)^\gamma \cdot \mathcal{L}_{\text{BCE}(p,y)}, \quad (6)$$

The Focal loss is parameterised by α and γ , which control the class weights and degree of down-weighting of easy examples, respectively (Fig. 2). When $\gamma = 0$, the Focal loss simplifies to the binary cross entropy loss.

For multi-class classification, we define the categorical Focal loss (\mathcal{L}_{CF}):

$$\mathcal{L}_{\text{CF}} = \boldsymbol{\alpha} (1 - (p_{t,c}))^\gamma \cdot \mathcal{L}_{\text{CCE}}, \quad (7)$$

where $\boldsymbol{\alpha}$ is now a vector of class weights, $p_{t,c}$ is a matrix of ground truth probabilities for each class, and \mathcal{L}_{CCE} is the categorical cross entropy loss as defined in Eq. (2).

2.3. Dice loss

The Sørensen–Dice index, known as the Dice similarity coefficient (DSC) when applied to Boolean data, is the most commonly used metric for evaluating segmentation accuracy. We can define DSC in terms of the per voxel classification of true positives (TP), false positives (FP) and false negatives (FN):

$$\text{DSC} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \quad (8)$$

The Dice loss (\mathcal{L}_{DSC}), can therefore be defined as:

$$\mathcal{L}_{\text{DSC}} = 1 - \text{DSC}. \quad (9)$$

Other variants of the Dice loss include the Generalised Dice loss (Crum et al., 2006; Sudre et al., 2017) where the class weights are corrected by the inverse of their volume, and the Generalised Wasserstein Dice loss (Fidon et al., 2017), which combines the Wasserstein metric with the Dice loss and is adapted for dealing with hierarchical data, such as the BraTS20 dataset (Menze et al., 2014).

However, even in its most simple formulation, the Dice loss is somewhat adapted to handle class imbalance, with equal weighting provided to each class.

2.4. Tversky loss

The Tversky index (Salehi et al., 2017) is closely related to Dice, but enables optimisation for output imbalance by assigning weights α and β to false positives and false negatives respectively:

$$\text{TI} = \frac{\sum_{i=1}^N p_{0i} g_{0i}}{\sum_{i=1}^N p_{0i} g_{0i} + \alpha \sum_{i=1}^N p_{0i} g_{1i} + \beta \sum_{i=1}^N p_{1i} g_{0i}}, \quad (10)$$

where p_{0i} is the probability of pixel i belonging to the foreground class and p_{1i} is the probability of pixel belonging to background class. g_{0i} is 1 for foreground and 0 for background and conversely g_{1i} takes values of 1 for background and 0 for foreground.

Using the Tversky index, we define the Tversky loss for C classes, \mathcal{L}_T as:

$$\mathcal{L}_T = \sum_{c=1}^C (1 - \text{TI}) \quad (11)$$

When the Dice loss function is applied to highly class imbalanced problems, the resulting segmentation often exhibits high precision but low recall rate (Salehi et al., 2017). By assigning a greater weight to false negatives, recall rate is improved leading to a better balance of precision and recall. Therefore, β is often set higher than α , most commonly $\beta = 0.7$ and $\alpha = 0.3$.

The asymmetric similarity loss is derived from the Tversky loss, but uses the F_β score and substitutes α for $\frac{1}{1+\beta^2}$ and β for $\frac{\beta^2}{1+\beta^2}$, adding the constraint that α and β must sum to 1 (Hashemi et al., 2018). In practice, α and β values for the Tversky loss are chosen such that they sum to 1, making both loss functions functionally equivalent.

2.5. Focal Tversky loss

Inspired by the Focal loss adaptation of the cross entropy loss, the Focal Tversky loss (Abraham & Khan, 2019) adapts the Tversky loss by applying a focal parameter.

Using the definition of TI from Eq. (10), the Focal Tversky loss is defined (\mathcal{L}_{FT}) as:

$$\mathcal{L}_{\text{FT}} = \sum_{c=1}^C (1 - \text{TI})^{\frac{1}{\gamma}}, \quad (12)$$

where $\gamma < 1$ increases the degree of focusing on harder examples. The Focal Tversky loss simplifies to the Tversky loss when $\gamma = 1$. However, contrary to the Focal loss, the optimal value reported was $\gamma = 4/3$, which

enhances rather than suppresses the loss of easy examples. Indeed, near the end of training where the majority of the examples are more confidently classified and the Tversky index approaches 1, enhancing the loss in this region maintains a higher loss which may prevent premature convergence to a suboptimal solution.

2.6. Combo loss

The Combo loss (Taghanaki et al., 2019) belongs to the class of compound losses, where multiple loss functions are minimised in unison. The Combo loss ($\mathcal{L}_{\text{combo}}$) is defined as a weighted sum of the Dice similarity coefficient in Eq. (8) and a modified form of the cross entropy loss (\mathcal{L}_{mCE}):

$$\mathcal{L}_{\text{combo}} = \alpha (\mathcal{L}_{\text{mCE}}) - (1 - \alpha) \cdot \text{DSC}, \quad (13)$$

where:

$$\mathcal{L}_{\text{mCE}} = -\frac{1}{N} \sum_{i=1}^N \beta (y_i - \log(p_i)) + (1 - \beta) [(1 - y_i) \ln(1 - p_i)] \quad (14)$$

and α in the range of $[0, 1]$ controls the relative contribution of the Dice and cross entropy terms to the loss, and β controls the relative weights assigned to false positives and negatives. A value of $\beta > \frac{1}{2}$ penalises false negative predictions more than false positives.

Confusingly, the term “Dice and cross entropy loss” has been used to refer to both the sum of cross entropy loss and DSC (Isensee et al., 2018; Taghanaki et al., 2019), as well as the sum of the cross entropy loss and Dice loss, such as in the DiceFocal loss and Dice and weighted cross entropy loss (Chen et al., 2019; Zhu et al., 2019b). Here, we decide to use the former

definition, which is consistent with both Combo loss and the loss function used in the state-of-the-art for the KiTS19 dataset (Isensee et al., 2018).

2.7. Hybrid Focal loss

Combo loss (Taghanaki et al., 2019) and DiceFocal loss (Zhu et al., 2019b) are two compound loss functions that inherit benefits from both Dice and cross entropy-based loss functions. However, neither exploits the full benefits in the context of class imbalance. Both the Combo loss and the DiceFocal loss, with a tunable β and α parameter respectively in their cross entropy component losses, are partially robust to output imbalance. However, both lack an equivalent for the Dice component loss, where positive and negative examples remain equally weighted. Similarly, the Dice component of both losses are not adapted to handle input imbalance, although the DiceFocal loss is better adapted with its focal parameter in the Focal loss component.

To overcome this, we previously proposed the Hybrid Focal loss function, which incorporates tunable parameters to handle output imbalance, as well as focal parameters to handle input imbalance, for both the Dice and cross entropy-based component losses (Yeung et al., 2021). By replacing the Dice loss with the Focal Tversky loss, and the cross entropy loss with the Focal loss, the Hybrid Focal loss (\mathcal{L}_{HF}) is defined as:

$$\mathcal{L}_{\text{HF}} = \lambda \mathcal{L}_{\text{F}} + (1 - \lambda) \mathcal{L}_{\text{FT}}, \quad (15)$$

where $\lambda \in [0, 1]$ and determines the relative weighting of the two component loss functions.

2.8. Unified Focal loss

The Hybrid Focal loss adapts both the Dice and cross entropy based losses to handle class imbalance. However, there are two main issues associated with using the Hybrid Focal loss in practice. Firstly, there are five hyperparameters to tune: α and γ from the Focal loss, α / β and γ from the Focal Tversky loss, and λ to control the relative weighting of the two component losses. While this allows a greater degree of flexibility, this comes at the cost of a significantly larger hyperparameter search space. The second issue is common to all focal loss functions, where the downweighting effect introduced by the focal parameter leads to an overall suppressive effect on the loss, which may affect convergence towards the end of training.

The Unified Focal loss addresses both issues, by grouping functionally equivalent hyperparameters together and exploiting asymmetry to focus the suppressive and enhancing effects of the focal parameters in the Focal loss and Focal Tversky loss, respectively.

Asymmetry in this context refers to assigning different losses to the foreground and background objects. The asymmetric Focal loss (\mathcal{L}_{aF}) removes the focal parameter for the component of the loss relating to the rare class r , while retaining suppression of the background elements (Li et al., 2019):

$$L_{aF} = -\frac{1}{N} y_{i:r} \log(p_{t,r}) - \frac{1}{N} \sum_{c \neq r} (1 - p_{t,c})^\gamma \log(p_{t,r}) \quad (16)$$

To handle output imbalance, we include a δ term and define a modified

asymmetric Focal loss (\mathcal{L}_{maF}) as:

$$L_{\text{maF}} = -\frac{\delta}{N} y_{i:r} \log(p_{t,r}) - \frac{1-\delta}{N} \sum_{c \neq r} (1-p_{t,c})^\gamma \log(p_{t,r}), \quad (17)$$

where δ controls the relative contribution of positive and negative examples.

Similarly, we modify the Focal Tversky loss in two stages. First we replace α and β in the Tversky Index with a single hyperparameter δ and define a modified Tversky Index (mTI):

$$\text{mTI} = \frac{\sum_{i=1}^N p_{0i} g_{0i}}{\sum_{i=1}^N p_{0i} g_{0i} + \delta \sum_{i=1}^N p_{0i} g_{1i} + (1-\delta) \sum_{i=1}^N p_{1i} g_{0i}}, \quad (18)$$

Next, we remove the focal parameter for the component of the loss relating to the background, retaining enhancement of the rare class r , and define the asymmetric Focal Tversky loss (\mathcal{L}_{aFT}) as:

$$L_{\text{aFT}} = \sum_{c \neq r} (1 - \text{mTI}) + \sum_{c=r} (1 - \text{mTI})^{1-\gamma} \quad (19)$$

The Unified Focal loss (\mathcal{L}_{UF}), is therefore defined as:

$$\mathcal{L}_{\text{UF}} = \lambda \mathcal{L}_{\text{aF}} + (1-\lambda) \mathcal{L}_{\text{aFT}}, \quad (20)$$

where $\lambda \in [0, 1]$ and determines the relative weighting of the two losses.

By grouping functionally equivalent hyperparameters, the five hyperparameters associated with the Hybrid Focal loss are reduced to three, with δ controlling the relative weighting of positive and negative examples, γ controlling both suppression of the background class and enhancement of the rare class, and finally λ determining the weights of the two component losses. Furthermore, the issue of loss suppression associated with the Focal

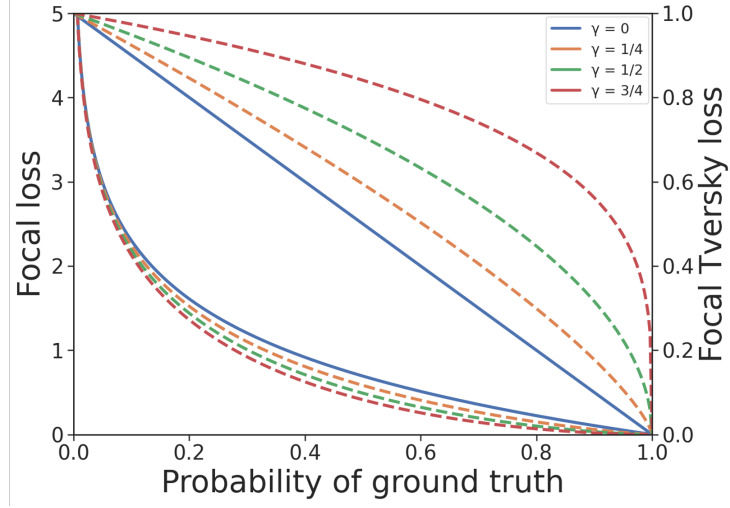


Figure 2: Effect of changing γ with the Unified Focal loss. The top and bottom group of curves relate to the Focal Tversky loss and Focal loss respectively. The dashed lines represent the foreground and background loss when changing γ for the Focal Tversky loss and Focal loss, respectively.

loss is mitigated by complementary pairing with the Focal Tversky loss, with the asymmetry enabling simultaneous background loss suppression and foreground loss enhancement, analogous to increasing the signal to noise ratio (Fig. 2).

By incorporating ideas from previous loss functions, the Unified Focal loss generalises Dice-based and cross entropy-based loss functions into a single framework. In fact, it can be shown that all Dice and cross entropy based loss functions described so far are special cases of the Unified Focal loss (Fig. 1). For example, by setting $\gamma = 0$ and $\delta = 0.5$, the Dice loss and the cross entropy loss are recovered when λ is set to 0 and 1 respectively.

3. Materials and methods

3.1. Dataset descriptions

We select three highly class imbalanced medical imaging datasets for our experiments: BUS2017, KiTS19 and BraTS20. To assess the degree of class imbalance, the percentage of foreground and background pixels/vowels were calculated for each dataset (Table 1).

Table 1: Percentage of foreground and background pixels/voxels per dataset. Values are calculated per image and averaged over the entire dataset.

Dataset	Foreground	Background
BUS2017	4.8% (breast lesion)	95.2%
BraTS20	0.2% (enhancing tumour)	99.8%
KiTS19	0.8% (kidney) 0.2% (tumour)	99.0%

3.1.1. BUS2017 dataset

The most commonly used screening tool for breast cancer assessment is digital mammography. However, dense breast tissue, often seen in younger patients, is poorly visualised on mammography. An important alternative is US imaging, which is an operator-dependent procedure requiring skilled radiologists, but has the advantage of no radiation exposure unlike mammography. BUS2017 dataset B consists of 163 ultrasound images and associated ground truth segmentations collected from the UDIAT Diagnostic Centre of the Parc Taulí Corporation, Sabadell, Spain. 110 images are benign lesions, consisting of 65 unspecified cysts, 39 fibroadenomas and 6 from other be-

nign types. The other 53 images depict cancerous masses, with the majority invasive ductal carcinomas.

3.1.2. BraTS20 dataset

Brain Tumour Segmentation 2020 (BraTS20) dataset is currently the largest, publicly available and fully-annotated dataset for medical image segmentation (Nazir et al., 2021), and comprises of 494 multimodal scans of patients with either low-grade glioma or high-grade glioblastoma (Bakas et al., 2017, 2018; Menze et al., 2014). The BraTS20 dataset provides images for the following MRI sequences: T1-weighted (T1), T1-weighted contrast-enhanced using gadolinium contrast agents (T1-CE), T2-weighted (T2) and fluid attenuated inverse recovery (FLAIR) sequence. Images were manually annotated, with regions associated with the tumour labelled as: necrotic and non-enhancing tumour core, peritumoural oedema or gadolinium-enhancing tumour. From the 494 scans provided, 125 scans are used for validation with reference segmentation masks withheld from public access, and therefore are excluded. We further exclude T1, T2 and FLAIR sequences to focus on gadolinium-enhancing tumour segmentation using the T1-CE sequence (Han et al., 2019; Rundo et al., 2019b), which not only appears to be the the most difficult class to segment (Henry et al., 2020), but is also the most clinically relevant for radiation therapy (Rundo et al., 2018, 2017). We further exclude another 27 scans without enhancing tumour regions, leaving 342 scans for use.

3.1.3. *KiTS19 dataset*

Kidney tumour segmentation is a challenging task due to the widespread presence of hypodense tissue, as well as highly heterogeneous appearance of tumours on CT (Linguraru et al., 2009; Rundo et al., 2020a). To evaluate our loss functions, we select the Kidney Tumour Segmentation 2019 (KiTS19) dataset (Heller et al., 2019), a highly class imbalanced, multi-class classification problem. Briefly, this dataset consists of 300 arterial phase abdominal CT scans from patients who underwent partial removal of the tumour and surrounding kidney or complete removal of the kidney including the tumour at the University of Minnesota Medical Center, USA. Kidney and tumour boundaries were manually delineated by two students, with class labels of either kidney, tumour or background assigned to each voxel resulting in a semantic segmentation task (Heller et al., 2019). 210 scans and their associated segmentations are provided for training, with the segmentation masks for the other 90 scans withheld from public access for testing. We therefore exclude the 90 scans without segmentation masks, and further exclude another 6 scans (case 15, 23, 37, 68, 125 and 133) due to concern over ground truth quality (Heller et al., 2021), leaving 204 scans for use.

3.2. *Experimental setup*

For our experiments, we make use of the Medical Image Segmentation with Convolutional Neural Networks (MIScnn) open-source Python library (Müller & Kramer, 2019). For both the KiTS19 and BraTS20 dataset, images and ground truth segmentation masks are provided in an anonymised NIFTI file format, while for BUS2017 these are provided in an anonymised png file format.

For the BUS2017 data, the mean image size is 760×570 pixels. To enable fair comparison, we follow the same preprocessing protocol as in (Abraham & Khan, 2019). Briefly, images are resized to 128×128 pixels, pixels are normalised to $[0, 1]$ using the z -score, and full-image analysis is performed with a batch size of 32. We use the original 2D U-Net architecture as defined in (Ronneberger et al., 2015), with a final softmax activation layer.

For the BraTS20 dataset, the original image size is $240 \times 240 \times 155$ voxels. The provided data are already pre-processed, with the skull stripped and images interpolated to the same isotropic resolution of 1mm^3 . We further normalise the pixel values to $[0, 1]$ using the z -score. We perform patch-wise analysis using random patches of size of $96 \times 96 \times 96$ voxels and patch-wise overlap of $48 \times 48 \times 48$ voxels, with a batch size of 2. For our model architecture, we use the standard 3D U-Net as described in (Çiçek et al., 2016) with a final softmax activation layer.

For the KiTS19 dataset, the original image size is 512×512 pixels in the axial plane, with an average of 216 slices in coronal plane. Pixel values are normalised to $[0, 1]$ using the z -score, Hounsfield units (HU) are clipped to $[-79, \dots, 304]$ HU and voxel spacing resampled to $3.22 \times 1.62 \times 1.62\text{mm}^3$. We perform patch-wise analysis using random patches of size of $80 \times 160 \times 160$ voxels and patch-wise overlap of $40 \times 80 \times 80$ voxels, with a batch size of 2. We use the same model architecture as for the KiTS19 dataset.

3.3. Implementation Details

For all datasets, we perform five-fold cross validation on remaining cases after exclusion. Since all scans belong to unique individuals, we perform a

single random assignment of scans to each fold and use the resulting configuration to evaluate all loss functions.

We evaluate the following loss functions: Focal loss, Dice loss, Tversky loss, Focal Tversky loss, Combo loss, Hybrid Focal loss and Unified Focal loss. We use optimal hyperparameters for each loss function as reported in the original studies. Specifically, we set $\alpha = 0.25$ and $\gamma = 2$ for the Focal loss (Lin et al., 2017), $\alpha = 0.3$, $\beta = 0.7$ for the Tversky loss (Salehi et al., 2017), $\alpha = 0.3$, $\beta = 0.7$ and $\gamma = 4/3$ for Focal Tversky loss (Abraham & Khan, 2019), $\alpha = \beta = 0.5$ for the Combo loss, and the same hyperparameters as for the individual Focal loss and Focal Tversky loss for the Hybrid Focal loss. For the Unified Focal loss, we perform hyperparameter tuning on the BUS2017 dataset, with values for $\delta \in [0.1, 0.9]$, $\gamma \in [0, 0.5]$ and $\lambda \in [0, 1]$.

Model parameters are initialised randomly, and we again make use of use MIScnn that leverages the ‘batchgenerators’ library to perform the following data augmentations: scaling, rotation, mirroring, gamma and elastic deformation.

We use two separate training schemes, one for hyperparameter tuning and loss function comparisons, and the second for attempting state-of-the-art results. For hyperparameter tuning, again we follow the training scheme used in (Abraham & Khan, 2019), training each model for 100 epochs using Stochastic Gradient Descent with Nesterov momentum ($\mu = 0.99$). The initial learning rate is set to 0.01, and follows a polynomial learning rate decay schedule:

$$\left(\frac{1 - \text{epoch}}{\text{epoch}_{\max}} \right)^{0.9} \quad (21)$$

For the KiTS19 and BraTS20 dataset, we train each model for 500 and 400 epochs respectively, using the Adam optimiser (Kingma & Ba, 2014) with an initial learning rate of 3.0×10^{-4} and minimum learning rate of 1.0×10^{-6} , with batch shuffling after each epoch.

When attempting for state-of-the-art results, for all datasets we train each model for 1000 epochs using Stochastic Gradient Descent with Nesterov momentum ($\mu = 0.99$), with an initial learning rate of 0.01 and polynomial learning rate decay schedule as previously defined. We also apply deep supervision as described in (Yeung et al., 2021), with weights w assigned to different output layers according to the stride length required to upsample back to the input dimensions:

$$w = 2^{-(\text{stride length} \times \text{stride width})}. \quad (22)$$

We supervise all layers including the final layer with the Unified Focal loss.

Validation loss is evaluated after each epoch, and the model with the lowest validation loss is selected as the final model. All experiments are programmed using Keras with TensorFlow backend and trained using NVIDIA P100 GPUs. Source code is available at: <https://github.com/mlyg/unified-focal-loss>.

3.4. Evaluation Metrics

To assess segmentation accuracy, we use three commonly used metrics (Wang et al., 2020): Dice similarity coefficient (DSC), recall and precision. DSC is defined as Eq. (8), and recall and precision are defined similarly per

pixel/voxel and according to Eqs. (23) and (24), respectively:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (23)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (24)$$

4. Experimental results

In this section, we first describe the results from hyperparameter tuning of the Unified Focal loss on the BUS2017 dataset, followed by loss function comparisons on the BraTS20 and KiTS19 datasets, and finally the results from training the original U-Net with deep supervision using the Unified Focal loss.

The results for the hyperparameter tuning experiments on the BUS2017 dataset are shown in Table 2.

Table 2: Hyperparameter tuning on the BUS2017 dataset. Values are in the form mean \pm 95% confidence intervals. Numbers in boldface denote the highest values for each metric.

$\gamma = 0, \lambda = 0.5$				$\delta = 0.6, \lambda = 0.5$				$\delta = 0.6, \gamma = 0.2$			
δ	DSC	Precision	Recall	γ	DSC	Precision	Recall	λ	DSC	Precision	Recall
0.1	0.707 \pm 0.055	0.872\pm0.032	0.639 \pm 0.063	0.05	0.798 \pm 0.020	0.802 \pm 0.036	0.849\pm0.016	0	0.555 \pm 0.138	0.764 \pm 0.056	0.523 \pm 0.163
0.2	0.716 \pm 0.034	0.852 \pm 0.058	0.680 \pm 0.046	0.1	0.795 \pm 0.020	0.828 \pm 0.060	0.821 \pm 0.063	0.1	0.701 \pm 0.061	0.761 \pm 0.102	0.743 \pm 0.081
0.3	0.716 \pm 0.061	0.835 \pm 0.049	0.689 \pm 0.093	0.15	0.795 \pm 0.025	0.809 \pm 0.067	0.828 \pm 0.034	0.2	0.704 \pm 0.059	0.784 \pm 0.067	0.712 \pm 0.096
0.4	0.755 \pm 0.057	0.803 \pm 0.061	0.766 \pm 0.064	0.2	0.814\pm0.027	0.829\pm0.046	0.838 \pm 0.038	0.3	0.776 \pm 0.043	0.775 \pm 0.072	0.834 \pm 0.030
0.5	0.776 \pm 0.058	0.821 \pm 0.056	0.786 \pm 0.079	0.25	0.810 \pm 0.011	0.824 \pm 0.045	0.849\pm0.035	0.4	0.778 \pm 0.020	0.763 \pm 0.040	0.850\pm0.031
0.6	0.793\pm0.035	0.835 \pm 0.061	0.814 \pm 0.040	0.3	0.770 \pm 0.034	0.784 \pm 0.054	0.810 \pm 0.036	0.5	0.814\pm0.027	0.829\pm0.046	0.838 \pm 0.038
0.7	0.788 \pm 0.056	0.774 \pm 0.088	0.857 \pm 0.036	0.35	0.721 \pm 0.106	0.771 \pm 0.042	0.760 \pm 0.165	0.6	0.777 \pm 0.034	0.810 \pm 0.035	0.806 \pm 0.053
0.8	0.755 \pm 0.039	0.740 \pm 0.076	0.844 \pm 0.040	0.4	0.740 \pm 0.040	0.788 \pm 0.073	0.763 \pm 0.090	0.7	0.771 \pm 0.027	0.810 \pm 0.046	0.788 \pm 0.034
0.9	0.739 \pm 0.065	0.672 \pm 0.092	0.902\pm0.049	0.45	0.761 \pm 0.027	0.771 \pm 0.058	0.808 \pm 0.047	0.8	0.758 \pm 0.034	0.776 \pm 0.036	0.813 \pm 0.075
				0.5	0.745 \pm 0.036	0.765 \pm 0.062	0.787 \pm 0.071	0.9	0.735 \pm 0.036	0.743 \pm 0.112	0.824 \pm 0.091
								1	0.744 \pm 0.042	0.715 \pm 0.071	0.843 \pm 0.049

The parameter δ controls the relative contribution of the positive and negative examples, with values of $\delta > \frac{1}{2}$ penalising false negative predictions

more than false positives. This is consistent with the data, where low values of δ are associated with the highest precision, and high values of delta associated with the highest recall. The highest DSC is observed with $\delta = 0.6$, slightly favouring recall over precision, and consistent with previous results observed with Tversky-based losses (Abraham & Khan, 2019; Salehi et al., 2017). γ controls the signal to noise ratio, by both enhancing the foreground contribution while suppressing the loss from the background elements. There appears to be a unimodal distribution, with the highest DSC observed with $\gamma = 0.2$. This suggests that increasing gamma improves performance up to a point, from which performance decreases, likely as a result of increased training instability at higher values of gamma. Finally, λ controls the relative contribution of the two component losses. At the extremes, the Unified Focal loss degenerates into the asymmetric Focal loss and asymmetric Focal Tversky loss at $\lambda = 1$ and $\lambda = 0$ respectively. Individually, the performance is poor, especially with the asymmetric Focal loss. However, the highest DSC was observed with $\lambda = 0.5$, where there is equal contribution of the asymmetric Focal Tversky loss and asymmetric Focal loss, suggesting the Unified Focal loss function outperforms its component losses. Based on the hyperparameter tuning experiments, the optimal hyperparameters $\delta = 0.6$, $\gamma = 0.2$ and $\lambda = 0.5$ are used in subsequent experiments.

The results of evaluating the seven loss functions on the BUS2017 dataset are shown in Table 3. The best performance was observed with the compound loss functions, with the Unified Focal loss after hyperparameter tuning associated with the highest DSC of 0.814 ± 0.027 , followed by the Hybrid Focal loss with a DSC of 0.795 ± 0.026 and Combo loss with a DSC of 0.791 ± 0.022 . The

Unified Focal loss was also associated with the highest recall of 0.838 ± 0.038 . The highest precision was observed with the Hybrid Focal loss and Focal Tversky loss with a precision of 0.866 ± 0.025 and 0.866 ± 0.024 respectively. While region-based losses are perform only slightly worse than compound loss functions, distribution-based losses were associated with the lowest DSC, with a comparable precision but significantly lower recall.

Table 3: Performance on the BUS2017 dataset. Values are in the form mean \pm 95% confidence intervals. Numbers in boldface denote the highest values for each metric.

Loss function	DSC	Precision	Recall
Focal loss	0.653 ± 0.038	0.799 ± 0.037	0.621 ± 0.042
Dice loss	0.775 ± 0.042	0.841 ± 0.060	0.765 ± 0.048
Tversky	0.785 ± 0.024	0.815 ± 0.033	0.808 ± 0.026
Focal Tversky loss	0.781 ± 0.017	0.866 ± 0.024	0.750 ± 0.026
Combo loss	0.791 ± 0.022	0.833 ± 0.040	0.800 ± 0.048
Hybrid Focal loss	0.795 ± 0.026	0.866 ± 0.025	0.777 ± 0.028
Unified Focal loss	0.814 ± 0.027	0.829 ± 0.046	0.838 ± 0.038

Next, we evaluated the loss functions on the BraTS20 dataset (Table 4).

The Unified Focal loss outperforms all other loss functions with a DSC of 0.778 ± 0.031 , precision of 0.784 ± 0.041 and recall of 0.815 ± 0.023 . The highest precision was associated with the Focal loss, with a precision of 0.847 ± 0.027 . A high precision at the cost of recall is characteristic of distribution-based loss functions in the context of class imbalance, where under-prediction of the rare class reduces false positives and therefore increases precision. The performance of the Hybrid Focal loss is similar to the Unified Focal loss,

with a DSC of 0.775 ± 0.033 . This is followed by the Combo loss and Focal Tversky loss, with a DSC of 0.748 ± 0.032 and 0.747 ± 0.050 respectively. The lowest performance was observed using the Tversky loss with a DSC of 0.719 ± 0.040 , associated with a high recall but at a cost of significantly lower precision of 0.692 ± 0.047 .

Table 4: Performance on the BraTS20 dataset. Values are in the form mean \pm 95% confidence intervals. Numbers in boldface denote the highest values for each metric.

Loss function	DSC	Precision	Recall
Focal loss	0.735 ± 0.027	0.847 ± 0.027	0.694 ± 0.023
Dice loss	0.725 ± 0.034	0.767 ± 0.050	0.739 ± 0.038
Tversky	0.719 ± 0.040	0.692 ± 0.047	0.811 ± 0.020
Focal Tversky loss	0.747 ± 0.050	0.776 ± 0.070	0.765 ± 0.029
Combo loss	0.748 ± 0.032	0.833 ± 0.039	0.716 ± 0.022
Hybrid Focal loss	0.775 ± 0.033	0.798 ± 0.032	0.795 ± 0.027
Unified Focal loss	0.778 ± 0.031	0.784 ± 0.041	0.815 ± 0.023

The seven loss functions were also evaluated on the KiTS19 dataset (Table 5). Similarly, the best performance was observed with the Unified Focal loss and Hybrid Focal loss, with a score of 0.942 ± 0.013 and 0.945 ± 0.017 for the DSC kidney respectively, and a score of 0.751 ± 0.058 and 0.751 ± 0.057 for the DSC tumour. Furthermore, the Hybrid Focal loss is associated with the highest kidney recall of 0.945 ± 0.017 , and the Unified Focal loss is associated with the highest tumour recall of 0.788 ± 0.054 . Despite a low DSC tumour score, Focal loss was associated with the highest kidney precision score of 0.952 ± 0.011 . The highest tumour precision scores were observed

Table 5: Performance on the KiTS19 dataset. Values are in the form mean \pm 95% confidence intervals. Numbers in boldface denote the highest values for each metric.

Loss function	DSC kidney	Precision kidney	Recall kidney	DSC tumour	Precision tumour	Recall tumour
Focal loss	0.940 \pm 0.020	0.952\pm0.011	0.935 \pm 0.021	0.576 \pm 0.079	0.718 \pm 0.036	0.549 \pm 0.086
Dice loss	0.938 \pm 0.018	0.948 \pm 0.012	0.934 \pm 0.019	0.695 \pm 0.066	0.794\pm0.034	0.684 \pm 0.084
Tversky loss	0.936 \pm 0.018	0.928 \pm 0.010	0.949 \pm 0.021	0.711 \pm 0.081	0.754 \pm 0.073	0.732 \pm 0.084
Focal Tversky loss	0.943 \pm 0.015	0.934 \pm 0.011	0.955 \pm 0.016	0.740 \pm 0.061	0.772 \pm 0.038	0.764 \pm 0.068
Combo loss	0.943 \pm 0.014	0.951 \pm 0.013	0.938 \pm 0.018	0.723 \pm 0.047	0.794\pm0.051	0.722 \pm 0.051
Hybrid Focal loss	0.945\pm0.017	0.936 \pm 0.012	0.957\pm0.017	0.751\pm0.057	0.791 \pm 0.026	0.778 \pm 0.069
Unified Focal loss	0.942 \pm 0.013	0.944 \pm 0.013	0.941 \pm 0.015	0.751\pm0.058	0.774 \pm 0.052	0.788\pm0.054

with both the Dice loss and Combo loss. Despite high precision scores for the Dice loss, Combo loss and Focal loss, these loss functions were associated with lower recall scores, and subsequently lower DSC values. In contrast, higher recall scores were obtained by the Tversky loss and its variants across both kidney and tumour segmentations, although this was balanced by lower precision scores. Comparing compound losses with their component losses, besides equivalent scores for tumour precision, the Combo loss outperformed the Dice loss across all other metrics. Similarly, the Hybrid Focal loss and Unified Focal loss outperformed both the Focal Tversky loss and Focal loss, except for the kidney precision score. Finally, comparisons between the compound losses showed better recall-precision balance with the Unified Focal loss and Hybrid Focal loss, outperforming the Combo loss for both the DSC and recall metrics.

Examples of image segmentations for each dataset are shown in Fig. 3. The high false negative rates associated with distribution-based losses are visible from the resulting segmentations, where the rare class is consistently under-predicted. This is most noticeable in the KiTS19 dataset, where the

boundary between the tumour and kidney are shifted in favour of kidney prediction, and is an expected consequence of the over-representation of the larger kidney class in the loss. In contrast, the Tversky loss is associated with over-prediction of the rare class across all datasets, with predictions of the rare class extending beyond its true boundary, reflecting its higher recall but lower precision scores. The segmentations observed with the compound loss functions appear most accurate, and this is consistent across datasets.

Finally, we train the original U-Net without modifications using the Unified Focal loss combined with deep supervision for 1000 epochs on each dataset. We compare with state-of-the-art models trained and evaluated on the same dataset. Importantly, this excludes some state-of-the-art models where performance is only reported for the challenges’ validation or test set. The results, compared against the state-of-the-art, are shown in Tables 6–8.

Table 6: Performance comparisons on the BUS2017 dataset. Values, where available, are in the form mean \pm 95% confidence intervals. Numbers in boldface denote the highest values for each metric.

Architecture	Loss function	Ensemble	Post-processing	DSC
Multi-input Attention U-Net (Abraham & Khan, 2019)	Focal Tversky loss	✗	✗	0.804 \pm 0.024
Two-stage U-Net (Amiri et al., 2020)	Cross entropy loss	✓	✗	0.805 \pm 0.210
RDAU-Net (Zhuang et al., 2019)	Dice loss	✗	✗	0.845
U-Net (with proposed loss)	Unified Focal loss	✗	✗	0.948\pm0.009

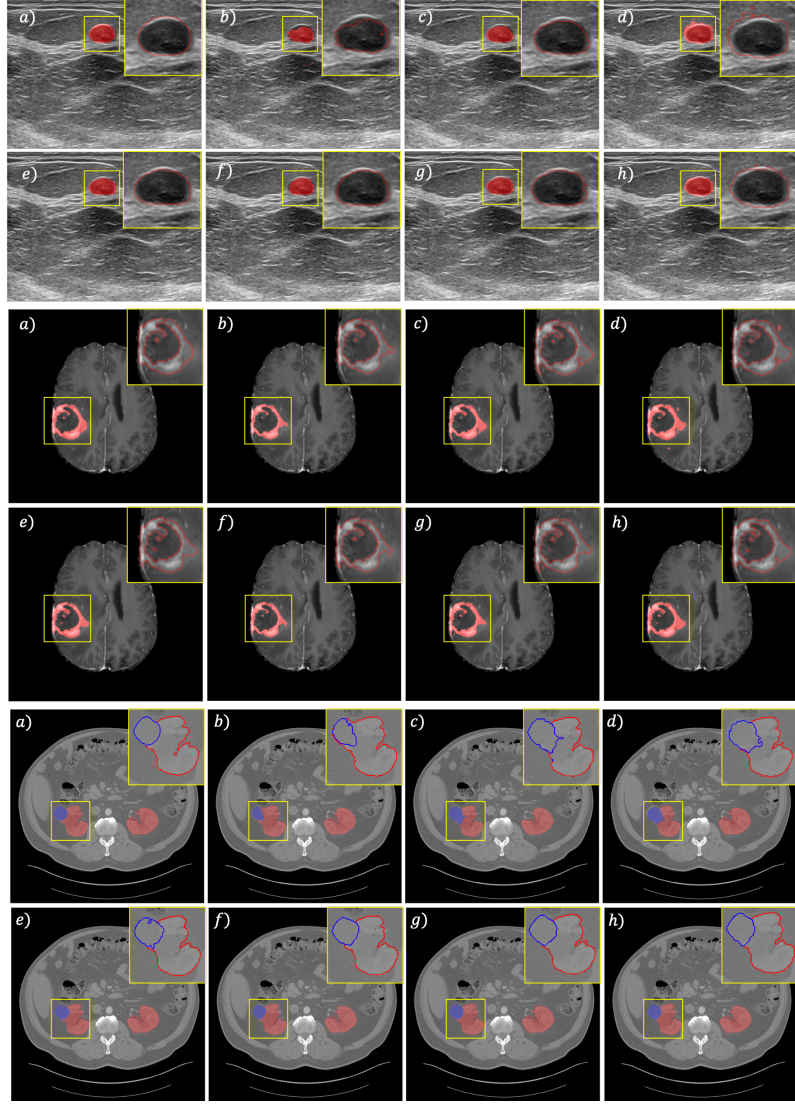


Figure 3: Image segmentations generated from the BUS2017 (top), BraTS20 (middle) and KiTS19 (bottom) datasets using (a) ground truth, (b) Focal loss, (c) Dice loss, (d) Tversky loss, (e) Focal Tversky loss, (f) Combo loss, (g) Hybrid Focal loss and (h) Unified Focal loss. For BUS2017 and BraTS20 datasets, the breast lesion and enhancing tumour region are highlighted in red respectively. For the KiTS19 dataset, the kidney is highlighted in red and the tumour in blue. A magnified contour of the segmentation is provided in the top right-hand corner of each image.

Table 7: Performance comparisons on the BraTS20 dataset. Values, where available, are in the form mean \pm 95% confidence intervals. Numbers in boldface denote the highest values for each metric.

Architecture	Loss function	Ensemble	Post-processing	DSC
Multiscale residual 3D-U-Net (Mora Ballestar & Vilaplana, 2020b)	Unified Dice loss	✓	✓	0.77
Cascaded 3D densely-connected U-Net (Ghaffari et al., 2020)	Multi-class Dice loss	✓	✓	0.78 \pm 0.26
V-Net (Mora Ballestar & Vilaplana, 2020a)	Generalised Dice loss	✗	✓	0.677
MDNet (Vu et al., 2020)	DSC + Cross entropy loss	✓	✓	0.781
U-Net (with proposed loss)	Unified Focal loss	✗	✗	0.800\pm0.030

Table 8: Performance comparisons on the KiTS19 dataset. Values, where available, are in the form mean \pm 95% confidence intervals. Numbers in boldface denote the highest values for each metric.

Architecture	Loss function	Ensemble	Post-processing	DSC kidney	DSC tumour
Multi-stage 2.5D Res U-Net (Santini et al., 2019)	Categorical cross entropy loss	✓	✓	0.98\pm0.01	0.73 \pm 0.25
BiSC U-Net (Wang et al., 2019)	Balanced cross entropy loss	✗	✗	0.954	0.741
SE-ResNeXT U-Net (Xie et al., 2020)	Not stated	✗	✗	0.968	0.743
Multi-scale 3D U-Net (Zhao & Zeng, 2019)	Exponential logarithmic loss	✗	✓	0.969	0.805
Triple-stage 3D U-Net (Hou et al., 2020)	DSC + weighted cross entropy loss	✓	✓	0.964	0.838
U-Net (with proposed loss)	Unified Focal loss	✗	✗	0.948 \pm 0.012	0.758 \pm 0.066

Across all datasets, the results obtained by training the original U-Net using the Unified Focal loss produces competitive, if not state-of-the-art, results even when compared to significantly more complicated models, with the highest DSC score for both BUS2017 and BraTS20 of 0.948 ± 0.009 and 0.800 ± 0.030 respectively. Results from KiTS19 are not state-of-the-art when compared to models such as the Triple-stage 3D U-Net (Hou et al., 2020), but are still competitive. Clearly, there is a limit to the performance by using a simple U-Net architecture with no post-processing, compared to state-of-the-art results such as the triple-stage 3D U-Net which involves ensembling three modified 3D U-Net architectures, as well as subsequent post-processing.

Comparing results across all three datasets, it is apparent that the better performing architectures make use of compound loss functions, with the worst performance associated with models using distribution-based losses, even following modifications, such as assigning class weights.

5. Discussion and Conclusions

In this study, we proposed a new compound loss function, the Unified Focal loss, which generalises Dice and cross entropy-based loss functions for handling class imbalance. We compared seven region-based and distribution-based loss functions on three highly class imbalanced datasets (namely BUS2017, BraTS20 and KiTS19), consistently achieving the highest Dice score associated with the Unified Focal loss. The difference in model performance across the numerous loss functions highlights the importance of loss function choice in class imbalanced image segmentation tasks. Most noticeable is the consistently poor performance using distribution-based losses, such as the cross entropy loss. The susceptibility of distribution-based losses is expected, given the tendency to prioritise precision over recall in order to maximise per pixel accuracy. On the other hand, the best performance is observed with compound loss functions, which holds true even when using state-of-the-art architectures. Without modification of the original U-Net architecture, we compared our results from training with the Unified Focal loss to current state-of-the-art methods on all three datasets. We achieved state-of-the-art results on both BUS2017 and BraTS20 datasets, and competitive results for the KiTS19 dataset.

There are several limitations associated with our study. Firstly, we per-

formed hyperparameter tuning on BUS2017, which although is class imbalanced, is still an order of magnitude less class imbalanced than the BraTS20 and KiTS19 datasets. It is unclear whether the optimal hyperparameters for BUS2017 are applicable to BraTS20 and KiTS19, and better results may be seen with higher values of γ . In fact, this may explain why similar performance was observed with the Hybrid Focal loss despite the addition of asymmetry to the Unified Focal loss, because the Hybrid Focal loss uses higher values of γ for both the Focal loss and Focal Tversky loss components. The less successful, yet competitive, performance on the KiTS19 dataset compared to state-of-the-art performance on the two binary datasets may also suggest the need for a different formulation when generalising to multi-class problems. It is not immediately clear whether to treat all foreground classes as the rare class as in these experiments, and better performance may be observed if foreground classes were treated differently, given that for example the kidney class in the KiTS19 dataset is four times more prevalent than the tumour class.

For our loss function comparisons, we focused our experiments on seven loss functions and illustrated how the Unified Focal loss encapsulates these loss functions into a single framework. However, it should be noted that the Unified Focal loss further generalises other loss functions, such as the Dice-Focal loss (Zhu et al., 2019b) and Asymmetric similarity loss (Hashemi et al., 2018). The comparisons presented in this study are also restricted to only a subset of region-based and distribution-based loss functions, and we further did not include any boundary-based loss functions (Kervadec et al., 2019; Zhu et al., 2019a), another class of loss functions that instead use distance-

based metrics to optimise contours rather than distributions or regions used by cross entropy and Dice-based losses, respectively.

We conclude by highlighting several areas for future research. To inform the loss function choice for class imbalanced segmentation, it is important to compare a greater number and variety of loss functions, especially from other loss function classes and with different class imbalanced datasets. We use the original U-Net architecture to simplify but also highlight the importance of loss functions on performance, but it would be useful to assess whether the performance gains generalise to state-of-the-art deep learning methods—such as the nnU-Net (Isensee et al., 2021)—and whether this is able to complement or even replace alternatives, such as training or sampling-based methods for handling class imbalance.

Acknowledgements

This work was partially supported by The Mark Foundation for Cancer Research and Cancer Research UK Cambridge Centre [C9685/A25177], the CRUK National Cancer Imaging Translational Accelerator (NCITA) [C42780/A27066] and the Wellcome Trust Innovator Award [RG98755]. Additional support was also provided by the National Institute of Health Research (NIHR) Cambridge Biomedical Research Centre [BRC-1215-20014] and the Cambridge Mathematics of Information in Healthcare (CMIH) [funded by the EPSRC grant EP/T017961/1]. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care.

CBS in addition acknowledges support from the Leverhulme Trust project

on ‘Breaking the non-convexity barrier’, the Philip Leverhulme Prize, the Royal Society Wolfson Fellowship, the EPSRC grants EP/S026045/1, EP/N014588/1, EP/T017961/1, European Union Horizon 2020 research and innovation programmes under the Marie Skłodowska-Curie grant agreement No. 777826 NoMADS and No. 691070 CHiPS, the Cantab Capital Institute for the Mathematics of Information and the Alan Turing Institute.

This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (www.csd3.cam.ac.uk), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/P020259/1), and DiRAC funding from the Science and Technology Facilities Council (www.dirac.ac.uk).

References

- Abraham, N., & Khan, N. M. (2019). A novel focal Tversky loss function with improved attention U-Net for lesion segmentation. In *Proc. 16th International Symposium on Biomedical Imaging (ISBI)* (pp. 683–687). IEEE. doi:10.1109/ISBI.2019.8759329.
- Amiri, M., Brooks, R., Behboodi, B., & Rivaz, H. (2020). Two-stage ultrasound image segmentation using u-net and test time augmentation. *International journal of computer assisted radiology and surgery*, 15, 981–988.
- Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S., & Yang, B. (2020). MedGAN: Medical image translation

- using GANs. *Comput. Med. Imaging Graph.*, 79, 101684. doi:10.1016/j.compmedimag.2019.101684.
- Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39, 2481–2495. doi:10.1109/TPAMI.2016.2644615.
- Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., Freymann, J. B., Farahani, K., & Davatzikos, C. (2017). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data*, 4, 170117. doi:10.1038/sdata.2017.117.
- Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R. T., Berger, C., Ha, S. M., Rozycki, M. et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv preprint arXiv:1811.02629*, .
- Chen, C., Dou, Q., Jin, Y., Chen, H., Qin, J., & Heng, P.-A. (2019). Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (pp. 447–456). Springer. doi:10.1007/978-3-030-32248-9_50.
- Chen, X., & Pan, L. (2018). A survey of graph cuts/graph search based medical image segmentation. *IEEE Rev. Biomed. Eng.*, 11, 112–124. doi:10.1109/RBME.2018.2798701.

- Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (pp. 424–432). Springer. doi:10.1007/978-3-319-46723-8_49.
- Crum, W. R., Camara, O., & Hill, D. L. G. (2006). Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Trans. Med. Imaging*, 25, 1451–1461. doi:10.1109/TMI.2006.880587.
- Eskildsen, S. F., Coupé, P., Fonov, V., Manjón, J. V., Leung, K. K., Guizard, N., Wassef, S. N., Østergaard, L. R., Collins, D. L., Initiative, A. D. N. et al. (2012). BEaST: brain extraction based on nonlocal segmentation technique. *NeuroImage*, 59, 2362–2373. doi:10.1016/j.neuroimage.2011.09.012.
- Fatemeh, Z., Nicola, S., Satheesh, K., & Eranga, U. (2020). Ensemble U-net-based method for fully automated detection and segmentation of renal masses on computed tomography images. *Med. Phys.*, 47, 4032–4044.
- Fidon, L., Li, W., Garcia-Peraza-Herrera, L. C., Ekanayake, J., Kitchen, N., Ourselin, S., & Vercauteren, T. (2017). Generalised wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks. In *Proc. International MICCAI Brainlesion Workshop* (pp. 64–76). Springer. doi:10.1007/978-3-319-75238-9_6.
- Ghaffari, M., Sowmya, A., & Oliver, R. (2020). Brain tumour seg-

- mentation using cascaded 3d densely-connected u-net. *arXiv preprint arXiv:2009.07563*, .
- Han, C., Rundo, L., Araki, R., Nagano, Y., Furukawa, Y. et al. (2019). Combining noise-to-image and image-to-image GANs: brain MR image augmentation for tumor detection. *IEEE Access*, 7, 156966–156977. doi:10.1109/ACCESS.2019.2947606.
- Hashemi, S. R., Salehi, S. S. M., Erdogmus, D., Prabhu, S. P., Warfield, S. K., & Gholipour, A. (2018). Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection. *IEEE Access*, 7, 1721–1735. doi:10.1109/ACCESS.2018.2886371.
- Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., & Larochelle, H. (2017). Brain tumor segmentation with deep neural networks. *Med. Image Anal.*, 35, 18–31.
- Heller, N., Isensee, F., Maier-Hein, K. H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M. et al. (2021). The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the KiTS19 challenge. *Med. Image Anal.*, 67, 101821. doi:10.1016/j.media.2020.101821.
- Heller, N., Sathianathan, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M. et al. (2019). The KiTS19 challenge data: 300 kidney tumor cases with clinical context. *arXiv preprint arXiv:1904.00445*, .

- Henry, T., Carre, A., Lerousseau, M., Estienne, T., Robert, C., Paragios, N., & Deutsch, E. (2020). Top 10 BraTS 2020 challenge solution: Brain tumor segmentation with self-ensembled, deeply-supervised 3D-Unet like neural networks. *arXiv preprint arXiv:2011.01045*, .
- Hou, X., Xie, C., Li, F., Wang, J., Lv, C., Xie, G., & Nan, Y. (2020). A triple-stage self-guided network for kidney tumor segmentation. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* (pp. 341–344). IEEE.
- Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods*, *18*, 203–211. doi:10.1038/s41592-020-01008-z.
- Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P. F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S. et al. (2018). nnU-net: Self-adapting framework for U-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, .
- Jadon, S. (2020). A survey of loss functions for semantic segmentation. In *Proc. Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* (pp. 1–7). IEEE. doi:10.1109/CIBCB48159.2020.9277638.
- Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., & Glocker, B. (2017). Efficient multi-scale

- 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.*, 36, 61–78.
- Ker, J., Wang, L., Rao, J., & Lim, T. (2018). Deep learning applications in medical image analysis. *IEEE Access*, 6, 9375–9389. doi:10.1109/ACCESS.2017.2788044.
- Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., & Ayed, I. B. (2019). Boundary loss for highly unbalanced segmentation. In *Proc. International Conference on Medical Imaging with Deep Learning (MIDL)* (pp. 285–296). PMLR.
- Kessler, D. A., MacKay, J. W., Crowe, V. A., Henson, F. M., Graves, M. J., Gilbert, F. J., & Kaggie, J. D. (2020). The optimisation of deep neural networks for segmenting multiple knee joint tissues from MRIs. *Comput. Med. Imaging Graph.*, 86, 101793. doi:10.1016/j.compmedimag.2020.101793.
- Khadidos, A., Sanchez, V., & Li, C.-T. (2017). Weighted level set evolution based on local edge features for medical image segmentation. *IEEE Trans. Image Process.*, 26, 1979–1991. doi:10.1109/TIP.2017.2666042.
- Kim, T., Lee, K., Ham, S., Park, B., Lee, S., Hong, D., Kim, G. B., Kyung, Y. S., Kim, C.-S., & Kim, N. (2020). Active learning for accuracy enhancement of semantic segmentation with CNN-corrected label curations: Evaluation on kidney segmentation in abdominal CT. *Sci. Rep.*, 10, 1–7.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, .

- Kofler, F., Berger, C., Waldmannstetter, D., Lipkova, J., Ezhov, I., Tetteh, G., Kirschke, J., Zimmer, C., Wiestler, B., & Menze, B. H. (2020). BraTS toolkit: Translating BraTS brain tumor segmentation algorithms into clinical and scientific practice. *Front. Neurosci.*, *14*. doi:10.3389/fnins.2020.00125.
- Li, Z., Kamnitsas, K., & Glocker, B. (2019). Overfitting of neural nets under class imbalance: Analysis and improvements for segmentation. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (pp. 402–410). Springer. doi:10.1007/978-3-030-32248-9_45.
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2017). Focal loss for dense object detection. In *Proc. International Conference on Computer Vision (ICCV)* (pp. 2999–3007). IEEE.
- Linguraru, M. G., Yao, J., Gautam, R., Peterson, J., Li, Z., Linehan, W. M., & Summers, R. M. (2009). Renal tumor quantification and classification in contrast-enhanced abdominal CT. *Pattern Recognit.*, *42*, 1149–1161. doi:10.1016/j.patcog.2008.09.018.
- Liu, L., Cheng, J., Quan, Q., Wu, F.-X., Wang, Y.-P., & Wang, J. (2020). A survey on U-shaped networks in medical image segmentations. *Neurocomputing*, *409*, 244–258. doi:10.1016/j.neucom.2020.05.070.
- Liu, Y., Yang, G., Hosseiny, M., Azadikhah, A., Mirak, S. A., Miao, Q., Raman, S. S., & Sung, K. (2020). Exploring uncertainty measures in

- bayesian deep attentive neural networks for prostate zonal segmentation. *IEEE Access*, 8, 151817–151828. doi:10.1109/ACCESS.2020.3017168.
- Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X., & Martel, A. L. (2021). Loss odyssey in medical image segmentation. *Med. Image Anal.*, (p. 102035).
- Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R. et al. (2014). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging*, 34, 1993–2024. doi:10.1109/TMI.2014.2377694.
- Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proc. Fourth International Conference on 3D Vision (3DV)* (pp. 565–571). IEEE. doi:10.1109/3DV.2016.79.
- Mora Ballestar, L., & Vilaplana, V. (2020a). Brain tumor segmentation using 3d-cnns with uncertainty estimation. *arXiv e-prints*, (pp. arXiv–2009).
- Mora Ballestar, L., & Vilaplana, V. (2020b). MRI brain tumor segmentation and uncertainty estimation using 3D-UNet architectures. *arXiv e-prints*, (pp. arXiv–2012).
- Müller, D., & Kramer, F. (2019). MIScnn: A framework for medical image segmentation with convolutional neural networks and deep learning. *arXiv preprint arXiv:1910.09308*, .
- Nazir, M., Shakil, S., & Khurshid, K. (2021). Role of deep learning in brain

- tumor detection and classification (2015 to 2020): A review. *Comput. Med. Imaging Graph.*, (p. 101940). doi:10.1016/j.compmedimag.2021.101940.
- Pal, N. R., & Pal, S. K. (1993). A review on image segmentation techniques. *Pattern Recognit.*, 26, 1277–1294. doi:10.1016/0031-3203(93)90135-J.
- Ren, T., Wang, H., Feng, H., Xu, C., Liu, G., & Ding, P. (2019). Study on the improved fuzzy clustering algorithm and its application in brain image segmentation. *Appl. Soft Comput.*, 81, 105503. doi:10.1016/j.asoc.2019.105503.
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (pp. 234–241). Springer. doi:10.1007/978-3-319-24574-4_28.
- Roth, H. R., Lu, L., Farag, A., Shin, H.-C., Liu, J., Turkbey, E. B., & Summers, R. M. (2015). Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 556–564). Springer. doi:10.1007/978-3-319-24553-9_68.
- Rueckert, D., & Schnabel, J. A. (2019). Model-based and data-driven strategies in medical image computing. *Proc. IEEE*, 108, 110–124. doi:10.1109/JPROC.2019.2943836.
- Rundo, L., Beer, L., Ursprung, S., Martin-Gonzalez, P., Markowetz, F., Brenton, J. D., Crispin-Ortuzar, M., Sala, E., & Woitek, R. (2020a). Tissue-

- specific and interpretable sub-segmentation of whole tumour burden on CT images by unsupervised fuzzy clustering. *Comput. Biol. Med.*, 120, 103751. doi:10.1016/j.combiomed.2020.103751.
- Rundo, L., Han, C., Nagano, Y. et al. (2019a). USE-Net: incorporating squeeze-and-excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets. *Neurocomputing*, 365, 31–43. doi:10.1016/j.neucom.2019.07.006.
- Rundo, L., Militello, C., Tangherloni, A., Russo, G., Vitabile, S., Gilardi, M. C., & Mauri, G. (2018). NeXt for neuro-radiosurgery: A fully automatic approach for necrosis extraction in brain tumor mri using an unsupervised machine learning technique. *Int. J. Imaging Syst. Technol.*, 28, 21–37. doi:10.1002/ima.22253.
- Rundo, L., Militello, C., Vitabile, S., Casarino, C., Russo, G., Midiri, M., & Gilardi, M. C. (2016). Combining split-and-merge and multi-seed region growing algorithms for uterine fibroid segmentation in MRg-FUS treatments. *Med. Biol. Eng. Comput.*, 54, 1071–1084. doi:10.1007/s11517-015-1404-6.
- Rundo, L., Militello, C., Vitabile, S., Russo, G., Sala, E., & Gilardi, M. C. (2020b). A survey on nature-inspired medical image analysis: a step further in biomedical data integration. *Fundam. Inform.*, 171, 345–365. doi:10.3233/FI-2020-1887.
- Rundo, L., Stefano, A., Militello, C., Russo, G., Sabini, M. G., D’Arrigo, C., Marletta, F., Ippolito, M., Mauri, G., Vitabile, S., & Gilardi, M. C. (2017).

- A fully automatic approach for multimodal PET and MR image segmentation in Gamma Knife treatment planning. *Comput. Methods Programs Biomed.*, 144, 77–96. doi:10.1016/j.cmpb.2017.03.011.
- Rundo, L., Tangherloni, A., Cazzaniga, P., Nobile, M. S., Russo, G., Gilardi, M. C. et al. (2019b). A novel framework for MR image segmentation and quantification by using MedGA. *Comput. Methods Programs Biomed.*, 176, 159–172. doi:10.1016/j.cmpb.2019.04.016.
- Salehi, S. S. M., Erdogmus, D., & Gholipour, A. (2017). Tversky loss function for image segmentation using 3D fully convolutional deep networks. In *Proc. International Workshop on Machine Learning in Medical Imaging* (pp. 379–387). Springer. doi:10.1007/978-3-319-67389-9_44.
- Sánchez-Peralta, L. F., Picón, A., Antequera-Barroso, J. A., Ortega-Morán, J. F., Sánchez-Margallo, F. M., & Pagador, J. B. (2020). Eigenloss: Combined PCA-Based loss function for polyp segmentation. *Mathematics*, 8, 1316.
- Santini, G., Moreau, N., & Rubeaux, M. (2019). Kidney tumor segmentation using an ensembling multi-stage deep learning approach. a contribution to the KiTS19 challenge. *arXiv preprint arXiv:1909.00735*, .
- Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., & Rueckert, D. (2019). Attention gated networks: learning to leverage salient regions in medical images. *Med. Image Anal.*, 53, 197–207. doi:10.1016/j.media.2019.01.012.

- Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Cardoso, M. J. (2017). Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 240–248). Springer. doi:10.1007/978-3-319-67558-9_28.
- Taghanaki, S. A., Zheng, Y., Zhou, S. K., Georgescu, B., Sharma, P., Xu, D., Comaniciu, D., & Hamarneh, G. (2019). Combo loss: Handling input and output imbalance in multi-organ segmentation. *Comput. Med. Imaging Graph.*, 75, 24–33. doi:10.1016/j.compmedimag.2019.04.005.
- Vu, M. H., Nyholm, T., & Löfstedt, T. (2020). Multi-decoder networks with multi-denoising inputs for tumor segmentation. *arXiv preprint arXiv:2012.03684*, .
- Wachinger, C., & Golland, P. (2014). Atlas-based under-segmentation. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (pp. 315–322). Springer. doi:10.1007/978-3-319-10404-1_40.
- Wang, C., He, Y., Qi, X., Zhao, Z., Yang, G., Zhu, X., Zhang, S., Dillenseger, J.-L., & Coatrieux, J.-L. (2019). BiSC-UNet: A fine segmentation framework for kidney and renal tumor. In *2019 Kidney Tumor Segmentation Challenge: KiTS19 (MICCAI 2019)* (pp. 1–8). University of Minnesota Libraries Publishing.
- Wang, S., & Summers, R. M. (2012). Machine learning and radiology. *Med. Image Anal.*, 16, 933–951. doi:10.1016/j.media.2012.02.005.

- Wang, Z., Wang, E., & Zhu, Y. (2020). Image segmentation evaluation: a survey of methods. *Artif. Intell. Rev.*, *53*, 5637–5674. doi:10.1007/s10462-020-09830-9.
- Xie, X., Li, L., Lian, S., Chen, S., & Luo, Z. (2020). Seru: A cascaded se-resnext u-net for kidney and tumor segmentation. *Concurrency and Computation: Practice and Experience*, *32*, e5738.
- Yap, M. H., Pons, G., Marti, J., Ganau, S., Sentis, M., Zwiggelaar, R., Davison, A. K., & Marti, R. (2017). Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE journal of biomedical and health informatics*, *22*, 1218–1226.
- Yeung, M., Sala, E., Schönlieb, C.-B., & Rundo, L. (2021). Advances in artificial intelligence to reduce polyp miss rates during colonoscopy. *arXiv preprint arXiv:2105.07467*, .
- Zhao, W., & Zeng, Z. (2019). Multi scale supervised 3d u-net for kidney and tumor segmentation. *arXiv preprint arXiv:1908.03204*, .
- Zhu, Q., Du, B., & Yan, P. (2019a). Boundary-weighted domain adaptive neural network for prostate mr image segmentation. *IEEE Trans. Med. Imaging*, *39*, 753–763. doi:10.1109/TMI.2019.2935018.
- Zhu, W., Huang, Y., Zeng, L., Chen, X., Liu, Y., Qian, Z., Du, N., Fan, W., & Xie, X. (2019b). AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Med. Phys.*, *46*, 576–589. doi:10.1002/mp.13300.

Zhuang, Z., Li, N., Joseph Raj, A. N., Mahesh, V. G., & Qiu, S. (2019). An RDAU-NET model for lesion segmentation in breast ultrasound images. *PloS One*, *14*, e0221535.