Computerized Medical Imaging and Graphics 95 (2022) 102026

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

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

Michael Yeung [a,b,*,1], Evis Sala [a,c], Carola-Bibiane Schönlieb [d], Leonardo Rundo [a,c,e,*,1]

[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
[e] Department of Information and Electrical Engineering and Applied Mathematics (DIEM), University of Salerno, Fisciano, SA 84084, Italy

ARTICLE INFO

Keywords:
Loss function
Class imbalance
Machine learning
Convolutional neural networks
Medical image segmentation

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 of the two. We propose the Unified Focal loss, a new hierarchical framework that generalises Dice and cross entropy-based losses for handling class imbalance. We evaluate our proposed loss function on five publicly available, class imbalanced medical imaging datasets: CVC-ClinicDB, Digital Retinal Images for Vessel Extraction (DRIVE), 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, across 2D binary, 3D binary and 3D multiclass segmentation tasks, demonstrating that our proposed loss function is robust to class imbalance and consistently outperforms the other loss functions. Source code is available at: https://github.com/mlyg/unified-focal-loss.

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 and 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 and 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 and Summers, 2012), unsupervised clustering (Ren et al., 2019) and atlas-based segmentation (Wachinger and Golland, 2014). In recent years, however, significant progress has been achieved using deep learning (Ker et al., 2018; Rueckert and Schnabel, 2019; Castiglioni et al., 2021).

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 (Y. Liu et al., 2020;

* Correspondence to: Department of Radiology, University of Cambridge, Box 218, Cambridge Biomedical Campus, Cambridge CB2 0QQ, United Kingdom.
E-mail addresses: mjyy2@cam.ac.uk (M. Yeung), es220@medschl.cam.ac.uk (E. Sala), cbs31@cam.ac.uk (C.-B. Schönlieb), lr495@cam.ac.uk, lrundo@unisa.it (L. Rundo).
[1] Present address: Department of Information and Electrical Engineering and Applied Mathematics (DIEM), University of Salerno, Fisciano (SA) 84084, Italy.

https://doi.org/10.1016/j.compmedimag.2021.102026
Received 26 May 2021; Received in revised form 18 November 2021; Accepted 4 December 2021
Available online 13 December 2021

L. Liu et al., 2020; Rundo et al., 2019a)—including the 3D U-Net (Cicek 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 (Kessler et al., 2020; Armanious et al., 2020).

To train deep neural networks, backpropagation updates model parameters in accordance with the optimisation goal defined by the loss function. The cross entropy loss is typically the most widely used loss function in classification problems (L. Liu et al., 2020; Y. Liu et al., 2020) and is applied in the U-Net (Ronneberger et al., 2015), 3D U-Net (Cicek et al., 2016) and SegNet (Badrinarayanan et al., 2017). In contrast, the Attention U-Net (Schlemper et al., 2019) and V-Net (Milletari et al., 2016) leverage the Dice loss, 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 the 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.

Taghanaki et al. (2019) distinguish between input and output imbalance, the former as aforementioned, and the latter referring to classification biases arising 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 loss function comparisons. 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-stripped dataset (Jadon, 2020) (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 compound-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 (Abraham and Khan, 2019). 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 Attention 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 and 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 (Kamnitsas et al., 2017; Havaei 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 labelling, 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. 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.

State-of-the-art solutions typically use unmodified versions of either the Dice loss, cross entropy loss or a combination of the two, and even when using available loss functions for handling class imbalance, such as the Focal Tversky loss, consistently improved performance has not been observed (Ma et al., 2021). Deciding which loss function to use is difficult because there is not only a significant number of loss functions available to choose from, but it is also unclear how each loss function relates to one another. Understanding the relationship between loss functions is the key for providing heuristics to inform loss function choice in class imbalanced situations.

In this paper, we propose the following contributions: .

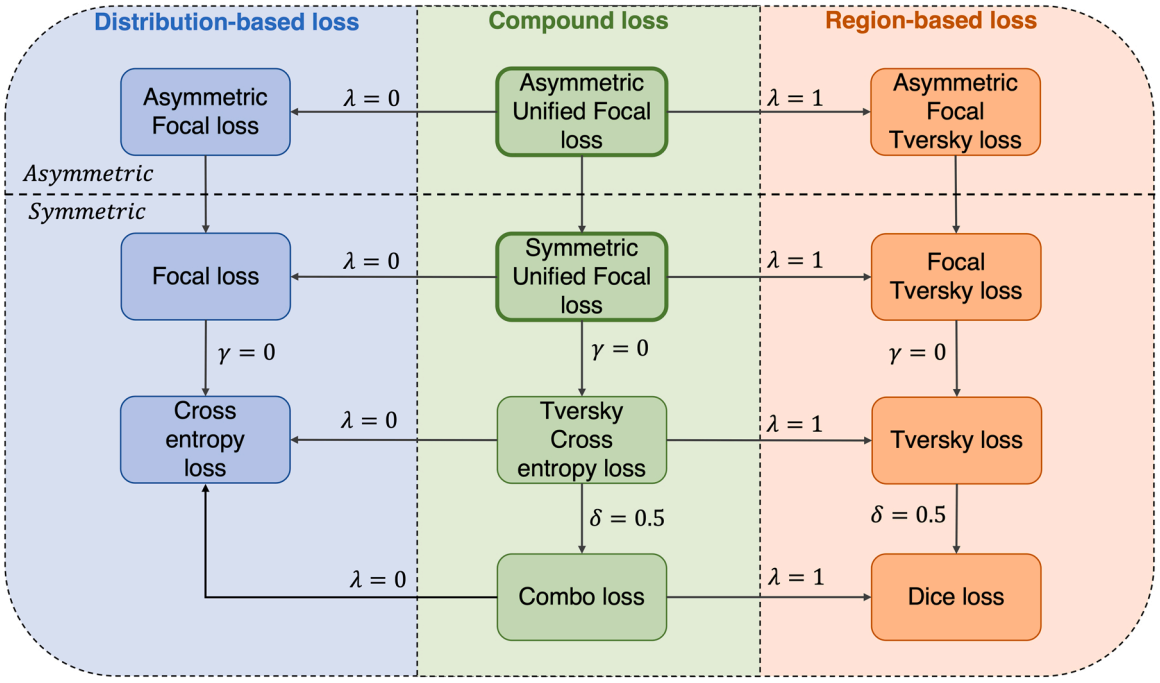

Fig. 1. Our proposed framework unifying various distribution-based, region-based and compound loss functions. The arrows and associated hyperparameter values indicate the required hyperparameter value to set for the preceding loss function in order to recover the resulting loss function.

(a) We summarise and extend the knowledge provided by previous studies that compare loss functions to address the context of class imbalance, by using five class imbalanced datasets with varying degrees of class imbalance, including 2D binary, 3D binary and 3D multi-class segmentation, across multiple imaging modalities.
(b) We define a hierarchical classification of Dice and cross entropy-based loss functions, and use this to derive the Unified Focal loss, that generalises Dice-based and cross entropy-based loss functions for handling class imbalanced datasets.
(c) Our proposed loss function consistently improves segmentation quality over six other related loss functions, is associated with a better recall-precision balance, and is robust to class imbalance.

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 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 ($\mathcal{L}_{\text{BCE}}$) is defined as the following:

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

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 ($\mathcal{L}_{\text{CCE}}$) is computed as:

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

where $y_{i,c}$ uses a one-hot encoding scheme of ground truth labels, $\text{p}_{i,c}$ 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 with 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:

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

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

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

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

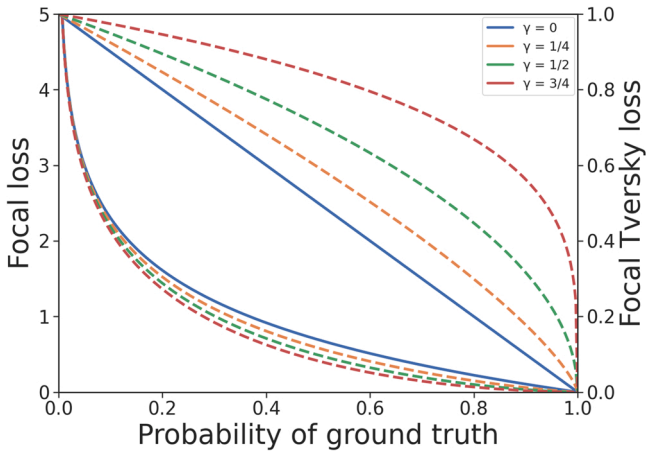

Fig. 2. Effect of changing γ with the Unified Focal loss. The top and bottom group of curves relate to the Focal Tversky loss and Focal loss respectively. The dashed lines represent different values of γ with the modified Focal Tversky loss and modified Focal loss components.

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

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

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

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

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

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

where $\boldsymbol{\alpha}$ is now a vector of class weights, $p_{t,c}$ is a matrix of ground truth probabilities for each class, and $\mathcal{L}_{\mathrm{CCE}}$ 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):

$$\mathrm{DSC} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}. \tag{8}$$

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

$$\mathcal{L}_{\mathrm{DSC}} = 1-\mathrm{DSC}. \tag{9}$$

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).

Even in its most simple formulation, the Dice loss is somewhat adapted to handle class imbalance. However, the Dice loss gradient is inherently unstable, most evident with highly class imbalanced data where gradient calculations involve small denominators (Wong et al., 2018; Bertels et al., 2019).

2.4. Tversky loss

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

$$\mathrm{TI} = \frac{\sum_{i=1}^{N} p_{0i}g_{0i}}{\sum_{i=1}^{N} p_{0i}g_{0i} + \alpha\sum_{i=1}^{N} p_{0i}g_{1i} + \beta\sum_{i=1}^{N} p_{1i}g_{0i}}, \tag{10}$$

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 ($\mathcal{L}_{\mathrm{T}}$) for C classes as:

$$\mathcal{L}_{T} = \sum_{c=1}^{C}(1-\mathrm{TI}). \tag{11}$$

When the Dice loss function is applied to class imbalanced problems, the resulting segmentation often exhibits high precision but low recall scores (Salehi et al., 2017). Assigning a greater weight to false negatives improves recall and results in 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 and 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 ($\mathcal{L}_{\mathrm{FT}}$) as:

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

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 ($\mathcal{L}_{\mathrm{combo}}$) is defined as a weighted sum of the DSC in Eq. 8 and a modified form of the cross entropy loss ($\mathcal{L}_{\mathrm{mCE}}$):

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

where:

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

and $\alpha \in [0,1]$ 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 $\beta > 0.5$ 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 (Taghanaki et al., 2019; Isensee et al., 2018), 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 (Zhu et al., 2019b; Chen et al., 2019). 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

The 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 the 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 ($\mathcal{L}_{\text{HF}}$) is defined as:

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

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 six 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 enhancing or suppressing effect introduced by the focal parameter is applied to all classes, which may affect the 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 modified Focal loss and Focal Tversky loss components, respectively.

Firstly, we replace α in the Focal loss and α and β in the Tversky Index with a common δ parameter to control output imbalance, and reformulate γ to enable simultaneous Focal loss suppression and Focal Tversky loss enhancement, naming these the modified Focal loss ($\mathcal{L}_{\text{mF}}$) and modified Focal Tversky loss ($\mathcal{L}_{\text{mFT}}$), respectively:

$$\mathcal{L}_{mF(p_t)} = \delta(1-p_t)^{1-\gamma} \cdot \mathcal{L}_{\text{BCE}(p,y)}, \tag{16}$$

$$\mathcal{L}_{\text{mFT}} = \sum_{c=1}^{C} (1-\text{mTI})^{\gamma}, \tag{17}$$

where,

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

The symmetric variant of the Unified Focal loss ($\mathcal{L}_{\text{sUF}}$) is therefore defined as:

$$\mathcal{L}_{\text{sUF}} = \lambda \mathcal{L}_{\text{mF}} + (1-\lambda)\mathcal{L}_{\text{mFT}}, \tag{19}$$

where $\lambda \in [0,1]$ and determines the relative weighting of the two losses. By grouping functionally equivalent hyperparameters, the six 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.

Although the Focal loss achieves suppression of the background class, the focal parameter is applied to all classes and therefore the loss contributed by the rare class is also suppressed. Asymmetry enables selective enhancement or suppression using the focal parameter by assigning different losses to each class, and this overcomes both the harmful suppression of the rare class and enhancement of the background class. The modified asymmetric Focal loss ($\mathcal{L}_{\text{maF}}$) 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):

$$\mathcal{L}_{\text{maF}} = -\frac{\delta}{N} y_{i,r} \log\left(p_{t,r}\right) - \frac{1-\delta}{N} \sum_{c \neq r} \left(1-p_{t,c}\right)^{\gamma} \log\left(p_{t,r}\right). \tag{20}$$

In contrast, for the modified Focal Tversky loss, 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 modified asymmetric Focal Tversky loss ($\mathcal{L}_{\text{maFT}}$) as:

$$\mathcal{L}_{\text{maFT}} = \sum_{c \neq r} (1-\text{mTI}) + \sum_{c=r} (1-\text{mTI})^{1-\gamma}. \tag{21}$$

The asymmetric variant of the Unified Focal loss ($\mathcal{L}_{\text{aUF}}$), is therefore defined as:

$$\mathcal{L}_{\text{aUF}} = \lambda \mathcal{L}_{\text{maF}} + (1-\lambda)\mathcal{L}_{\text{maFT}}. \tag{22}$$

The issue of loss suppression associated with the Focal 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. By clarifying the relationship between the loss functions, the Unified Focal loss is much easier to optimise than separately trialling the different loss functions, and it is also more powerful because it is robust to both input and output imbalances. Importantly, given that the Dice loss and cross entropy loss both are efficient operations, and applying the focal parameter adds negligible time complexity, the Unified Focal loss is not expected to significantly increase training time over its component loss functions.

In practice, optimisation of the Unified Focal loss can be further simplified to a single hyperparameter. Given the different effect of the focal parameter on each component loss, the role of λ is partially redundant, and therefore we recommend settings $\lambda = 0.5$, which assigns equal weight to each component loss and is supported by empirical evidence (Taghanaki et al., 2019). Furthermore, we recommend setting $\delta = 0.6$, to correct the Dice loss tendency to produce high precision, low recall segmentations with class imbalance. This is less than $\delta = 0.7$ in the Tversky loss, to account for the effect from the cross entropy-based

Table 1
Details of datasets and training setup used for our experiments.

Dataset	Segmentation	#Images	Input size	#Training	#Validation	#Test	%Foreground
CVC-ClinicDB	Colorectal polyp	612	$288 \times 384 \times 3$	392	98	122	9.3
DRIVE	Retinal vessel	40	$512 \times 512 \times 3$	16	4	20	8.7
BUS2017	Breast tumour	163	$128 \times 128 \times 3$	104	26	33	4.8
BraTS20	Enhancing tumour	342	$96 \times 96 \times 96$	219	55	68	0.2
KiTS19	Kidney / Tumour	204	$80 \times 160 \times 160$	130	33	41	0.8 / 0.2

component. This heuristic reduction of the hyperparameter search space to the single γ parameter makes the Unified Focal loss both powerful and easy to optimise. We provide further empirical evidence behind these heuristics for the Unified Focal loss in the Supplementary Materials.

3. Materials and methods

3.1. Dataset descriptions and evaluation metrics

We select five class imbalanced medical imaging datasets for our experiments: CVC-ClinicDB, DRIVE, BUS2017, KiTS19 and BraTS20. To assess the degree of class imbalance, the percentage of foreground pixels/vowels were calculated per image and averaged over the entire dataset (Table 1).

3.1.1. CVC-ClinicDB dataset

Colonoscopy is the gold-standard screening tool for colorectal cancer, but is associated with significant polyp miss rates, presenting an opportunity to leverage computer-aided systems to support clinicians in reducing the number of polyps missed (Kim et al., 2017). We use the CVC-ClinicDB dataset, which consists of 612 frames containing polyps with image resolution 288×384 pixels, generated from 23 video sequences from 13 different patients using standard colonoscopy interventions with white light (Bernal et al., 2015).

3.1.2. DRIVE dataset

Degenerative retinal diseases display characteristic features on fundoscopy that may be used to aid diagnosis. In particular, retinal vessel abnormalities such as changes in tortuosity or neovascularisation provide important clues for staging and treatment planning. We select the DRIVE dataset (Staal et al., 2004), which consists of 40 coloured fundus photographs obtained from diabetic retinopathy screening in the Netherlands, captured using 8 bits per colour plane of resolution 768×584. 33 photographs display no signs of diabetic retinopathy, while 7 photographs show signs of mild diabetic retinopathy.

3.1.3. 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 with mean image size of 760×570 pixels 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 benign types. The other 53 images depict cancerous masses, with the majority invasive ductal carcinomas.

3.1.4. BraTS20 dataset

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 (Menze et al., 2014; Bakas et al., 2017, 2018). 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. To define a binary segmentation task, we further exclude T1, T2 and FLAIR sequences to focus on gadolinium-enhancing tumour segmentation using the T1-CE sequence (Rundo et al., 2019b; Han et al., 2019), which not only appears to be the most difficult class to segment (Henry et al., 2020), but is also the most clinically relevant for radiation therapy (Rundo et al., 2017, 2018). We further exclude another 27 scans without enhancing tumour regions, leaving 342 scans, with image resolution $240 \times 240 \times 155$ voxels, for use.

3.1.5. 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 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. The image size is 512×512 pixels in the axial plane, with an average of 216 slices in coronal plane. 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.

3.1.6. Evaluation metrics

To assess segmentation accuracy, we use four commonly used metrics (Wang et al., 2020): DSC, Intersection over Union (IoU), recall and precision. DSC is defined in Eq. 8, and IoU, recall and precision are similarly defined per pixel/voxel and according to Eqs. 23, 24 and 25, respectively:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}, \tag{23}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{24}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{25}$$

3.2. Implementation details

All experiments are programmed using Keras with TensorFlow

backend and run on NVIDIA P100 GPUs. We made use of the Medical Image Segmentation with Convolutional Neural Networks (MIScnn) open-source Python library (Müller and Kramer, 2019).

Images from the CVC-ClinicDB, DRIVE and BUS2017 datasets are provided in an anonymised tiff, jpeg and png file formats respectively. For both the KiTS19 and BraTS20 dataset, images and ground truth segmentation masks are provided in an anonymised NIfTI file format. For all datasets, except for the DRIVE dataset which is originally partitioned into 20 training images and 20 testing images, we randomly partitioned each dataset into 80% development and 20% test set, and further divided the development set into 80% training set and 20% validation set. All images were normalised to [0,1] using the z-score. We made use of the 'batchgenerators' library to apply on-the-fly data augmentation with probability 0.15, including: scaling ($0.85 - 1.25\times$), rotation (-15° to $+15^{\circ}$), mirroring (vertical and horizontal axes), elastic deformation ($\alpha \in [0,900]$ and $\sigma \in [9.0, 13.0]$) and brightness ($0.5 - 2\times$).

For 2D binary segmentation, we used the CVC-ClinicDB, DRIVE and BUS2017 datasets and perform full-image analysis, with images resized as described in (Table 1). For 3D binary segmentation, we used the BraTS20 dataset. Here, images were pre-processed, with the skull stripped and images interpolated to the same isotropic resolution of 1 mm^3, and we performed patch-wise analysis using random patches of size of $96 \times 96 \times 96$ voxels for training with patch-wise overlap of $48 \times 48 \times 48$ voxels for inference. For 3D multiclass segmentation, we used the KiTS19 dataset. Hounsfield units (HU) were clipped to [− 79, …, 304] HU and voxel spacing resampled to $3.22 \times 1.62 \times 1.62$ mm^3 (Müller and Kramer, 2019). We performed patch-wise analysis using random patches of size of $80 \times 160 \times 160$ voxels for training and patch-wise overlap of $40 \times 80 \times 80$ voxels for inference.

For the 2D segmentation tasks, we used the original 2D U-Net architecture (Ronneberger et al., 2015), and for the 3D segmentation tasks, we used the 3D U-Net (Cicek et al., 2016). Model parameters were initialised using Xavier initialisation (Glorot and Bengio, 2010), and we added instance normalisation and a final softmax activation layer (Zhou and Yang, 2019). We trained using the stochastic gradient descent optimiser with a batch size of 2 and initial learning rate of 0.1. For convergence criteria, we used ReduceLROnPlateau to reduce the learning rate by 0.1 if the validation loss did not improve after 10 epochs, and the EarlyStopping callback to terminate training if the validation loss did not improve after 20 epochs. Validation loss was evaluated after each epoch, and the model with the lowest validation loss was selected as the final model.

We evaluate the following loss functions: cross entropy loss, Focal loss, Dice loss, Tversky loss, Focal Tversky loss, Combo loss, and symmetric and asymmetric variants of the Unified Focal loss. We used 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 the Focal Tversky loss (Abraham and Khan, 2019) and $\alpha = \beta = 0.5$ for the Combo loss. For the Unified Focal loss, we set $\lambda = 0.5$, $\delta = 0.6$, and performed hyperparameter tuning with $\gamma \in [0.1, 0.9]$ for the 2D segmentation tasks, and set $\gamma = 0.5$ for the 3D segmentation tasks.

To test for statistical significance, we used the Wilcoxon rank-sum test. A statistically significant difference was defined as $p < 0.05$.

4. Experimental results

In this section, we first describe the results from the 2D binary segmentation using the CVC-ClinicDB, DRIVE and BUS2017 datasets, followed by 3D binary segmentation using the BraTS20 dataset, and conclude with 3D multiclass segmentation with the KiTS19 dataset.

Table 2
Results on the CVC-ClinicDB dataset. Values are in the form mean ± 95% confidence intervals. Numbers in boldface denote the highest values for each metric. The best values for the Unified Focal losses are reported from the hyperparameter tuning.

Loss function	DSC	IoU	Precision	Recall
CE	0.889 ± 0.025	0.820 ± 0.029	0.921 ± 0.026	0.878 ± 0.026
Focal	0.868 ± 0.027	0.790 ± 0.031	0.844 ± 0.032	**0.933 ± 0.017**
DSC	0.867 ± 0.029	0.792 ± 0.034	0.895 ± 0.030	0.875 ± 0.031
Tversky	0.874 ± 0.025	0.796 ± 0.030	0.864 ± 0.029	0.909 ± 0.025
Focal Tversky	0.894 ± 0.026	0.831 ± 0.030	0.896 ± 0.026	0.919 ± 0.023
Combo	0.895 ± 0.025	0.831 ± 0.030	**0.927 ± 0.023**	0.885 ± 0.028
Unified Focal (Sym)	**0.909 ± 0.024**	**0.852 ± 0.028**	0.917 ± 0.026	0.919 ± 0.020
Unified Focal (Asym)	**0.909 ± 0.023**	0.851 ± 0.028	0.910 ± 0.026	0.932 ± 0.016

Table 3
Results on the DRIVE dataset. Values are in the form mean ± 95% confidence intervals. Numbers in boldface denote the highest values for each metric. The best values for the Unified Focal losses are reported from the hyperparameter tuning.

Loss function	DSC	IoU	Precision	Recall
CE	0.789 ± 0.008	0.652 ± 0.011	**0.874 ± 0.017**	0.742 ± 0.021
Focal	0.781 ± 0.008	0.642 ± 0.011	0.724 ± 0.019	**0.853 ± 0.016**
DSC	0.799 ± 0.007	0.666 ± 0.009	0.817 ± 0.018	0.787 ± 0.021
Tversky	0.794 ± 0.007	0.658 ± 0.009	0.750 ± 0.019	0.848 ± 0.018
Focal Tversky	0.798 ± 0.007	0.664 ± 0.010	0.765 ± 0.020	0.839 ± 0.019
Combo	0.796 ± 0.007	0.661 ± 0.010	0.836 ± 0.017	0.763 ± 0.021
Unified Focal (Sym)	0.801 ± 0.006	0.669 ± 0.009	0.816 ± 0.018	0.792 ± 0.021
Unified Focal (Asym)	**0.803 ± 0.006**	**0.671 ± 0.008**	0.793 ± 0.018	0.818 ± 0.020

Table 4
Results on the BUS2017 dataset. Values are in the form mean ± 95% confidence intervals. Numbers in boldface denote the highest values for each metric. The best values for the Unified Focal losses are reported from the hyperparameter tuning.

Loss function	DSC	IoU	Precision	Recall
CE	0.751 ± 0.086	0.653 ± 0.090	0.730 ± 0.098	0.851 ± 0.075
Focal	0.603 ± 0.092	0.480 ± 0.090	0.659 ± 0.113	0.710 ± 0.102
DSC	0.767 ± 0.085	0.672 ± 0.088	0.788 ± 0.094	0.808 ± 0.075
Tversky	0.808 ± 0.070	0.716 ± 0.078	0.780 ± 0.081	0.904 ± 0.039
Focal Tversky	0.799 ± 0.081	0.712 ± 0.085	0.758 ± 0.090	0.912 ± 0.062
Combo	0.759 ± 0.087	0.665 ± 0.092	0.746 ± 0.094	0.849 ± 0.073
Unified Focal (Sym)	0.814 ± 0.063	0.716 ± 0.070	0.768 ± 0.076	**0.923 ± 0.027**
Unified Focal (Asym)	**0.824 ± 0.063**	**0.731 ± 0.071**	**0.797 ± 0.074**	0.908 ± 0.037

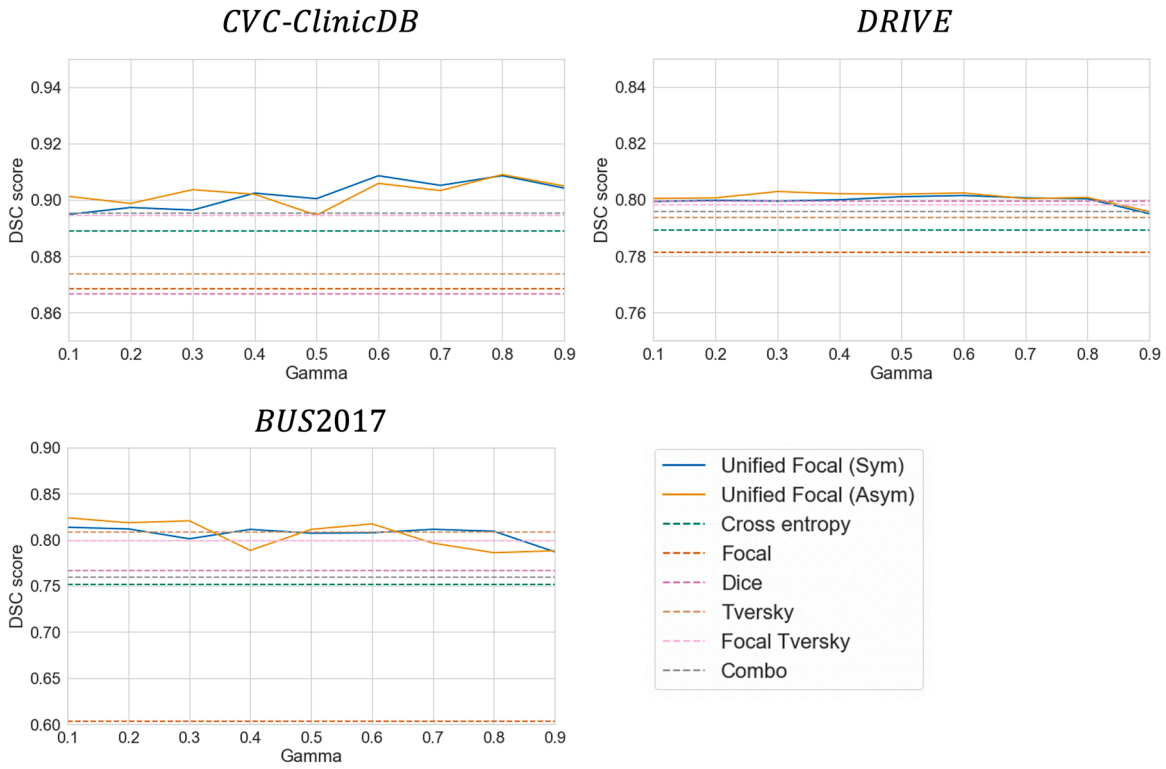

Fig. 3. Evaluating the stability of γ with the Unified Focal loss on the DSC performance for each dataset. The solid lines represent the symmetric and asymmetric variants of the Unified Focal loss, and for reference, the dashed lines represent the DSC performance of the other loss functions.

4.1. 2D binary segmentation

The results for the 2D binary segmentation experiments are shown in Tables 2, 3, and 4.

Across all three datasets, the best performance was consistently observed with the asymmetric variant of the Unified Focal loss, achieving a DSC of 0.909 ± 0.023, 0.803 ± 0.006 and 0.824 ± 0.063 on the CVC-ClinicDB, DRIVE and BUS2017 datasets respectively. This was followed by the symmetric variant of the Unified Focal loss, which achieved the best IoU score of 0.852 ± 0.028 on the CVC-ClinicDB dataset, and comparable DSC scores to the asymmetric variant with a DSC of 0.909 ± 0.024, 0.801 ± 0.006 and 0.814 ± 0.063 on the CVC-ClinicDB, DRIVE and BUS2017 datasets. No statistically significant difference in performance was observed between the two variants of the Unified Focal loss on these datasets. Generally, the worst performance was observed with cross entropy-based loss functions, with the Focal loss performing significantly worse than the cross entropy loss on the CVC-ClinicDB ($p = 0.04$) and BUS2017 ($p = 0.004$) datasets, and significantly worse than the asymmetric variant of the Unified Focal loss across the three datasets (CVC-ClinicDB: $p = 2\times10^{-6}$, DRIVE: $p = 110^{-4}$ and BUS2017: $p = 5\times10^{-5}$). No significant differences were observed between the Dice-based losses.

To evaluate the performance stability of the γ hyperparameter, we display the DSC performance for each value of $\gamma \in [0.1, 0.9]$ for the three datasets in Fig. 3.

For both the symmetric and asymmetric variants, the Unified Focal loss displays consistently strong performance across the range of $\gamma \in [0.1, 0.9]$. This is most evident with the CVC-ClinicDB dataset, where improved performance over the other loss functions is observed across the entire range of hyperparameter values. The worst performance occurred at high values such as $\gamma = 0.9$, while middle values, such as $\gamma = 0.5$, provided robust performance benefits across datasets.

To enable a qualitative comparison, example segmentations are shown in Fig. 4.

There is a clear visual difference between the segmentations generated using different loss functions. The segmentations from cross entropy-based loss functions are associated with a greater proportion of false negative predictions compared to the Dice-based loss functions. The highest quality segmentations were produced by the compound loss functions, with the best segmentations produced using the Unified Focal loss. This is particularly clear with the asymmetric variant of the Unified Focal loss in the CVC-ClinicDB example.

4.2. 3D binary segmentation

The results for the 3D binary segmentation experiments are shown in Tables 5.

The best performance was observed with the Unified Focal loss, specifically the asymmetric variant with a DSC of 0.787 ± 0.049, IoU of 0.683 ± 0.050, precision of 0.795 ± 0.048 and recall of 0.800 ± 0.056. This was followed by the symmetric variant of the Unified Focal loss, with no significant difference between the two loss functions. In contrast, the asymmetric Unified Focal loss displayed significantly improved performance compared to all the other loss functions (cross entropy loss: $p = 0.02$, Focal loss: $p = 0.03$, Dice loss: $p = 6\times10^{-10}$, Tversky loss: $p = 5\times10^{-11}$, Focal Tversky loss: $p = 0.02$ and Combo loss: $p = 1\times10^{-4}$).

Axial slices taken from an example segmentation are shown in Fig. 5.

From the results, there is a clear recall bias on this dataset, and this is reflected by the proportion of false positive predictions with each segmentation prediction. The compound loss functions displayed the best recall-precision balance, and this is evident by the significantly reduced false positive predictions visible in the segmentations produced using these loss functions.

4.3. 3D multiclass segmentation

The results for the 3D multiclass segmentation experiments are shown in Tables 6.

The Unified Focal loss achieves the best performance, with DSC of

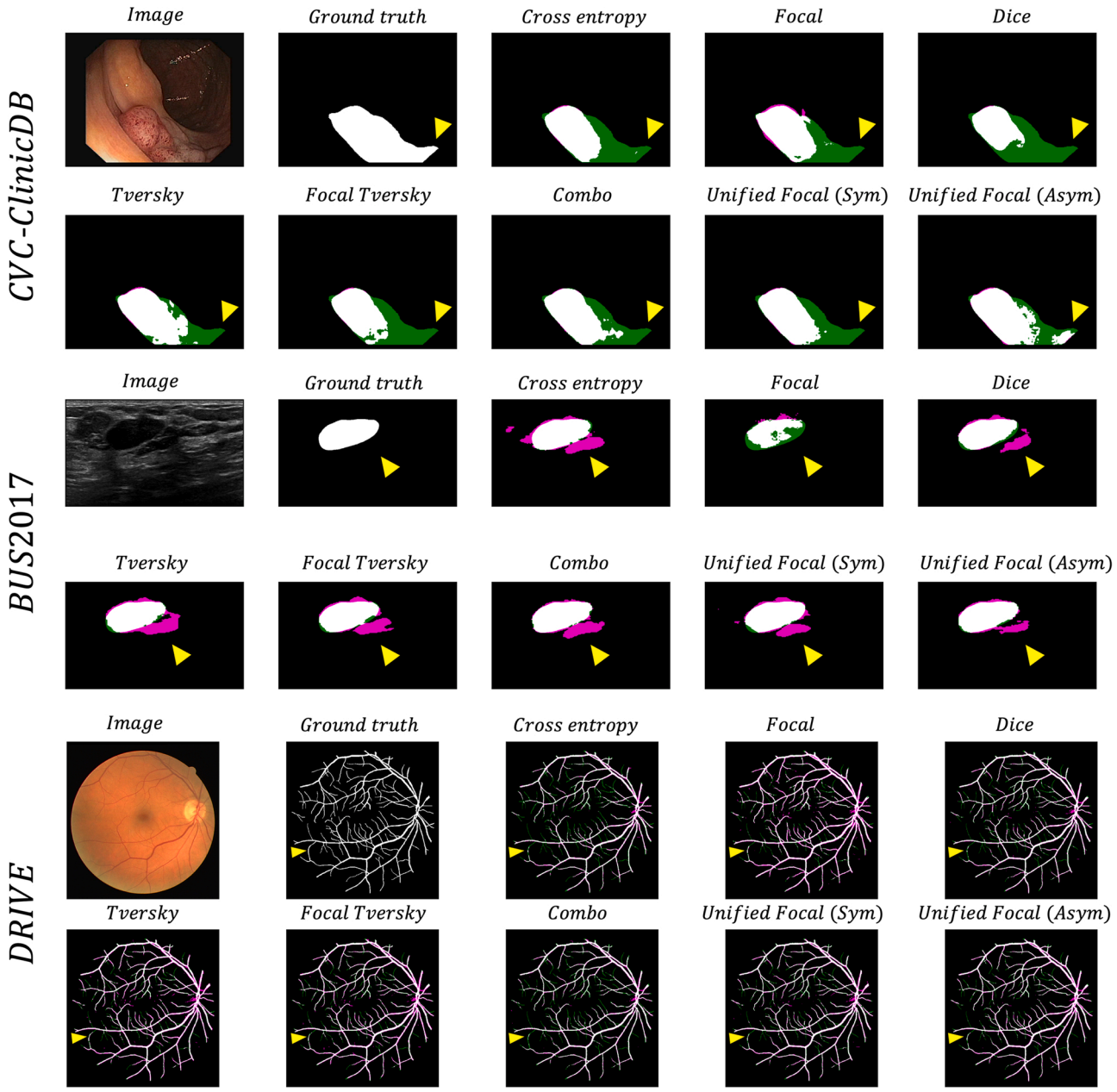

Fig. 4. Example segmentations, for each loss function for each of the three datasets. The image and ground truth are provided for reference. The false positive are highlighted in magenta, and the false negatives are highlighted in green. The yellow arrows highlight example areas where segmentation quality differs.

0.943 ± 0.011 and 0.634 ± 0.079 with the asymmetric variant, and DSC of 0.943 ± 0.013 and 0.614 ± 0.079 with the symmetric variant, for the kidney and kidney tumour segmentation respectively. For kidney segmentation, the asymmetric variant of the Unified Focal loss achieves significantly improved performance compared to the cross entropy loss ($p = 0.03$), Focal loss ($p = 0.004$), Tversky loss ($p = 0.001$), and Focal Tversky loss ($p = 0.03$). The worst performance for kidney segmentation was observed using Dice-based losses, with the Tversky loss followed by the Focal Tversky loss. In contrast, the worst performance for kidney tumour segmentation was observed using cross entropy-based losses, with significantly better DSC performance using the Dice loss compared to the cross entropy loss ($p = 0.01$). For kidney tumour segmentation, the asymmetric variant of the Unified Focal loss achieves significantly better DSC performance compared to the cross entropy loss ($p = 6\times10^{-5}$), Focal loss ($p = 1\times10^{-4}$), Dice loss ($p < 0.05$) and Tversky loss ($p = 4\times10^{-4}$).

Axial slices taken from an example segmentation are shown in Fig. 5.

While the kidneys are generally well segmented with only subtle differences between the loss functions, the tumour segmentations vary considerably in quality. The low tumour recall scores with the cross entropy-based loss functions are reflected in the segmentations, where the boundary between the tumour and kidney are shifted in favour of kidney prediction. The highest quality segmentation is observed with the Unified Focal loss, with visibly the most accurate contour of the tumour.

5. Discussion and conclusions

In this study, we proposed a new hierarchical framework to encompass various Dice and cross entropy-based loss functions, and used this to derive the Unified Focal loss, which generalises Dice and cross entropy-based loss functions for handling class imbalance. We compared the Unified Focal loss against six other loss functions on five class imbalanced datasets with varying degrees of class imbalance (CVC-

Table 5
Results on the BraTS20 dataset. Values are in the form mean ± 95% confidence intervals. Numbers in boldface denote the highest values for each metric. $\gamma = 0.5$ for the Unified Focal losses.

Loss function	DSC	IoU	Precision	Recall
CE	0.716 ± 0.061	0.604 ± 0.058	**0.826 ± 0.055**	0.682 ± 0.068
Focal	0.738 ± 0.055	0.623 ± 0.052	0.706 ± 0.055	0.805 ± 0.063
DSC	0.620 ± 0.054	0.482 ± 0.049	0.629 ± 0.064	0.683 ± 0.058
Tversky	0.580 ± 0.056	0.443 ± 0.051	0.525 ± 0.061	0.740 ± 0.062
Focal Tversky	0.747 ± 0.052	0.632 ± 0.050	0.717 ± 0.056	**0.822 ± 0.053**
Combo	0.686 ± 0.056	0.563 ± 0.055	0.668 ± 0.063	0.757 ± 0.056
Unified Focal (Sym)	0.780 ± 0.049	0.673 ± 0.049	0.803 ± 0.049	0.792 ± 0.056
Unified Focal (Asym)	**0.787 ± 0.049**	**0.683 ± 0.050**	0.795 ± 0.048	0.800 ± 0.056

ClinicDB, DRIVE, BUS2017, BraTS20 and KiTS19) involving 2D binary, 3D binary and 3D multiclass segmentation. The Unified Focal loss consistently achieved the highest DSC and IoU scores across the five datasets, with slightly better performance observed using the asymmetric variant over the symmetric variant. We demonstrated that the optimisation of the Unified Focal loss can be simplified to tuning a single γ hyperparameter, which we observed is stable and therefore easy to optimise (Fig. 3).

The significant difference in model performance using different loss functions highlights the importance of the loss function choice in class imbalanced image segmentation tasks. Most noticeable is the poor performance using distribution-based loss functions with the segmentation of the kidney tumour class on the highly class imbalanced KiTS19 dataset (Table 6). This susceptibility to class imbalance is expected given the greater representation of classes occupying a larger region in cross entropy-based losses. Generally, the Dice-based and compound loss functions performed better with class imbalanced data, but one notable exception was the BraTS20 dataset, where the Dice loss and Tversky loss performed significantly worse than the other loss functions. This likely reflects the unstable gradient issue associated with the Dice loss, resulting in suboptimal convergence and resulting poor performance. Compound loss functions such as the Combo loss and Unified Focal loss

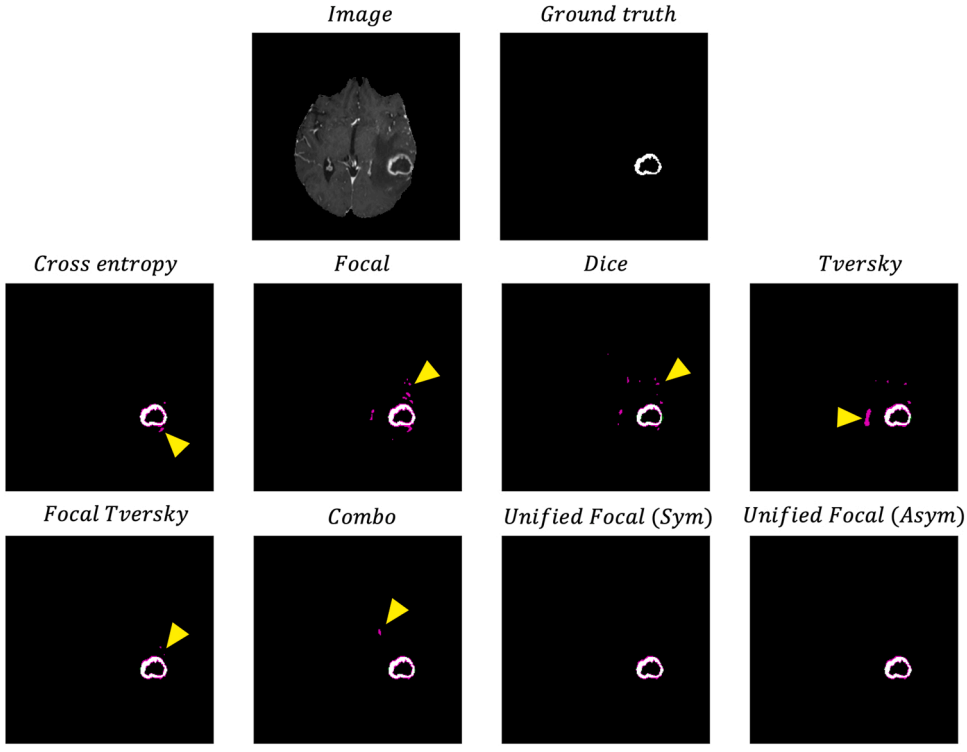

Fig. 5. Axial slice from an example segmentation for each loss function for the BraTS20 dataset. The image and ground truth are provided for reference. The false positive are highlighted in magenta, and the false negatives are highlighted in green. The yellow arrows highlight example areas where segmentation quality differs.

Table 6
Results on the KiTS19 dataset. Values are in the form mean ± 95% confidence intervals. Numbers in boldface denote the highest values for each metric. $\gamma = 0.5$ for the Unified Focal losses.

	Kidney				Tumour			
Loss function	DSC	IoU	Precision	Recall	DSC	IoU	Precision	Recall
CE	0.928 ± 0.016	0.869 ± 0.024	0.933 ± 0.010	0.928 ± 0.026	0.336 ± 0.107	0.262 ± 0.091	0.585 ± 0.133	0.303 ± 0.103
Focal	0.928 ± 0.011	0.868 ± 0.019	0.907 ± 0.011	**0.952 ± 0.018**	0.349 ± 0.110	0.276 ± 0.094	0.556 ± 0.136	0.317 ± 0.105
DSC	0.931 ± 0.015	0.875 ± 0.024	0.936 ± 0.012	0.930 ± 0.024	0.536 ± 0.074	0.402 ± 0.069	0.594 ± 0.089	0.585 ± 0.077
Tversky	0.914 ± 0.017	0.846 ± 0.026	0.894 ± 0.023	0.940 ± 0.016	0.420 ± 0.087	0.308 ± 0.075	0.411 ± 0.097	0.616 ± 0.087
Focal Tversky	0.926 ± 0.013	0.864 ± 0.022	0.909 ± 0.017	0.946 ± 0.017	0.520 ± 0.089	0.401 ± 0.081	0.513 ± 0.095	0.619 ± 0.095
Combo	0.935 ± 0.015	0.881 ± 0.024	**0.954 ± 0.008**	0.920 ± 0.025	0.554 ± 0.081	0.425 ± 0.074	0.616 ± 0.091	0.586 ± 0.088
Unified Focal (Sym)	**0.943 ± 0.013**	**0.894 ± 0.020**	0.949 ± 0.007	0.940 ± 0.021	0.614 ± 0.079	0.488 ± 0.077	**0.667 ± 0.082**	0.657 ± 0.084
Unified Focal (Asym)	**0.943 ± 0.011**	**0.894 ± 0.019**	0.942 ± 0.015	0.946 ± 0.014	**0.634 ± 0.079**	**0.510 ± 0.078**	0.656 ± 0.083	**0.695 ± 0.084**

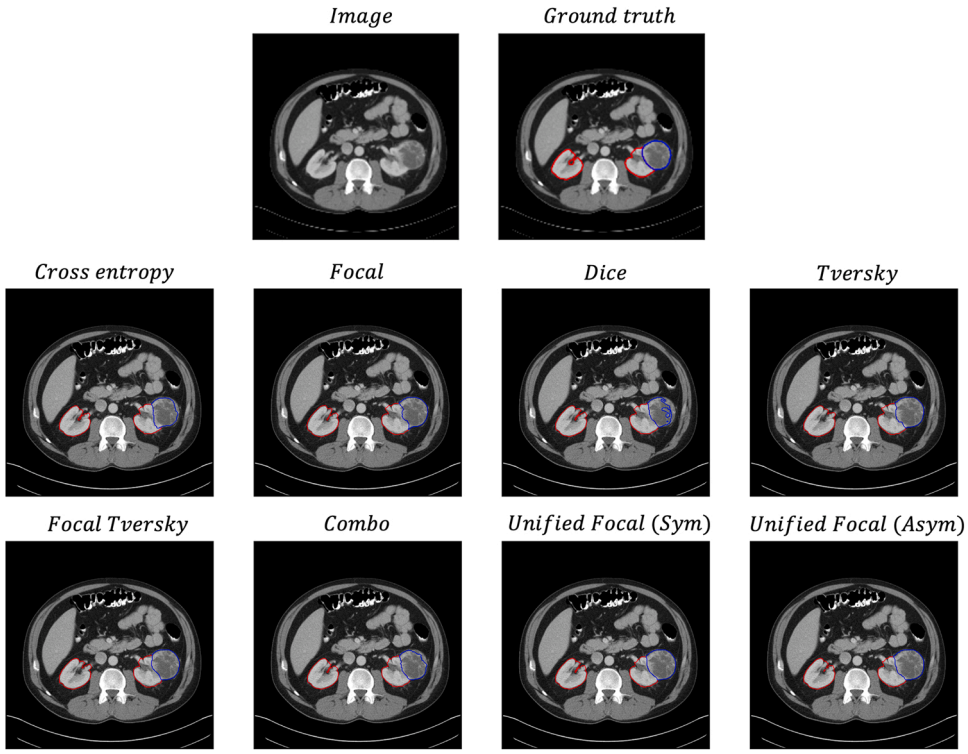

Fig. 6. Axial slice from an example segmentation for each loss function for the KiTS19 dataset. The image and ground truth are provided for reference. The red contour corresponds to the kidneys, and the blue contour to the tumour.

performed consistently well across datasets, benefiting from the increased gradient stability with the cross entropy-based component, and the robustness to class imbalance from the Dice-based component. The qualitative assessment correlates with the performance metrics, with the highest quality segmentations observed using the Unified Focal loss (Fig. 4–6). As expected, no difference in training time was observed between any of the loss functions used in these experiments.

There are several limitations associated with our study. Firstly, we have restricted our framework and comparisons to include only a subset of the most popular variants of the Dice-based and cross entropy-based loss functions. However, it should be noted that the Unified Focal loss also generalises other loss functions that were not included, such as the DiceFocal loss (Zhu et al., 2019b) and Asymmetric similarity loss (Hashemi et al., 2018). One major class of loss functions that were not included were boundary-based loss functions (Kervadec et al., 2019; Zhu et al., 2019a), which are 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. Secondly, it is not immediately clear how to optimise the γ hyperparameter in multiclass segmentation tasks. In our experiments, we treated both the kidney and the kidney tumour as the rare class and assigned $\gamma = 0.5$. Better performance may be observed by assigning different γ values to each class, given that for example the kidney class in the KiTS19 dataset is four times more prevalent than the tumour class. However, we still achieved improved performance using the Unified Focal loss over the other loss functions even with this simplification.

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.

CRediT authorship contribution statement

Michael Yeung: Conceptualisation, Methodology, Software, Validation, Formal analysis, Investigation, Writing – Original Draft, Writing – Review & Editing, Visualization. **Evis Sala:** Conceptualisation, Methodology, Investigation, Writing – Original Draft, Writing – Review & Editing, Funding acquisition. **Carola-Bibiane Schönlieb:** Conceptualisation, Methodology, Formal analysis, Investigation, Writing – Review & Editing, Funding acquisition. **Leonardo Rundo:** Conceptualisation, Methodology, Formal analysis, Investigation, Writing – Original Draft, Writing – Review & Editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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, UK [215733/Z/19/Z]. 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, European Union Horizon 2020 research and innovation programmes under the Marie Skodowska-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).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.compmedimag.2021.102026.

References

Abraham, N., Khan, N.M., 2019. A novel focal Tversky loss function with improved attention U-Net for lesion segmentation. Proc. 16th International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 683–687. https://doi.org/10.1109/ISBI.2019.8759329.

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 https://doi.org/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. https://doi.org/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. https://doi.org/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:1811.02629,.

Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F., 2015. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Comput. Med. Imaging Graph. 43, 99–111.

Bertels, J., Robben, D., Vandermeulen, D., Suetens, P., 2019. Optimization with soft dice can lead to a volumetric bias. International MICCAI Brainlesion Workshop. Springer, pp. 89–97.

Castiglioni, I., Rundo, L., Codari, M., DiLeo, G., Salvatore, C., Interlenghi, M., et al., 2021. AI applications to medical images: From machine learning to deep learning. Phys. Med. 83, 9–24. https://doi.org/10.1016/j.ejmp.2021.02.006.

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. Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 447–456. https://doi.org/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. https://doi.org/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. Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 424–432. https://doi.org/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. https://doi.org/10.1109/TMI.2006.880587.

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. Proc. International MICCAI Brainlesion Workshop. Springer, pp. 64–76. https://doi.org/10.1007/978-3-319-75238-9_6.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks.In Proceedings of the thirteenth international conference on artificial intelligence and statistics (249–256). JMLR Workshop and Conference Proceedings.

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. https://doi.org/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. https://doi.org/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 https://doi.org/10.1016/j.media.2020.101821.

Heller, N., Sathianathen, 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: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:2011.01045,.

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. https://doi.org/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:1809.10486,.

Jadon, S., 2020. A survey of loss functions for semantic segmentation. Proc. Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). IEEE, pp. 1–7. https://doi.org/10.1109/CIBCB48159.2020.9277638.

Jadon, Shruti, 2020. A survey of loss functions for semantic segmentation. 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). In this issue.

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. https://doi.org/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) (.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 https://doi.org/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. https://doi.org/10.1109/TIP.2017.2666042.

Kim, N.H., Jung, Y.S., Jeong, W.S., Yang, H.-J., Park, S.-K., Choi, K., Park, D.I., 2017. Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies. Intest. Res. 15, 411.

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.

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. https://doi.org/10.3389/fnins.2020.00125.

Li, Z., Kamnitsas, K., Glocker, B., 2019. Overfitting of neural nets under class imbalance: Analysis and improvements for segmentation. Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 402–410. https://doi.org/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. Proc. International Conference on Computer Vision (ICCV). IEEE, pp. 2999–3007.

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. https://doi.org/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. https://doi.org/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. https://doi.org/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., 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. https://doi.org/10.1109/TMI.2014.2377694.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-Net: fully convolutional neural networks for volumetric medical image segmentation. Proc. Fourth International Conference on 3D Vision (3DV). IEEE, pp. 565–571. https://doi.org/10.1109/3DV.2016.79.

Müller, D., & Kramer, F. , 2019. MIScnn: A framework for medical image segmentation with convolutional neural networks and deep learning. 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., 101940 https://doi.org/10.1016/j.compmedimag.2021.101940.
Pal, N.R., Pal, S.K., 1993. A review on image segmentation techniques. Pattern Recognit. 26, 1277–1294. https://doi.org/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 https://doi.org/10.1016/j.asoc.2019.105503.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional networks for biomedical image segmentation. Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 234–241. https://doi.org/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. International conference on medical image computing and computer-assisted intervention. Springer, pp. 556–564. https://doi.org/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. https://doi.org/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 https://doi.org/10.1016/j.compbiomed.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. https://doi.org/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. https://doi.org/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 MRgFUS treatments. Med. Biol. Eng. Comput. 54, 1071–1084. https://doi.org/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. https://doi.org/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 Prog. Biomed. 144, 77–96. https://doi.org/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 Prog. Biomed. 176, 159–172. https://doi.org/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. Proc. International Workshop on Machine Learning in Medical Imaging. Springer, pp. 379–387. https://doi.org/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.
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. https://doi.org/10.1016/j.media.2019.01.012.
Staal, J., Abràmoff, M.D., Niemeijer, M., Viergever, M.A., Van Ginneken, B., 2004. Ridge-based vessel segmentation in color images of the retina. IEEE Trans. Med. Imaging 23, 501–509.
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. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 240–248. https://doi.org/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. https://doi.org/10.1016/j.compmedimag.2019.04.005.
Wachinger, C., Golland, P., 2014. Atlas-based under-segmentation. Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 315–322. https://doi.org/10.1007/978-3-319-10404-1_40.
Wang, S., Summers, R.M., 2012. Machine learning and radiology. Med. Image Anal. 16, 933–951. https://doi.org/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. https://doi.org/10.1007/s10462-020-09830-9.
Wong, K.C., Moradi, M., Tang, H., Syeda-Mahmood, T., 2018. 3d segmentation with exponential logarithmic loss for highly unbalanced object sizes. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 612–619.
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 J. Biomed. Health Inform. 22, 1218–1226.
Yeung, M., Sala, E., Schönlieb, C.-B., Rundo, L., 2021. Focus U-Net: a novel dual attention-gated CNN for polyp segmentation during colonoscopy. Comput. Biol. Med. 137, 104815 https://doi.org/10.1016/j.compbiomed.2021.104815.
Zhou, X.-Y., Yang, G.-Z., 2019. Normalization in training U-Net for 2-D biomedical semantic segmentation. IEEE Robot. Autom. Lett. 4, 1792–1799.
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. https://doi.org/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. https://doi.org/10.1002/mp.13300.
Zhuang, Z., Li, N., JosephRaj, A.N., Mahesh, V.G., Qiu, S., 2019. An RDAU-NET model for lesion segmentation in breast ultrasound images. PLoS One 14, e0221535.