




Calibrating the Dice Loss to Handle Neural Network Overconfidence for Biomedical Image Segmentation

Michael Yeung^{1,2,6}  · Leonardo Rundo^{1,3,4} · Yang Nan² · Evis Sala^{1,3} · Carola-Bibiane Schönlieb⁵ · Guang Yang²

Received: 20 January 2022 / Revised: 30 October 2022 / Accepted: 31 October 2022
© The Author(s) 2022

Abstract

The Dice similarity coefficient (DSC) is both a widely used metric and loss function for biomedical image segmentation due to its robustness to class imbalance. However, it is well known that the DSC loss is poorly calibrated, resulting in overconfident predictions that cannot be usefully interpreted in biomedical and clinical practice. Performance is often the only metric used to evaluate segmentations produced by deep neural networks, and calibration is often neglected. However, calibration is important for translation into biomedical and clinical practice, providing crucial contextual information to model predictions for interpretation by scientists and clinicians. In this study, we provide a simple yet effective extension of the DSC loss, named the DSC++ loss, that selectively modulates the penalty associated with overconfident, incorrect predictions. As a standalone loss function, the DSC++ loss achieves significantly improved calibration over the conventional DSC loss across six well-validated open-source biomedical imaging datasets, including both 2D binary and 3D multi-class segmentation tasks. Similarly, we observe significantly improved calibration when integrating the DSC++ loss into four DSC-based loss functions. Finally, we use softmax thresholding to illustrate that well calibrated outputs enable tailoring of recall-precision bias, which is an important post-processing technique to adapt the model predictions to suit the biomedical or clinical task. The DSC++ loss overcomes the major limitation of the DSC loss, providing a suitable loss function for training deep learning segmentation models for use in biomedical and clinical practice. Source code is available at <https://github.com/mlyg/DicePlusPlus>.

Keywords Biomedical imaging · Image segmentation · Machine learning · Cost function

✉ Michael Yeung
michael.yeung21@imperial.ac.uk

Leonardo Rundo
lrundo@unisa.it

Yang Nan
y.nan20@imperial.ac.uk

Evis Sala
es220@medschl.cam.ac.uk

Carola-Bibiane Schönlieb
cbs31@cam.ac.uk

Guang Yang
g.yang@imperial.ac.uk

² National Heart & Lung Institute, Imperial College London, Dovehouse St, London SW3 6LY, UK

³ Cancer Research UK Cambridge Centre, University of Cambridge, Robinson Way, Cambridge CB2 0RE, UK

⁴ Department of Information and Electrical Engineering and Applied Mathematics (DIEM), University of Salerno, Fisciano, Salerno 84084, Italy

⁵ Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Wilberforce Rd, Cambridge CB3 0WA, UK

⁶ Department of Computing, Imperial College London, London, UK

¹ Department of Radiology, University of Cambridge, Hills Rd, Cambridge CB2 0QQ, UK

Introduction

Image segmentation describes a per-pixel classification task, involving partitioning an image into semantic regions based on regional pixel characteristics [1]. However, class imbalance is frequently observed in biomedical image segmentation tasks, where objects, such as tumours or cell nuclei often occupy a small area relative to the background tissue [2]. This can hinder per-pixel classification accuracy and could result in poor segmentation results on biomedical images. To evaluate segmentation quality, the two most popular metrics used are the Dice similarity coefficient (DSC) and the Jaccard Index. Both metrics measure spatial overlap and are therefore generally robust to class imbalance [3, 4].

To incorporate automated image segmentation methods for biomedical applications, not only is segmentation quality important, but it is necessary that predictions are well calibrated [5–7]. Calibration measures how similar the probabilities assigned to model predictions reflect the real-world underlying uncertainty. In the context of medical image segmentation, a well calibrated model is expected to output predictions with probabilities that match the confidence of an expert human annotator

performing manual delineation, or similarly, to match the distribution of segmentations produced by a group of expert annotators.

Importantly, even small differences in imaging hardware or image acquisition parameters may lead to a domain shift that could significantly affect neural network performance, and without proper calibration, result in overconfident predictions that could provide false reassurance and cause potential harm [8]. Calibration also provides crucial contextual information to the corresponding segmentation output, which is useful for guiding clinical decision making, such as planning for surgical resection or image-guided interventions.

The cross entropy (CE) loss is the most widely used loss function for classification, favoured because of its well calibrated prediction outputs, but it is susceptible to class imbalance and regularly underperforms in these situations, particularly when very small segmentation targets are involved [9, 10]. In contrast, the DSC loss is, similar to its respective evaluation metric, robust to class imbalance and has been successfully applied to a variety of biomedical image segmentation tasks [11–13]. However, it is well known that optimising the DSC loss results in poorly calibrated, overconfident predictions (Fig. 1) [6, 14, 15].

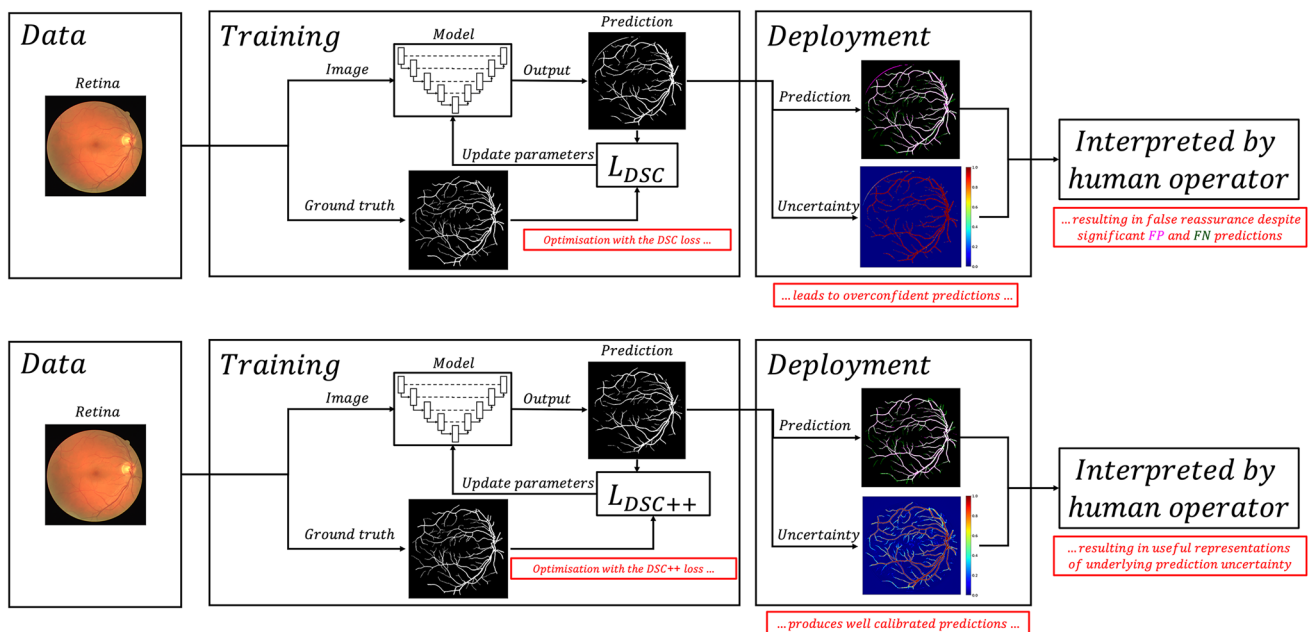


Fig. 1 Deep learning-based biomedical image segmentation pipeline. During training, model predictions are compared to ground truth annotations, with model parameters iteratively updated based on the optimisation goal defined by the loss function. During deployment, the model is used for inference, generating a segmentation mask and associated softmax values, which are accessible by the scientist or clinician. Top: Using the DSC loss results in overconfident model pre-

dictions, demonstrated by the extreme softmax values illustrated by the heatmap, despite significant false positive (FP) and false negative (FN) predictions. Bottom: In contrast, using the DSC++ loss produces well calibrated predictions that, with a lower certainty, capture the more difficult-to-segment small-diameter retinal vessels. The colours corresponding to the softmax values are shown by the colour-bar on the right

The dichotomy between the CE loss, which provides well calibrated but often suboptimal segmentations, and the DSC loss—which produces higher quality segmentations but results in poorly calibrated predictions—suggests that neither loss function is appropriate for clinical use.

To overcome these challenges in biomedical image segmentation, considerable research has focused on either modifying the CE loss to improve robustness to class imbalance, or improving the calibration of networks trained using the DSC loss. The Focal loss is a variant of the CE loss that addresses the issue of class imbalance by down-weighting the contribution of easy examples enabling learning of harder examples [16]. Similarly, the exponentially weighted CE loss down-weights correctly predicted samples, but is better suited for smaller degrees of class imbalance [17]. In contrast to directly modifying the CE loss, approaches to improve DSC loss calibration generally focus on modifying the network or applying post hoc calibration. Performing dropout during inference, known as Monte Carlo (MC) dropout, was shown to approximate Bayesian neural networks and improve calibration [6, 7, 18]. Other network modifications where improved calibration was observed include Platt scaling, which involves fitting a logistic regression model using model outputs, as well as auxiliary networks, which are a generalised version of Platt scaling that instead uses a convolutional layer [19, 20]. Avoiding network modifications, deep ensembles involve averaging predictions from multiple, randomly initialised networks, outperforming MC dropout for both performance and calibration [6, 7, 21]. However, ensembling of multiple networks is not only computationally expensive to train, but significantly increases inference time and is therefore of limited use in real-time applications. Finally, improved calibration was observed by initially training a network using the DSC loss, followed by fine-tuning using the CE loss [7].

Despite various modifications to the CE loss, the segmentation performance remains generally worse than using the DSC loss [10]. In contrast, while the modifications to improve the DSC loss calibration result in comparable calibration to the CE loss, they require pipeline modifications, limiting uptake by the research community as well as clinical applicability.

The main contributions of this work may be summarised as follows:

1. We identify the reason for the poor calibration observed with networks trained using the DSC loss, and provide a reformulation, named the DSC++ loss, which directly addresses the issue.
2. We demonstrate significantly improved calibration using the DSC++ loss over the DSC loss, measured using the

negative log likelihood (NLL) and Brier score, across six well-validated open-source datasets, including 2D binary and 3D multi-class segmentation tasks.

3. We demonstrate that the DSC++ loss may be readily incorporated to improve the calibration of other DSC-based loss functions.
4. We combine softmax thresholding with the DSC++ loss to enable tailoring of the recall-precision bias for the biomedical or clinical task.

Material and Methods

In this section, we first introduce the CE loss and its variant, the Focal loss, followed by the DSC loss. We then identify the cause of the poor calibration using the DSC loss, and use this to derive the DSC++ loss. After introducing softmax thresholding, the section finally concludes with details on the experimental setup and implementation.

CE Loss

CE measures the difference between two probability distributions y and p . The CE loss is among the most widely used loss function in machine learning, and in the context of image segmentation, y and p represent the true and predicted distributions over class labels for a given pixel, respectively. The CE loss, (\mathcal{L}_{CE}), is defined as:

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

where $y_{i,c}$ uses a one-hot encoding scheme corresponding to the ground truth labels, $p_{i,c}$ is a matrix of predicted values generated by the model for each class, and where indices i and c iterate over all pixels and classes, respectively. The CE loss is a strictly proper scoring rule, superficially equivalent to the NLL, and therefore yields consistent probabilistic predictions [22].

Focal Loss

The Focal loss (\mathcal{L}_F) is an extension of the cross entropy loss developed to address the issue of class imbalance in classification tasks [16].

The Focal loss uses a modulating factor γ to reduce the contribution of easy examples to the overall loss:

$$\mathcal{L}_F = \alpha(1 - (p_{i,c}))^\gamma \cdot \mathcal{L}_{CE}, \quad (2)$$

where α is a vector of class weights, $p_{i,c}$ is a matrix of ground truth probabilities for each class, and \mathcal{L}_{CE} is the cross entropy loss as defined in Eq. (1). The Focal loss is equivalent to the cross entropy loss when $\gamma = 1$.

DSC Loss

CE and the Focal loss are based on pixel-wise error, and therefore in class imbalanced situations, using the CE-based losses results in over-representation of larger objects in the loss, and consequently under-segmentation of smaller objects. Often the segmentation target in biomedical imaging tasks occupies a small area relative to the size of the image, limiting its use as a segmentation quality metric or loss function [10].

In contrast, the DSC is a spatial overlap index and is therefore robust to class imbalance, and is defined as:

$$\text{DSC} = \frac{1}{C} \sum_{c=1}^C \frac{2 \sum_{i=1}^N p_{i,c} y_{i,c}}{\sum_{i=1}^N p_{i,c} + \sum_{i=1}^N y_{i,c}}, \quad (3)$$

where the DSC loss (\mathcal{L}_{DSC}) is:

$$\mathcal{L}_{\text{DSC}} = 1 - \text{DSC}. \quad (4)$$

DSC++ Loss

The optimisation goal, for both the CE and the DSC loss, is for the neural network to produce confident, and correct, predictions matching the ground truth label. However, neural network overconfidence is a well known phenomenon associated with the DSC loss, but not with the CE loss. To understand this difference, we provide an equivalent definition of the DSC loss \mathcal{L}_{DSC} (Eq. (4)), in terms of true positive (TP), false negative (FN) and false positive predictions (FP):

$$\mathcal{L}_{\text{DSC}} = 1 - \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}, \quad (5)$$

noting that the DSC score is the harmonic mean of precision and recall, where:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (6)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (7)$$

When both classes are present in equal frequency, the errors associated with the FP and FN predictions are not biased towards a particular class. However, in class imbalanced situations, high precision, low recall solutions are favoured, with over-prediction of the dominant class [23]. Combined with an optimisation goal that favours confident predictions, this results in networks producing extremely confident, and often incorrect, predictions of the dominant class in regions of uncertainty.

To overcome this issue, we reformulate the DSC loss to more heavily penalise overconfident predictions. First, we define another equivalent formulation of the \mathcal{L}_{DSC} , identical in structure to Eq. (5):

$$\mathcal{L}_{\text{DSC}} = 1 - \frac{1}{C} \sum_{c=1}^C \frac{2 \sum_{i=1}^N p_{0i,c} y_{0i,c}}{2 \sum_{i=1}^N p_{0i,c} y_{0i,c} + \sum_{i=1}^N p_{0i,c} y_{1i,c} + \sum_{i=1}^N p_{1i,c} y_{0i,c}}, \quad (8)$$

where $p_{0i,c}$ is the probability of pixel i belonging to class c , and $p_{1i,c}$ is the probability of pixel not belonging to class c . Similarly, y_{0i} is 1 for class c and 0 for all other classes, and conversely y_{1i} takes values of 0 for class c and 1 for all other classes.

To penalise overconfidence for uncertain regions, we apply the focal parameter, γ , directly to the FP and FN predictions, defining the DSC++ loss ($\mathcal{L}_{\text{DSC}++}$):

$$\mathcal{L}_{\text{DSC}++} = 1 - \frac{1}{C} \sum_{c=1}^C \frac{2 \sum_{i=1}^N p_{0i,c} y_{0i,c}}{2 \sum_{i=1}^N p_{0i,c} y_{0i,c} + \sum_{i=1}^N (p_{0i,c} y_{1i,c})^\gamma + \sum_{i=1}^N (p_{1i,c} y_{0i,c})^\gamma}. \quad (9)$$

The DSC++ loss achieves selective penalisation of the overconfident predictions by transforming the penalty from a linear to an exponentially weighted penalty. When $\gamma = 1$, the DSC++ loss is identical to the DSC loss. When $\gamma > 1$, overconfident predictions are more heavily penalised, with increasing values of γ resulting in successively larger penalties applied. Higher γ values therefore favour low confidence predictions. The optimal γ value balances the maintenance of confident, correct predictions while simultaneously suppressing confident but incorrect predictions.

Softmax Thresholding

While the softmax function is not a proxy for uncertainty, the distribution of well calibrated softmax outputs is closely related to the underlying uncertainty, even for out-of-distribution data [24, 25]. To generate a class labelled segmentation output, the argmax function assigns each pixel with the associated class based on the highest softmax value. Rather than using the argmax function, we use a variable threshold that enables manual adjustment of model outputs to favour either precision or recall. Here, we define the output of a model using an indicator function, describing a per-pixel operation that compares the softmax output for the segmentation target, s , to a given softmax threshold \mathcal{T} :

$$I_s = \begin{cases} 1 & \text{if } s < \mathcal{T} \\ 0 & \text{otherwise} \end{cases}. \quad (10)$$

With this generalisation, the argmax function may be restated as a special case where $\mathcal{T} = 0.5$. Higher values of \mathcal{T} favour precision, while lower values favour recall.

Dataset Descriptions and Evaluation Metrics

To evaluate our proposed loss function, we select six public, well-validated biomedical image segmentation datasets. For retinal vessel segmentation, we use the Digital Retinal Images for Vessel Extraction (DRIVE) dataset [26]. The DRIVE dataset consists of 40 coloured fundus photographs obtained from diabetic retinopathy screening in the Netherlands, with an image resolution of 768×584 pixels. The Breast UltraSound 2017 (BUS2017) dataset consists of 163 ultrasound images of breast lesions with an average image size of 760×570 pixels collected from the UDIAT Diagnostic Centre of the Parc Taulí Corporation, Sabadell, Spain [27]. Furthermore, we include the 2018 Data Science Bowl (2018DSB) dataset, which contains 670 light microscopy images for nuclei segmentation [28]. For skin lesion segmentation, we use the ISIC2018: Skin Lesion Analysis Towards Melanoma Detection grand challenge dataset. This dataset contains 2,594 images of skin lesions with an average size of 2166×3188 pixels [29]. For colorectal polyp segmentation, we use the CVC-ClinicDB dataset, which consists of 612 frames containing polyps with image resolution 288×368 pixels, generated from 23 video sequences from 13 different patients using standard colonoscopy interventions with white light [30]. Finally, for 3D multi-class segmentation, we use the Kidney Tumour Segmentation 2019 (KiTS19) dataset [31]. This dataset contains 300 arterial phase abdominal CT scans, with voxel-level kidney and kidney tumour annotations. We exclude the 90 scans without associated segmentation masks, and further exclude another 6 scans (case 15, 23, 37, 68, 125 and 133) due to issues with the ground truth quality [32].

For all the experiments, except for the DRIVE dataset, which is already partitioned into 20 training and 20 test images, we randomly partitioned the other five datasets into 80% development and 20% test set. For all datasets, we further partitioned the development set into 80% training set and 20% validation set. Except for the CVC-ClinicDB and KiTS19 datasets, image resolutions are downsampled using bilinear interpolation. For KiTS19, we performed on-the-fly random sampling of patch size $80 \times 160 \times 160$, with patch-wise overlap of $40 \times 80 \times 80$. A summary of the

datasets, image resolutions and data partitions are presented in Table 1.

To assess the loss functions, we select two evaluation metrics each for calibration and performance. For calibration, we use the NLL and Brier score, both strictly proper scoring rules. The NLL is equivalent to the CE loss in Eq. (1), while the Brier score (Brier) computes the mean squared error between predicted probability scores and the true class labels:

$$\text{Brier} = \frac{1}{C} \frac{1}{N} \sum_{i=1}^C \sum_{j=1}^N (y_i - p_i)^2. \quad (11)$$

For both metrics, a lower score corresponds to better calibration.

For performance, we use the DSC as previously defined, and the Intersection over Union (IoU), also known as the Jaccard Index:

$$\text{Jaccard} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \quad (12)$$

Contrary to the calibration metrics, a higher DSC or Jaccard score corresponds to better performance.

Implementation Details

For our experiments, we leveraged the Medical Image Segmentation with Convolutional Neural Networks (MiS CNN) open-source Python library [33]. This is based on the Keras library using the Tensorflow backend, and all experiments were carried out using NVIDIA P100 GPUs.

Images were resized as described previously and normalised per-image using the z-score. We applied 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).

To investigate the effect of altering γ on the DSC++ loss, we perform a grid search, evaluating values $\gamma \in [0.5, 5]$.

To evaluate the loss functions, we trained the standard U-Net, with model parameters initialised using the Xavier initialisation [34]. We trained each model with instance

Table 1 Summary of the dataset details and training setup used in these experiments

Dataset	Segmentation	#Images	Image resolution	#Training	#Validation	#Test
DRIVE	Retinal vessel	40	512×512	16	4	20
BUS2017	Breast tumour	163	128×128	104	26	33
2018DSB	Cell nucleus	670	256×256	428	108	134
ISIC2018	Skin lesion	2596	512×512	1661	417	518
CVC-ClinicDB	Colorectal polyp	612	288×384	392	98	122
KiTS19	Kidney/Kidney tumour	204	$80 \times 160 \times 160$	130	33	41

For KiTS19, the image resolution refers to the patch size used for training

normalisation, using the stochastic gradient descent optimiser with a batch size of 1 and initial learning rate of 0.1 [35]. For convergence criteria, we used ReduceLROnPlateau to reduce the learning rate by 0.1 if the validation loss did not improve after 25 epochs, and the EarlyStopping callback to terminate training if the validation loss did not improve after 50 epochs. To compromise for the large patch size used for training on the KiTS19 dataset, we used a stricter convergence criteria of 5 epochs and 10 epochs for the ReduceLROnPlateau and EarlyStopping callbacks respectively.

To evaluate the effect of substituting the DSC loss for the DSC++ loss in several DSC-based variants commonly used to achieve state-of-the-art results, we selected the Tversky loss, Focal Tversky loss, Combo loss and Unified Focal loss [10, 23, 36, 37].

The Combo loss ($\mathcal{L}_{\text{Combo}}$) is a compound loss function defined as the weighted sum of the DSC and modified CE loss (\mathcal{L}_{mCE}) [37]:

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

where:

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

The parameters α and β take values in the range [0, 1], controlling the relative contribution of the DSC and CE terms to the loss, and the relative weights assigned to false positives and negatives, respectively. Optimising models with the Combo loss has been observed to improve performance, as well as produce visually more consistent segmentations over models trained using the component losses [38].

To overcome the high precision, low recall bias associated with the DSC loss, the Tversky loss ($\mathcal{L}_{\text{Tversky}}$) modifies the weights associated with the FP and FN predictions [23]:

$$\mathcal{L}_{\text{Tversky}} = \sum_{c=1}^C (1 - \text{TI}), \quad (15)$$

where the Tversky index (TI) is defined as:

$$\text{TI} = \frac{\sum_{i=1}^N p_{0i} y_{0i}}{\sum_{i=1}^N p_{0i} y_{0i} + \alpha \sum_{i=1}^N p_{0i} y_{1i} + \beta \sum_{i=1}^N p_{1i} y_{0i}}, \quad (16)$$

where α and β control the FP and FN weightings, respectively.

To handle class imbalanced data, the Focal Tversky loss (\mathcal{L}_{FT}) applies a focal parameter γ to alter the weights associated with difficult to classify examples [36]:

$$\mathcal{L}_{\text{FT}} = \sum_{c=1}^C (1 - \text{TI})^\gamma, \quad (17)$$

$\gamma < 1$ increases the degree of focusing on harder examples.

Finally, the Unified Focal loss (\mathcal{L}_{UF}) generalises distribution-based and region-based loss functions into a single framework [10], and is defined as the weighted sum of the Asymmetric Focal loss (\mathcal{L}_{AF}) and Asymmetric Focal Tversky loss (\mathcal{L}_{AFT}):

$$\mathcal{L}_{\text{UF}} = \lambda \mathcal{L}_{\text{AF}} + (1 - \lambda) \mathcal{L}_{\text{AFT}}, \quad (18)$$

where:

$$\mathcal{L}_{\text{AF}} = -\frac{\delta}{N} y_{i:r} \log(p_{i,r}) - \frac{1 - \delta}{N} \sum_{c \neq r} (1 - p_{i,c})^\gamma \log(p_{i,r}), \quad (19)$$

$$\mathcal{L}_{\text{AFT}} = \sum_{c \neq r} (1 - \text{TI}) + \sum_{c=r} (1 - \text{TI})^{1-\gamma}, \quad (20)$$

where the TI is redefined as:

$$\text{TI} = \frac{\sum_{i=1}^N p_{0i} y_{0i}}{\sum_{i=1}^N p_{0i} y_{0i} + (1 - \delta) \sum_{i=1}^N p_{0i} y_{1i} + \delta \sum_{i=1}^N p_{1i} y_{0i}}. \quad (21)$$

The three hyperparameters are λ , which controls the relative weights of the two component losses, δ , which controls the relative weighting of positive and negative examples, and γ , which controls the relative weighting of easy and difficult examples.

We used the optimal hyperparameters as described in the original papers, detailed in Table 2. For each loss function, we substituted the DSC component of the loss for the DSC++ loss, setting $\gamma = 2$.

To test for statistical significance, we used the Wilcoxon rank sum test. A statistically significant difference was defined as $p < 0.05$. We use bootstrapping to calculate the standard errors for each metric. To evaluate effect of softmax thresholding, we selected thresholds $\mathcal{T} \in [0.05, 0.95]$ using the DSC and DSC++ loss on the DRIVE dataset.

Table 2 Hyperparameter settings used in these experiments for the DSC and cross entropy-based loss functions

Loss	Hyperparameter				
	α	β	γ	δ	λ
Focal	0.5	-	2	-	-
Tversky	0.3	0.7	-	-	-
Focal Tversky	0.3	0.7	$\frac{4}{3}$	-	-
Combo	0.5	0.5	-	-	-
Unified Focal	-	-	0.5	0.6	0.5

Results

In this section, we first describe the results for the hyperparameter experiments using the DSC++ loss, before comparing the DSC loss, CE loss, Focal loss and DSC++ loss on five 2D binary segmentation tasks, followed by on one 3D multi-class segmentation task. Next, we compare the performance and calibration of various DSC-based loss functions with and without the DSC++ modification. Finally, we compare the effects of softmax thresholding using the DSC and DSC++ loss on recall-precision bias.

DSC++ Loss Hyperparameter Tuning

The results for the hyperparameter experiments using the DSC++ loss are shown in Table 3.

Most noticeable is the significant decrease in the NLL with values of $\gamma > 1$. The NLL decreases with increasing γ , appearing to plateau at $\gamma = 2$. Similarly, Brier score decreases with increasing γ , with the lowest Brier scores at γ values of 2 and 2.5. With $\gamma = 2$, there is a statistically significance difference in NLL ($p = 6 \times 10^{-8}$) and Brier ($p = 2 \times 10^{-7}$) scores compared to the DSC loss. However, above this range, increasing gamma leads to an increase in NLL and Brier score. In terms of performance

Table 3 Calibration and performance of the DSC++ loss on the DRIVE dataset with different γ values

Gamma	Uncertainty		Performance	
	NLL (\downarrow)	Brier (\downarrow)	Dice (\uparrow)	Jaccard (\uparrow)
0.5	0.281 (± 0.019)	0.033 (± 0.001)	0.804 (± 0.003)	0.673 (± 0.004)
1.0	0.204 (± 0.014)	0.031 (± 0.001)	0.804 (± 0.003)	0.672 (± 0.05)
1.5	0.067 (± 0.005)	0.026 (± 0.001)	0.804 (± 0.004)	0.673 (± 0.005)
2.0	0.041 (± 0.003)	0.024 (± 0.001)	0.808 (± 0.003)	0.678 (± 0.004)
2.5	0.038 (± 0.002)	0.024 (± 0.001)	0.804 (± 0.003)	0.673 (± 0.05)
3.0	0.038 (± 0.002)	0.027 (± 0.001)	0.797 (± 0.004)	0.664 (± 0.006)
3.5	0.035 (± 0.002)	0.031 (± 0.001)	0.804 (± 0.004)	0.672 (± 0.005)
4.0	0.038 (± 0.002)	0.034 (± 0.001)	0.796 (± 0.004)	0.661 (± 0.006)
4.5	0.039 (± 0.002)	0.042 (± 0.001)	0.795 (± 0.004)	0.660 (± 0.005)
5.0	0.041 (± 0.002)	0.048 (± 0.001)	0.794 (± 0.004)	0.658 (± 0.006)

The standard errors are shown in brackets. The best scores are denoted in bold

metrics, the highest DSC and Jaccard scores were observed with $\gamma = 2$, at 0.808 and 0.678 respectively. There is no statistically significant difference between the DSC and Jaccard scores using different γ values, suggesting that the improved calibration score does not come at the cost to performance.

To understand whether γ improves calibration scores by reducing model overconfidence, Fig. 2 shows an example of the softmax probability outputs for an example test set image.

With increasing γ values, there is a reduction in overconfident model predictions, in comparison to the DSC loss ($\gamma = 1$), where model predictions are concentrated at the extremes. Importantly, the low confidence areas are concentrated around the difficult to segment smaller retinal vessels, providing a plausible approximation of the underlying uncertainty.

Loss Function Comparisons

2D Binary Segmentation

The CE loss, Focal loss, DSC loss and DSC++ loss were evaluated on five, 2D binary biomedical imaging datasets. Based on the results from the hyperparameter investigation, we set $\gamma = 2$ for the DSC++ loss. The results are shown in Table 4.

Firstly, there was a statistically significant difference between the NLL using DSC++ loss compared to the DSC loss, across all datasets (DRIVE: $p = 6 \times 10^{-8}$, BUS2017: $p = 0.01$, 2018DSB: $p = 8 \times 10^{-13}$, ISIC2018: $p = 1 \times 10^{-12}$ and CVC-ClinicDB: $p = 2 \times 10^{-11}$). There was no significant difference between the NLL values using the CE, Focal or DSC++ loss. The DSC++ loss achieved the lowest Brier score for all five datasets, with statistically significant differences observed on the DRIVE ($p = 2 \times 10^{-7}$), ISIC2018 ($p = 0.01$) and CVC-ClinicDB ($p = 0.04$) datasets compared to the DSC loss. The DSC++ loss achieved the highest DSC score for four out of the five datasets, and the highest Jaccard score for three out of the five datasets. In contrast, the CE-based loss achieved the lowest performance scores across all datasets, with the Focal loss achieving the lowest Dice and Jaccard score for four out of the five datasets. A statistically significant difference ($p < 0.05$) was observed on the DRIVE dataset for both the DSC and Jaccard scores between the DSC++ and CE-based losses.

Example segmentations using each loss function for the five datasets are shown in Fig. 3. Visually, the best segmentations are observed using the DSC++ loss. While model predictions derived from CE-based losses appear well calibrated, the segmentation quality is generally poor. In contrast, the DSC loss, despite very confident predictions,

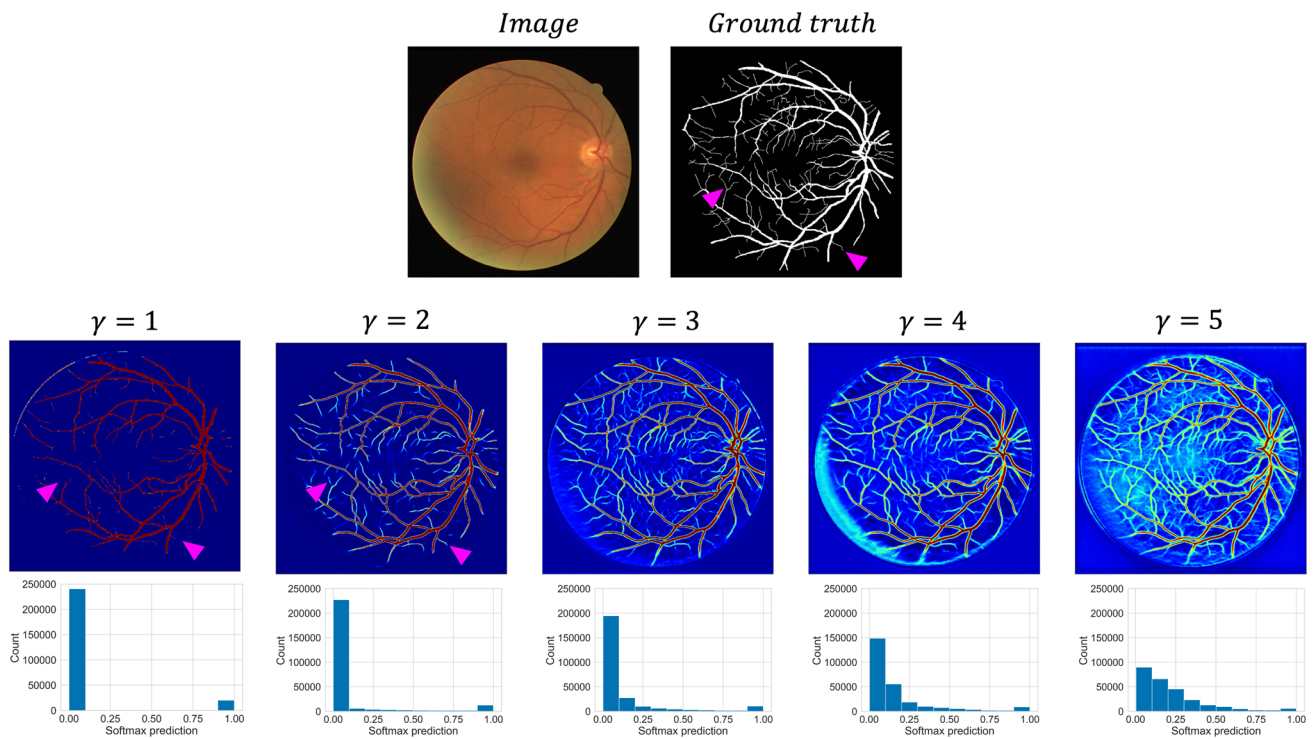


Fig. 2 The effect of altering γ on the softmax prediction outputs. Top: input image and ground truth segmentation. The pink arrows highlight example areas where segmentation quality differs. Middle: the softmax predictions for each model trained using the DSC++ loss

with different γ value are displayed as heatmaps. Bottom: Histogram plots showing the softmax predictions and corresponding number of pixels

Table 4 Calibration and performance of different loss functions on five biomedical imaging datasets

Dataset	Loss	Calibration		Performance	
		NLL (\downarrow)	Brier (\downarrow)	Dice (\uparrow)	Jaccard (\uparrow)
Drive	CE	0.051 (± 0.003)	0.024 (± 0.001)	0.798 (± 0.004)	0.664 (± 0.005)
	Focal	0.048 (± 0.002)	0.036 (± 0.001)	0.795 (± 0.005)	0.660 (± 0.007)
	DSC	0.204 (± 0.013)	0.031 (± 0.001)	0.804 (± 0.003)	0.672 (± 0.005)
	DSC++	0.041 (± 0.003)	0.024 (± 0.001)	0.808 (± 0.003)	0.678 (± 0.004)
BUS2017	CE	0.020 (± 0.005)	0.014 (± 0.003)	0.787 (± 0.037)	0.690 (± 0.041)
	Focal	0.019 (± 0.004)	0.020 (± 0.003)	0.770 (± 0.041)	0.673 (± 0.042)
	DSC	0.137 (± 0.046)	0.022 (± 0.005)	0.784 (± 0.038)	0.688 (± 0.042)
	DSC++	0.034 (± 0.016)	0.013 (± 0.004)	0.842 (± 0.031)	0.756 (± 0.034)
2018DSB	CE	0.033 (± 0.003)	0.019 (± 0.002)	0.912 (± 0.006)	0.845 (± 0.009)
	Focal	0.044 (± 0.004)	0.028 (± 0.002)	0.904 (± 0.007)	0.832 (± 0.010)
	DSC	0.167 (± 0.019)	0.025 (± 0.002)	0.916 (± 0.006)	0.852 (± 0.009)
	DSC++	0.033 (± 0.004)	0.019 (± 0.002)	0.916 (± 0.006)	0.850 (± 0.009)
ISIC2018	CE	0.083 (± 0.010)	0.036 (± 0.003)	0.863 (± 0.008)	0.787 (± 0.009)
	Focal	0.068 (± 0.005)	0.041 (± 0.002)	0.865 (± 0.008)	0.793 (± 0.009)
	DSC	0.373 (± 0.037)	0.044 (± 0.003)	0.883 (± 0.006)	0.812 (± 0.008)
	DSC++	0.086 (± 0.011)	0.034 (± 0.003)	0.882 (± 0.006)	0.811 (± 0.008)
CVC-ClinicDB	CE	0.041 (± 0.008)	0.015 (± 0.002)	0.870 (± 0.014)	0.796 (± 0.017)
	Focal	0.028 (± 0.004)	0.014 (± 0.002)	0.893 (± 0.013)	0.828 (± 0.015)
	DSC	0.167 (± 0.033)	0.019 (± 0.003)	0.884 (± 0.014)	0.817 (± 0.016)
	DSC++	0.037 (± 0.007)	0.013 (± 0.002)	0.894 (± 0.013)	0.829 (± 0.015)

The standard errors are shown in brackets. The best scores are denoted in bold

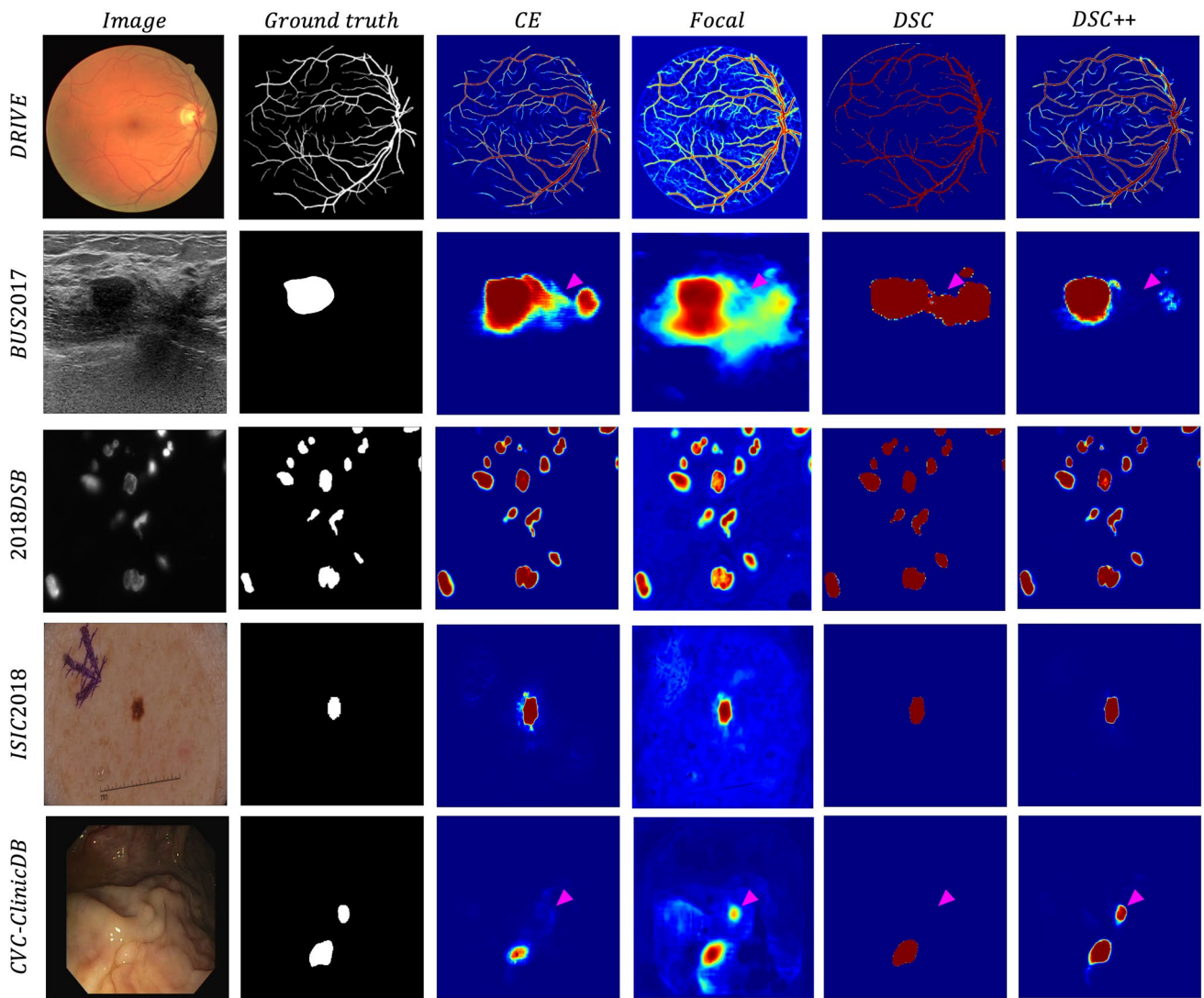


Fig. 3 Example segmentations, with softmax predictions visualised as a heatmap, for each loss function for each of the five datasets. The image and ground truth are provided for reference. The pink arrows highlight example areas where segmentation quality differs

Table 5 Calibration and performance of different loss functions on the KiTS19 dataset

Loss	Kidney				Kidney tumour			
	Calibration		Performance		Calibration		Performance	
	NLL (↓)	Brier (↓)	Dice (↑)	Jaccard (↑)	NLL (↓)	Brier (↓)	Dice (↑)	Jaccard (↑)
CE	0.012 (±0.005)	0.007 (±0.003)	0.896 (±0.012)	0.819 (±0.017)	0.031 (±0.003)	0.012 (±0.002)	0.188 (±0.035)	0.124 (±0.025)
Focal	0.012 (±0.004)	0.008 (±0.003)	0.911 (±0.009)	0.841 (±0.014)	0.020 (±0.002)	0.011 (±0.002)	0.301 (±0.043)	0.213 (±0.035)
DSC	0.050 (±0.022)	0.008 (±0.003)	0.818 (±0.019)	0.710 (±0.026)	0.124 (±0.021)	0.014 (±0.003)	0.232 (±0.035)	0.153 (±0.027)
DSC++	0.017 (±0.007)	0.007 (±0.003)	0.911 (±0.008)	0.841 (±0.013)	0.045 (±0.006)	0.012 (±0.002)	0.429 (±0.041)	0.311 (±0.036)

The standard errors are shown in brackets. The best scores are denoted in bold

produces considerable false positive predictions such as in the BUS2017 example, as well as false negative predictions as seen in the CVC-ClinicDB example.

3D Multi-class Segmentation

The performance of the CE loss, Focal loss, DSC loss and DSC++ loss was further evaluated on the KiTS19 dataset, a 3D multi-class segmentation task. The results are shown in Table 5.

The DSC++ achieved significantly better calibration scores compared to the DSC loss across both classes (Kidney: $p = 3 \times 10^{-8}$, Kidney tumour: $p = 2 \times 10^{-8}$). In contrast, there was no significant difference in calibration scores between the DSC++ loss and CE-based losses. In terms of segmentation quality, the DSC++ achieved the best performance with a DSC score of 0.911 and 0.429 for the kidney and kidney tumour segmentation respectively. The DSC score on the kidney tumour class using the DSC++ loss significantly outperformed the other loss functions (DSC: $p = 2 \times 10^{-6}$, CE: $p = 4 \times 10^{-7}$, Focal: $p = 0.0002$).

Example segmentations using each loss function on the KiTS19 dataset is shown in Fig. 4. The DSC++ loss

produces accurate and well calibrated segmentations, for both kidney and kidney tumour class. The DSC loss produces false positive predictions with high confidence, most noticeable with the kidney tumour class. The CE-based losses produce poor quality kidney tumour segmentation, with associated over-segmentation of the kidney.

Incorporating the DSC++ Loss into Other Dice-Based Loss Functions

The DSC loss forms the basis for several other region-based loss functions, and therefore we investigate the effect of integrating the DSC++ loss modification into these loss functions. The results are shown in Table 6.

The DSC-based variants appear to all inherit the poorly calibrated nature of the DSC loss, except for the two compound loss functions, the Combo loss and the Unified Focal loss, which also incorporate the CE-based variants. Using the DSC++ loss led to significant improvements in calibration for all loss functions compared, for both the NLL and Brier scores. Similarly, the highest performance, measured using the DSC and Jaccard scores, was obtained using the DSC++ variants.

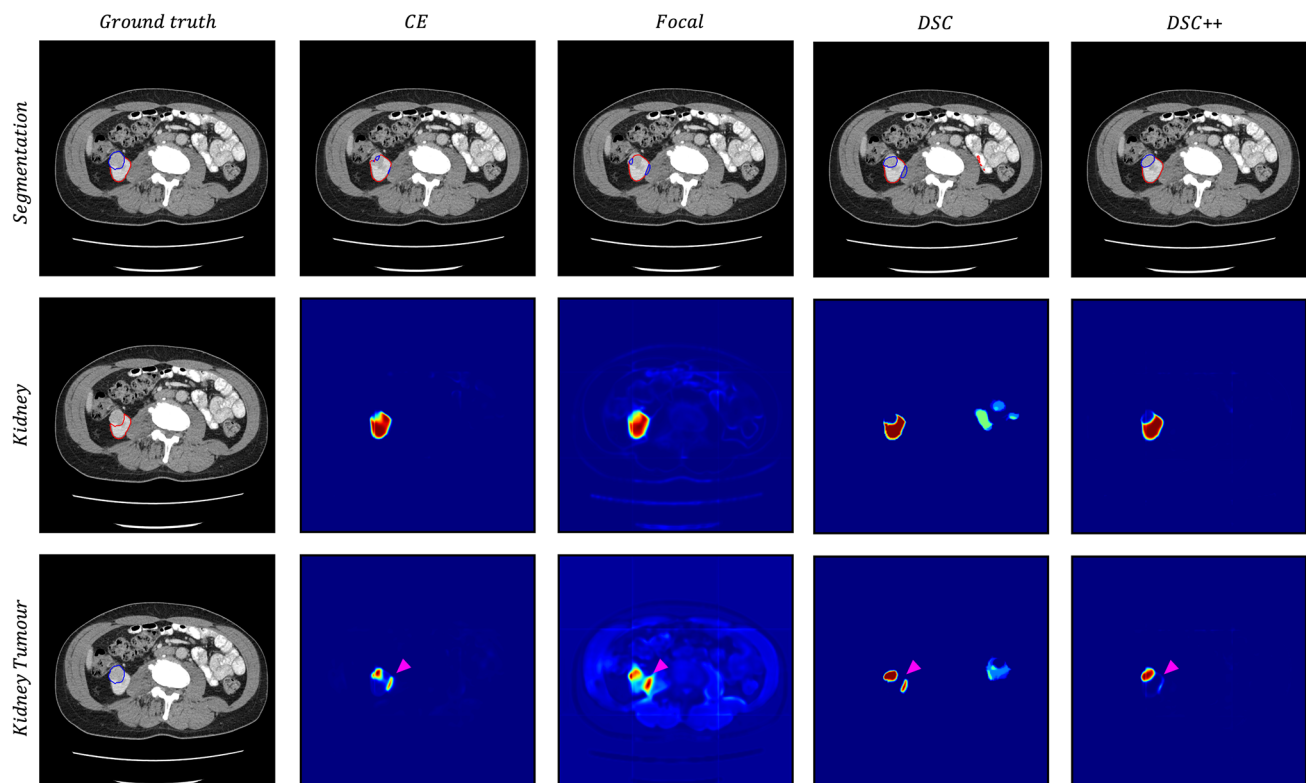


Fig. 4 Example segmentation, with softmax predictions visualised as a heatmap, for each loss function. The image and ground truth are provided for reference. The pink arrows highlight example areas where segmentation quality differs

Table 6 Calibration and performance of the DSC-based loss functions, using either the original loss functions (Tversky, Focal Tversky, Combo and Unified Focal) or substituting the DSC component of the loss for the DSC++ loss (Tversky++, Focal Tversky++, Combo++ and Unified Focal++)

Loss	Calibration		Performance	
	NLL (\downarrow)	Brier (\downarrow)	Dice (\uparrow)	Jaccard (\uparrow)
Tversky	0.144 (± 0.011)	0.034 (± 0.001)	0.807 (± 0.003)	0.676 (± 0.004)
Tversky++	0.033 (± 0.003)	0.025 (± 0.001)	0.810 (± 0.003)	0.681 (± 0.004)
Focal Tversky	0.142 (± 0.011)	0.033 (± 0.001)	0.807 (± 0.003)	0.677 (± 0.004)
Focal Tversky++	0.036 (± 0.003)	0.024 (± 0.001)	0.810 (± 0.003)	0.680 (± 0.004)
Combo	0.063 (± 0.004)	0.025 (± 0.001)	0.802 (± 0.004)	0.669 (± 0.005)
Combo++	0.050 (± 0.003)	0.024 (± 0.001)	0.802 (± 0.003)	0.670 (± 0.005)
Unified Focal	0.056 (± 0.004)	0.026 (± 0.001)	0.810 (± 0.003)	0.680 (± 0.004)
Unified Focal++	0.039 (± 0.003)	0.024 (± 0.001)	0.810 (± 0.003)	0.681 (± 0.004)

γ is set to 2 for the DSC++ variants. The standard errors are shown in brackets. The best scores are denoted in bold

Softmax Thresholding

The effect of softmax thresholding on the performance of the DSC and DSC++ loss for the DRIVE dataset are shown in Fig. 5. The DSC loss predictions display almost no variation across the entire range of softmax thresholds. In contrast, there are significant variations in recall and precision scores

using the DSC++ loss. Importantly, considerable increases in recall or precision between $T = 0.3$ and $T = 0.7$ did not affect the DSC score. The DSC++ loss enables models to be tailored to provide either very high recall or precision values, with little effect on the DSC score. For example, the model achieved a precision of 0.923 and DSC of 0.748 at $T = 0.8$, and recall of 0.923 and DSC of 0.761 at $T = 0.2$.

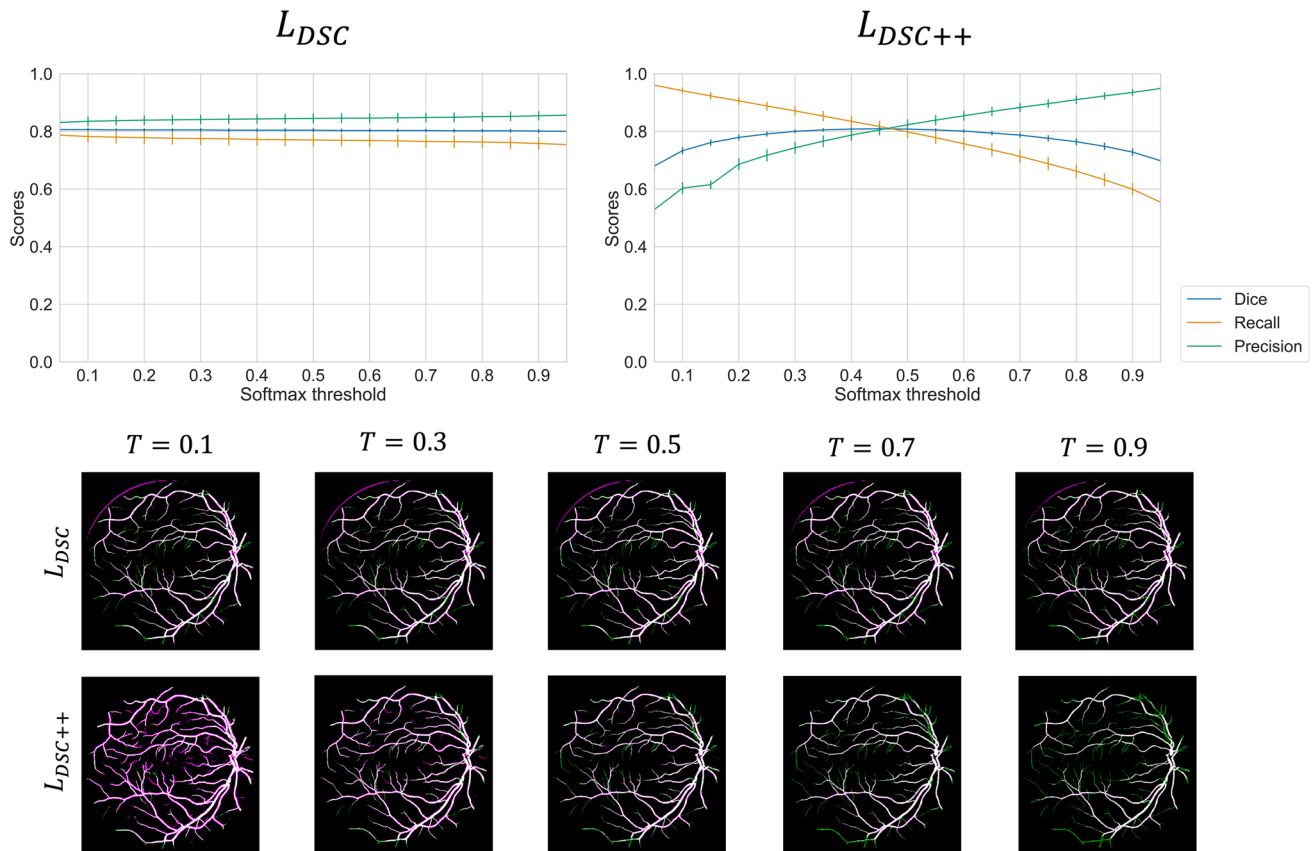


Fig. 5 The effect of softmax thresholding on the recall and precision using models trained with the DSC and DSC++ loss on the DRIVE dataset. Top: Recall, precision and DSC scores at different softmax thresholds for the DSC and DSC++ loss. The vertical bars represent

the 95% confidence intervals. Bottom: Example segmentation output at different softmax thresholds. The false positives are highlighted in magenta, and the false negatives are highlighted in green

Discussion

In this work, we identified the weights associated with the FP and FN predictions as the reason for the poor calibration associated with the DSC loss, and used this to provide a reformulation, known as the DSC++ loss, which uses a γ parameter to more heavily penalise overconfident predictions. We observed significantly improved calibration using the DSC++ loss over the DSC loss, measured using the NLL and Brier scores, across six well-validated open-source datasets, including both 2D binary and 3D multi-class segmentation tasks. Furthermore, we demonstrated that the variants of the DSC loss inherit poor calibration, while those using DSC++ variants led to significant improvements in calibration. Finally, we evaluated the effect of softmax thresholding on the DSC loss and DSC++ loss, where little variation in recall or precision was observed with the DSC loss, in comparison to the significant variation achievable using the DSC++ loss.

Modifying the loss function, rather than the network or training setup, is the most intuitive solution to improve calibration. This is because with optimal training, it is the loss function that primarily determines the calibration quality of the resulting segmentation outputs. Optimisation with the DSC loss will encourage overconfident predictions (Fig. 1), and therefore methods—such as MC dropout or deep ensembling—may improve calibration, but do not address the direct cause of the issue. Importantly, both MC dropout and deep ensembling significantly increase inference time, with the latter requiring additional computational resources to handle predictions from multiple networks. Furthermore, MC dropout requires modifying networks to include dropout layers, and this may not be compatible with certain architectures.

We also explored the synergistic effect of softmax thresholding, together with well calibrated outputs, to enable tailoring towards high recall or high precision output states (Fig. 5). For biomedical or clinical use, generally high recall is favoured, especially when the role of automatic segmentation systems is to support human operators in reducing false negative predictions, for example with polyp identification during colonoscopy [39]. As shown in Fig. 5, it is possible to identify even the small-diameter retinal vessels when recall is prioritised. It is possible to optimise models to produce high recall or precision outputs, such as the Tversky loss modification of the DSC loss [23]. However, after model training, it is not possible to further modify the recall-precision bias, which would instead require the training of a new model. Softmax thresholding is used during post-processing and is therefore independent of the model, enabling flexible and reversible control over the recall-precision bias. Even without softmax thresholding, the uncertainty associated with well calibrated predictions can highlight regions of interest which

may be missed when interpreting poorly calibrated predictions (Figs. 3 and 4).

Given the widespread use of these functions, it is important to consider whether there are any reasons to not replace them with these alternatives. The one apparent limitation of using the DSC++ loss over the DSC loss is additional hyperparameter tuning required. However, we investigated a large range of γ values (Table 3 and Fig. 2), and observed that performance was not significantly affected, while the calibration improves significantly, even with small values of γ . Moreover, we selected a γ value of 2 based on results from the DRIVE dataset, and this appeared to generalise well across the other five datasets, with consistently significant improvements to calibration (Tables 4 and 5). Therefore, even small γ parameter values appear to be effective, and optimal choices for γ generalise well across datasets, suggesting that the γ parameter is relatively easy to optimise.

It is less clear whether the DSC++ loss should be favoured above other loss functions. Besides calibration, the DSC++ loss suffers from the same limitations as the DSC loss, namely the unstable gradient, resulting from gradient calculations involving small denominators [14, 40]. While there is currently little empirical evidence relating the unstable gradient to suboptimal performance, it has been suggested that incorporating the CE loss helps to mitigate the unstable gradients generated by the DSC loss [41]. Our experiments confirm previous results that compound loss functions generally perform better [9, 10]. However, even if the DSC++ cannot replace these loss functions, we have shown that replacing the DSC component of loss functions with the DSC++ loss leads to significant improvements in calibration, as well as evidence of better performance (Table 6).

In future work, we will investigate the effect of gradient instability on the performance of the DSC++ loss. It would be important to evaluate the performance on highly class imbalanced datasets, where gradient stabilisation may be expected to be more important. Furthermore, it would be useful to evaluate networks trained using the DSC++ loss on out-of-distribution data, to test whether the model predictions remain well calibrated.

Conclusion

In this study, we identified the main reason behind neural network overconfidence when training deep learning-based image segmentation models using the DSC loss, and provided a simple yet effective modification, named the DSC++ loss, that directly addresses the issue. After evaluating the performance and calibration of both the DSC loss and DSC++ loss across six well-validated biomedical imaging

datasets, as well as systematically analysing the softmax predictions, it is clear that the DSC loss is not suitable for training neural networks for use in biomedical or clinical practice. In contrast, the DSC++ loss, together with its synergistic effect using softmax thresholding, produce model outputs that are useful to interpret, and readily adjustable to provide high recall or precision outputs. Compared with previous methods used to improve the calibration of networks trained using the DSC loss, the DSC++ loss provides the most intuitive, readily accessible solution that is an important contribution towards the goal of deploying deep learning image segmentation systems into biomedical or clinical practice.

Data Availability The data used in this study are all openly available [26–31].

Declarations

Competing Interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Pal, N.R., Pal, S.K.: A review on image segmentation techniques. *Pattern Recognit.* **26**(9), 1277–1294 (1993). [https://doi.org/10.1016/0031-3203\(93\)90135-J](https://doi.org/10.1016/0031-3203(93)90135-J)
2. Roth, H.R., Lu, L., Farag, A., Shin, H.-C., Liu, J., Turkbey, E.B., Summers, R.M.: Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In: *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 556–564 (2015). https://doi.org/10.1007/978-3-319-24553-9_68. Springer
3. Reinke, A., Eisenmann, M., Tizabi, M.D., Sudre, C.H., Rädtsch, T., Antonelli, M., Arbel, T., Bakas, S., Cardoso, M.J., Cheplygina, V., et al.: Common limitations of image processing metrics: A picture story. *arXiv preprint arXiv:2104.05642* (2021)
4. Fidon, L., Li, W., Garcia-Peraza-Herrera, L.C., Ekanayake, J., Kitchen, N., Ourselin, S., Vercauteren, T.: Generalised wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks. In: *International MICCAI Brain Lesion Workshop*, pp. 64–76 (2017). Springer
5. Sander, J., de Vos, B.D., Wolterink, J.M., Išgum, I.: Towards increased trustworthiness of deep learning segmentation methods on cardiac mri. In: *Medical Imaging 2019: Image Processing*, vol. 10949, p. 1094919 (2019). International Society for Optics and Photonics
6. Mehrtaş, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Trans. Med. Imaging* **39**(12), 3868–3878 (2020)
7. Rousseau, A.-J., Becker, T., Bertels, J., Blaschko, M.B., Valkenburg, D.: Post training uncertainty calibration of deep networks for medical image segmentation. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1052–1056 (2021). IEEE
8. Ghafoorian, M., Mehrtaş, A., Kapur, T., Karssemeijer, N., Marchiori, E., Pesteie, M., Guttmann, C.R., de Leeuw, F.-E., Tempany, C.M., Van Ginneken, B., et al.: Transfer learning for domain adaptation in mri: Application in brain lesion segmentation. In: *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 516–524 (2017). Springer
9. Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X., Martel, A.L.: Loss odyssey in medical image segmentation. *Med. Image Anal.*, 102035 (2021)
10. Yeung, M., Sala, E., Schönlieb, C.-B., Rundo, L.: Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics*, 102026 (2021)
11. Milletari, F., Navab, N., Ahmadi, S.-A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: *Proc. Fourth International Conference on 3D Vision (3DV)*, pp. 565–571 (2016). <https://doi.org/10.1109/3DV.2016.79>. IEEE
12. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: 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, Cham, Switzerland (2017)
13. Eelbode, T., Bertels, J., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., Blaschko, M.B.: Optimization for medical image segmentation: theory and practice when evaluating with dice score or jaccard index. *IEEE Trans. Med. Imaging* **39**(11), 3679–3690 (2020)
14. Bertels, J., Robben, D., Vandermeulen, D., Suetens, P.: Optimization with soft dice can lead to a volumetric bias. In: *International MICCAI Brainlesion Workshop*, pp. 89–97 (2019). Springer
15. Bertels, J., Robben, D., Vandermeulen, D., Suetens, P.: Theoretical analysis and experimental validation of volume bias of soft dice optimized segmentation maps in the context of inherent uncertainty. *Med. Image Anal.* **67**, 101833 (2021)
16. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: *Proc. International Conference on Computer Vision (ICCV)*, pp. 2999–3007 (2017). IEEE
17. Dong, Y., Shen, X., Jiang, Z., Wang, H.: Recognition of imbalanced underwater acoustic datasets with exponentially weighted cross-entropy loss. *Appl. Acoust.* **174**, 107740 (2021)
18. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: *International Conference on Machine Learning*, pp. 1050–1059 (2016). PMLR
19. Platt, J., et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* **10**(3), 61–74 (1999)
20. DeVries, T., Taylor, G.W.: Leveraging uncertainty estimates for predicting segmentation quality. *arXiv preprint arXiv:1807.00502* (2018)
21. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474* (2016)
22. Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc* **102**(477), 359–378 (2007)

23. Salehi, S.S.M., Erdogmus, D., Gholipour, A.: 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 (2017). https://doi.org/10.1007/978-3-319-67389-9_44. Springer
24. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning, pp. 1321–1330 (2017). PMLR
25. Pearce, T., Brintrup, A., Zhu, J.: Understanding softmax confidence and uncertainty. arXiv preprint [arXiv:2106.04972](https://arxiv.org/abs/2106.04972) (2021)
26. Staal, J., Abramoff, M.D., Niemeijer, M., Viergever, M.A., Van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. *IEEE Trans. Med. Imaging* **23**(4), 501–509 (2004)
27. Yap, M.H., Pons, G., Martí, J., Ganau, S., Sentís, M., Zwiggelaar, R., Davison, A.K., Martí, R.: Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE J Biomed Health Inform* **22**(4), 1218–1226 (2017)
28. Caicedo, J.C., Goodman, A., Karhohs, K.W., Cimini, B.A., Ackerman, J., Haghighi, M., Heng, C., Becker, T., Doan, M., McQuin, C., *et al*: Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nat. Methods* **16**(12), 1247–1253 (2019)
29. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., *et al*: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint [arXiv:1902.03368](https://arxiv.org/abs/1902.03368) (2019)
30. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput Med Imaging Graph* **43**, 99–111 (2015)
31. Heller, N., Sathianathan, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., *et al*: The KiTS19 challenge data: 300 kidney tumor cases with clinical context. arXiv preprint [arXiv:1904.00445](https://arxiv.org/abs/1904.00445) (2019)
32. Heller, N., Isensee, F., Maier-Hein, K.H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., *et al*: 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 (2021). <https://doi.org/10.1016/j.media.2020.101821>
33. Müller, D., Kramer, F.: Miscnn: a framework for medical image segmentation with convolutional neural networks and deep learning. *BMC Med. Imaging* **21**(1), 1–11 (2021)
34. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241 (2015). Springer
35. Zhou, X.-Y., Yang, G.-Z.: Normalization in training U-Net for 2-D biomedical semantic segmentation. *IEEE Robot. Autom. Lett.* **4**(2), 1792–1799 (2019)
36. Abraham, N., Khan, N.M.: A novel focal tversky loss function with improved attention u-net for lesion segmentation. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 683–687 (2019). IEEE
37. Taghanaki, S.A., Zheng, Y., Zhou, S.K., Georgescu, B., Sharma, P., Xu, D., Comaniciu, D., Hamarneh, G.: Combo loss: Handling input and output imbalance in multi-organ segmentation. *Comput. Med. Imaging Graph.* **75**, 24–33 (2019). <https://doi.org/10.1016/j.compmedimag.2019.04.005>
38. Drozdal, M., Vorontsov, E., Chartrand, G., Kadoury, S., Pal, C.: The importance of skip connections in biomedical image segmentation. In: Deep Learning and Data Labeling for Medical Applications, pp. 179–187. Springer, Cham, Switzerland (2016)
39. Nogueira-Rodríguez, A., Domínguez-Carbajales, R., López-Fernández, H., Iglesias, Á., Cubiella, J., Fdez-Riverola, F., Reboiro-Jato, M., Glez-Peña, D.: Deep neural networks approaches for detecting and classifying colorectal polyps. *Neurocomputing* **423**, 721–734 (2021)
40. Wong, K.C., Moradi, M., Tang, H., Syeda-Mahmood, T.: 3d segmentation with exponential logarithmic loss for highly unbalanced object sizes. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 612–619 (2018). Springer
41. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**(2), 203–211 (2021)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.