The Dice Loss in the Context of Missing or Empty Labels: Introducing Φ and ϵ

Sofie Tilborghs[1,2(✉)], Jeroen Bertels[1,2], David Robben[1,2,3], Dirk Vandermeulen[1,2], and Frederik Maes[1,2]

[1] Department of Electrical Engineering, ESAT/PSI, KU Leuven, Leuven, Belgium
`{sofie.tilborghs,jeroen.bertels}@kuleuven.be`
[2] Medical Imaging Research Center, UZ Leuven, Leuven, Belgium
[3] icometrix, Kolonel Begaultlaan 1b/12, Leuven, Belgium

Abstract. Albeit the Dice loss is one of the dominant loss functions in medical image segmentation, most research omits a closer look at its derivative, i.e. the real motor of the optimization when using gradient descent. In this paper, we highlight the peculiar action of the Dice loss in the presence of missing or empty labels. First, we formulate a theoretical basis that gives a general description of the Dice loss and its derivative. It turns out that the choice of the reduction dimensions Φ and the smoothing term ϵ is non-trivial and greatly influences its behavior. We find and propose heuristic combinations of Φ and ϵ that work in a segmentation setting with either missing or empty labels. Second, we empirically validate these findings in a binary and multiclass segmentation setting using two publicly available datasets. We confirm that the choice of Φ and ϵ is indeed pivotal. With Φ chosen such that the reductions happen over a single batch (and class) element and with a negligible ϵ, the Dice loss deals with missing labels naturally and performs similarly compared to recent adaptations specific for missing labels. With Φ chosen such that the reductions happen over multiple batch elements or with a heuristic value for ϵ, the Dice loss handles empty labels correctly. We believe that this work highlights some essential perspectives and hope that it encourages researchers to better describe their exact implementation of the Dice loss in future work.

1 Introduction

The *Dice loss* was introduced in [5] and [13] as a loss function for binary image segmentation taking care of the class imbalance between foreground and background often present in medical applications. The *generalized Dice loss* [16] extended this idea to multiclass segmentation tasks, thereby taking into account the class imbalance that is present across different classes. In parallel, the Jaccard loss was introduced in the wider computer vision field for the same purpose [14,17]. More recently, it has been shown that one can use either Dice or Jaccard loss during training to effectively optimize both metrics at test time [6].

S. Tilborghs and J. Bertels—Contributed equally to this work.

L. Wang et al. (Eds.): MICCAI 2022, LNCS 13435, pp. 527–537, 2022.
https://doi.org/10.1007/978-3-031-16443-9_51

The use of the Dice loss in popular and state-of-the-art methods such as No New-Net [9] has only fueled its dominant usage across the entire field of medical image segmentation. Despite its fast and wide adoption, research that explores the underlying mechanisms is remarkably limited and mostly focuses on the loss value itself building further on the concept of *risk minimization* [8]. Regarding model calibration and inherent uncertainty, for example, some intuitions behind the typical hard and poorly calibrated predictions were exposed in [4], thereby focusing on the potential volume bias as a result of using the Dice loss. Regarding semi-supervised learning, adaptations to the original formulations were proposed to deal with "missing" labels [7,15], i.e. a label that is missing in the ground truth even though it is present in the image.

In this work, we further contribute to a deeper understanding of the specific implementation of the Dice loss, especially in the context of missing and empty labels. In contrast to missing labels, "empty" labels are labels that are not present in the image (and hence also not in the ground truth). We will first take a closer look at the derivative, i.e. the real motor of the underlying optimization when using gradient descent, in Sect. 2. Although [13] and [16] report the derivative, it is not being discussed in detail, nor is any reasoning behind the choice of the reduction dimensions Φ given (Sect. 2.1). When the smoothing term ϵ is mentioned, no details are given and its effect is underestimated by merely linking it with numerical stability [16] and convergence issues [9]. In fact, we find that both Φ and ϵ are intertwined, and that their choice is non-trivial and pivotal in the presence of missing or empty labels. To confirm and validate these findings, we set up two empirical settings with missing or empty labels in Sects. 3 and 4. Indeed, we can make or break the segmentation task depending on the exact implementation of the Dice loss.

2 Bells and Whistles of the Dice Loss: Φ and ϵ

In a CNN-based setting, the weights $\theta \in \Theta$ are often updated using gradient descent. For this purpose, the loss function ℓ computes a real valued cost $\ell(Y, \tilde{Y})$ based on the comparison between the ground truth Y and its prediction $\tilde{Y}$ in each iteration. Y and $\tilde{Y}$ contain the values $y_{b,c,i}$ and $\tilde{y}_{b,c,i}$, respectively, pointing to the value for a semantic class $c \in \mathcal{C} = [\mathrm{C}]$ at an index $i \in \mathcal{I} = [\mathrm{I}]$ (e.g. a voxel) of a batch element $b \in \mathcal{B} = [\mathrm{B}]$ (Fig. 1). The exact update of each θ depends on $d\ell(Y, \tilde{Y})/d\theta$, which can be computed via the generalized chain rule. With $\omega = (b, c, i) \in \Omega = \mathcal{B} \times \mathcal{C} \times \mathcal{I}$, we can write:

$$\frac{d\ell(Y, \tilde{Y})}{d\theta} = \sum_{b \in \mathcal{B}} \sum_{c \in \mathcal{C}} \sum_{i \in \mathcal{I}} \frac{\partial \ell(Y, \tilde{Y})}{\partial \tilde{y}_{b,c,i}} \frac{\partial \tilde{y}_{b,c,i}}{\partial \theta} = \sum_{\omega \in \Omega} \frac{\partial \ell(Y, \tilde{Y})}{\partial \tilde{y}_{\omega}} \frac{\partial \tilde{y}_{\omega}}{\partial \theta}. \tag{1}$$

The Dice similarity coefficient (DSC) over a subset $\phi \subset \Omega$ is defined as:

$$\mathrm{DSC}(Y_{\phi}, \tilde{Y}_{\phi}) = \frac{2|Y_{\phi} \cap \tilde{Y}_{\phi}|}{|Y_{\phi}| + |\tilde{Y}_{\phi}|}. \tag{2}$$

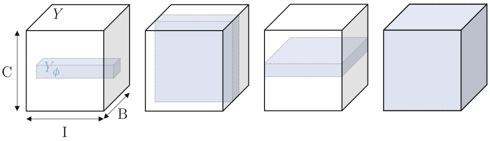

Fig. 1. Schematic representation of Y, having a batch, class and image dimension, respectively with $|\mathcal{B}| = \mathrm{B}$, $|\mathcal{C}| = \mathrm{C}$ and $|\mathcal{I}| = \mathrm{I}$ (similarly for $\tilde{Y}$). The choice of Φ, i.e. a family of subsets ϕ over Ω defines the extent of the reductions in $\mathrm{sDSC}(Y_\phi, \tilde{Y}_\phi)$. From left to right, we see how the choice of Φ, and thus an example subset ϕ in blue, is different between the image-wise ($\mathrm{DL}_{\mathbb{I}}$), class-wise ($\mathrm{DL}_{\mathbb{CI}}$), batch-wise ($\mathrm{DL}_{\mathbb{BI}}$) and all-wise ($\mathrm{DL}_{\mathbb{BCI}}$) implementation of DL.

This formulation of $\mathrm{DSC}(Y_\phi, \tilde{Y}_\phi)$ requires Y and $\tilde{Y}$ to contain values in $\{0, 1\}$. In order to be differentiable and handle values in $[0, 1]$, relaxations such as the *soft* DSC (sDSC) are used [5,13]. Furthermore, in order to allow both Y and $\tilde{Y}$ to be empty, a smoothing term ϵ is added to the nominator and denominator such that $\mathrm{DSC}(Y_\phi, \tilde{Y}_\phi) = 1$ in case both Y and $\tilde{Y}$ are empty. This results in the more general formulation of the Dice loss (DL) computed over a number of subsets $\Phi = \{\phi\}$:

$$\mathrm{DL}(Y, \tilde{Y}) = 1 - \frac{1}{|\Phi|} \sum_{\phi \in \Phi} \mathrm{sDSC}(Y_\phi, \tilde{Y}_\phi) = 1 - \frac{1}{|\Phi|} \sum_{\phi \in \Phi} \frac{2 \sum_{\varphi \in \phi} y_\varphi \tilde{y}_\varphi + \epsilon}{\sum_{\varphi \in \phi} (y_\varphi + \tilde{y}_\varphi) + \epsilon}. \quad (3)$$

Note that typically all ϕ are equal in size and define a partition over the domain Ω, such that $\bigcup_{\phi \in \Phi} \phi = \Omega$ and $\bigcap_{\phi \in \Phi} \phi = 0$. In $d\mathrm{DL}(Y, \tilde{Y})/d\theta$ from Eq. 1, the derivative $\partial \mathrm{DL}(Y, \tilde{Y})/\partial \tilde{y}_\omega$ acts as a scaling factor. In order to understand the underlying optimization mechanisms we can thus analyze $\partial \mathrm{DL}(Y, \tilde{Y})/\partial \tilde{y}_\omega$. Given that all ϕ are disjoint, this can be written as:

$$\frac{\partial \mathrm{DL}(Y, \tilde{Y})}{\partial \tilde{y}_\omega} = -\frac{1}{|\Phi|} \left(\frac{2 y_\omega}{\sum_{\varphi \in \phi^\omega} (y_\varphi + \tilde{y}_\varphi) + \epsilon} - \frac{2 \sum_{\varphi \in \phi^\omega} y_\varphi \tilde{y}_\varphi + \epsilon}{\left(\sum_{\varphi \in \phi^\omega} (y_\varphi + \tilde{y}_\varphi) + \epsilon \right)^2} \right), \quad (4)$$

with ϕ^ω the subset that contains ω. As such, it becomes clear that the specific action of DL depends on the exact configuration of the partition Φ of Ω and the choice of ϵ. Next, we describe the most common choices of Φ and ϵ in practice. Then, we investigate their effects in the context of missing or empty labels. Finally, we present a simple heuristic to tune both.

2.1 Configuration of Φ and ϵ in Practice

In Fig. 1, we depict four straightforward choices for Φ. We define these as the *image-wise*, *class-wise*, *batch-wise* or *all-wise* DL implementation, respectively $\mathrm{DL}_{\mathbb{I}}$, $\mathrm{DL}_{\mathbb{CI}}$, $\mathrm{DL}_{\mathbb{BI}}$ and $\mathrm{DL}_{\mathbb{BCI}}$, thus referring to the dimensions over which a complete reduction (i.e. the summations $\sum_{\varphi \in \phi}$ in Eq. 3 and Eq. 4) is performed.

We see that in all cases, a complete reduction is performed over the set of image indices $\mathcal{I}$, which is in line with all relevant literature that we consulted. Furthermore, while in most implementations $\mathrm{B} > 1$, only in [11] the exact usage of the batch dimension is described. In fact, they experimented with both $\mathrm{DL}_{\mathbb{I}}$ and $\mathrm{DL}_{\mathbb{BI}}$, and found the latter to be superior for head and neck organs at risk segmentation in radiotherapy. Based on the context, we assume that most other contributions [5,6,9,10,13,18] used $\mathrm{DL}_{\mathbb{I}}$, although we cannot rule out the use of $\mathrm{DL}_{\mathbb{BI}}$. Similarly, we assume that in [16] $\mathrm{DL}_{\mathbb{CI}}$ was used (with additionally weighting the contribution of each class inversely proportional to the object size), although we cannot rule out the use of $\mathrm{DL}_{\mathbb{BCI}}$.

Note that in Eq. 3 and Eq. 4 we have assumed the choice for $\varPhi$ and ϵ to be fixed. As such, the loss value or gradients only vary across different iterations due to a different sampling of Y and $\tilde{Y}$. Relaxing this assumption allows us to view the *leaf Dice loss* from [7] as a special case of choosing $\varPhi$. Being developed in the context of missing labels, the partition $\varPhi$ of $\varOmega$ is altered each iteration by substituting each ϕ with $\emptyset$ if $\sum_{\varphi}^{\phi} y_{\varphi} = 0$. Similarly, the *marginal Dice loss* from [15] adapts $\varPhi$ every iteration by treating the missing labels as background and summing the predicted probabilities of unlabeled classes to the background prediction before calculating the loss.

Based on our own experience, ϵ is generally chosen to be small (e.g. 10^{-7}). However, most research does not include ϵ in their loss formulation, nor do they mention its exact value. We do find brief mentions related to convergence issues [9] (without further information) or numerical stability in the case of empty labels [10,16] (to avoid division by zero in Eq. 3 and Eq. 4).

2.2 Effect of $\varPhi$ and ϵ on Missing or Empty Labels

When inspecting the derivative given in Eq. 4, we notice that in a way $\partial\mathrm{DL}/\partial\tilde{y}_{\omega}$ does not depend on $\tilde{y}_{\omega}$ itself. Instead, the contributions of $\tilde{y}_{\varphi}$ are aggregated over the reduction dimensions, resulting in a global effect of prediction $\tilde{Y}_{\phi}$. Consequently, the derivative in a subset ϕ takes only two distinct values corresponding to $y_{\omega} = 0$ or $y_{\omega} = 1$. This is in contrast to the derivative shown in [13] who used a L^2 norm-based relaxation, which causes the gradients to be different for every ω if $\tilde{y}_{\omega}$ is different. If we work further with the L^1 norm-based relaxation (following the vast majority of implementations) and assuming that $\sum_{\varphi \in \phi^{\omega}} \tilde{y}_{\varphi} \gg \epsilon$, we see that $\partial\mathrm{DL}/\partial\tilde{y}_{\omega}$ will be negligible for missing or empty ground truth labels. Exploiting this property, we can either avoid having to implement specific losses for missing labels, or we can learn to predict empty maps with a good configuration of $\varPhi$. Regarding the former, we simply need to make sure $\sum_{\varphi \in \phi^{\omega}} y_{\varphi} = 0$ for each map that contains missing labels which can be achieved by using the image-wise implementation $\mathrm{DL}_{\mathbb{I}}$. Regarding the latter, non-zero gradients are required for empty maps. Hence, we want to choose ϕ large enough to avoid $\sum_{\varphi \in \phi^{\omega}} y_{\varphi} = 0$ for which a batch-wise implementation $\mathrm{DL}_{\mathbb{BI}}$ is suitable.

2.3 A Simple Heuristic for Tuning ϵ to Learn from Empty Maps

We hypothesized that we can learn to predict empty maps by using the batch-wise implementation $\mathrm{DL}_{\mathbb{BI}}$. However, due to memory constraints and trade-off with receptive field, it is often not possible to go for large batch sizes. In the limits when $\mathrm{B} = 1$ we find that $\mathrm{DL}_{\mathbb{I}} = \mathrm{DL}_{\mathbb{BI}}$, and thus the gradients of empty maps will be negligible. Hence, we want to mimic the behavior of $\mathrm{DL}_{\mathbb{BI}}$ with $\mathrm{B} \gg 1$, but using $\mathrm{DL}_{\mathbb{I}}$. This can be achieved by tuning ϵ to increase the derivative for empty labels $y_\omega = 0$. A very simple strategy would be to let $\partial \mathrm{DL}(Y, \tilde{Y})/\partial \tilde{y}_\omega$ for $y_\omega = 0$ be equal in case of (i) $\mathrm{DL}_{\mathbb{BI}}$ with infinite batch size such that $\sum_{\varphi \in \phi^\omega} y_\varphi \neq 0$ and negligible ϵ and (ii) $\mathrm{DL}_{\mathbb{I}}$ with non-negligible epsilon and $\sum_{\varphi \in \phi^\omega} y_\varphi = 0$. If we set $\sum_{\varphi \in \phi^\omega} \tilde{y}_\varphi = \hat{v}$ we get:

$$\frac{2\sum_{\varphi \in \phi^\omega} y_\varphi \tilde{y}_\varphi}{\left(\sum_{\varphi \in \phi^\omega} (y_\varphi + \tilde{y}_\varphi)\right)^2} = \frac{\epsilon}{\left(\sum_{\varphi \in \phi^\omega} \tilde{y}_\varphi + \epsilon\right)^2} \Rightarrow \frac{2a\hat{v}}{(b\hat{v})^2} = \frac{\epsilon}{(\hat{v} + \epsilon)^2}, \qquad (5)$$

with a and b variables to express the intersection and union as a function of $\hat{v}$. We can easily see that when we assume the overlap to be around 50%, thus $a \approx 1/2$, and $\sum_{\varphi \in \phi^\omega} y_\varphi \approx \sum_{\varphi \in \phi^\omega} \tilde{y}_\varphi = \hat{v}$, thus $b \approx 2$, we can find $\epsilon \approx \hat{v}$. It is further reasonable to assume that after some iterations $\hat{v} \approx \mathbb{E} \sum_{\varphi \in \phi^\omega} y_\varphi$, thus setting $\epsilon = \hat{v}$ will allow DL to learn empty maps.

3 Experimental Setup

To confirm empirically the observed effects of Φ and ϵ on missing or empty labels (Sect. 2.2), and to test our simple heuristic choice of ϵ (Sect. 2.3), we perform experiments using three implementations of DL on two different public datasets.

Setups $\mathbb{I}$, $\mathbb{BI}$ and $\mathbb{I}_\epsilon$: In $\mathbb{I}$ and $\mathbb{BI}$, respectively $\mathrm{DL}_{\mathbb{I}}$ and $\mathrm{DL}_{\mathbb{BI}}$ are used to calculate the Dice loss (Sect. 2.1). The difference between $\mathbb{I}$ and $\mathbb{I}_\epsilon$ is that we use a negligible value for epsilon $\epsilon = 10^{-7}$ in $\mathbb{I}$ and use the heuristic from Sect. 2.3 to set $\epsilon = \mathbb{E} \sum_{\varphi \in \phi^\omega} y_\varphi$ in $\mathbb{I}_\epsilon$. From Sect. 2.2, we expect $\mathbb{I}$ (any B) and $\mathbb{BI}$ ($\mathrm{B} = 1$) to successfully ignore missing labels during training, still segmenting these at test time. Vice versa, we expect $\mathbb{BI}$ ($\mathrm{B} > 1$) and $\mathbb{I}_\epsilon$ (any B) to successfully learn what maps should be empty and thus output empty maps at test time.

BRATS: For our purpose, we resort to the binary segmentation of whole brain tumors on pre-operative MRI in BRATS 2018 [1,2,12]. The BRATS 2018 training dataset consists of 75 subjects with a lower grade glioma (LGG) and 210 subjects with a glioblastoma (HGG). To construct a partially labeled dataset for the missing and empty label tasks, we substitute the ground truth segmentations of the LGGs with empty maps during training. In light of missing labels, we would like the CNN to successfully segment LGGs at test time. In light of empty maps, we would like the CNN to output empty maps for LGGs at test time.

Based on the ground truths of the entire dataset, in $\mathbb{I}_\epsilon$ we need to set $\epsilon = 8,789$ or $\epsilon = 12,412$ when we use the partially or fully labeled dataset for training, respectively.

ACDC: The ACDC dataset [3] consists of cardiac MRI of 100 subjects. Labels for left ventricular (LV) cavity, LV myocardium and right ventricle (RV) are available in end-diastole (ED) and end-systole (ES). To create a structured partially labeled dataset, we remove the myocardium labels in ES. This is a realistic scenario since segmenting the myocardium only in ED is common in clinical practice. More specifically, ED and ES were sampled in the ratio 3/1 for $\mathbb{I}_\epsilon$, resulting in ϵ being equal to 13,741 and 19,893 on average for the myocardium class during partially or fully labeled training, respectively. For LV and RV, ϵ was 21,339 and 18,993, respectively. We ignored the background map when calculating DL. Since we hypothesize that $\text{DL}_\mathbb{I}$ is able to ignore missing labels, we compare $\mathbb{I}$ to the marginal Dice loss [15] and the leaf Dice loss [7], two loss functions designed in particular to deal with missing labels.

Implementation Details: We start from the exact same preprocessing, CNN architecture and training parameters as in No New-Net [9]. The images of the BRATS dataset were first resampled to an isotropic voxel size of $2 \times 2 \times 2\,\text{mm}^3$, such that we could work with a smaller output segment size of $80 \times 80 \times 48$ voxels as to be able to vary B in $\{1, 2, 4, 8\}$. Since we are working with a binary segmentation task we have $\text{C} = 1$ and use a single sigmoid activation in the final layer. For ACDC, the images were first resampled to $192 \times 192 \times 48$ with a voxel size of $1.56 \times 1.56 \times 2.5\,\text{mm}^3$. The aforementioned CNN architecture was modified to use batch normalization and pReLU activations. To compensate the anisotropic voxel size, we used a combination of $3 \times 3 \times 3$ and $3 \times 3 \times 1$ convolutions and omitted the first max-pooling for the third dimension. These experiments were only performed for $\text{B} = 2$. In this multiclass segmentation task, we use a softmax activation in the final layer to obtain four output maps.

Statistical Performance: All experiments were performed under a five-fold cross-validation scheme, making sure each subject was only present in one of the five partitions. Significant differences were assessed with non-parametric bootstrapping, making no assumptions on the distribution of the results [2]. Results were considered statistically significant if the p-value was below 5%.

4 Results

Table 1 reports the mean DSC and mean volume difference (ΔV) between the fully labeled validation set and the predictions for tumor (BRATS) and myocardium (ACDC). For both the label that was always available (HGG or MYO_{ED}) and the label that was not present in the partially labeled training dataset (LGG or MYO_{ES}), we can make two observations. First, configurations

Table 1. Mean DSC and mean ΔV. HGG and MYO_{ED} are always present during training while LGG and MYO_{ES} are replaced by empty maps under partial labeling. Configurations that we expect to learn to predict empty maps are highlighted (since we used a fully labeled validation set, we expect lower DSC and ΔV). Comparing partial with full labeling, inferior ($p < 0.05$) results are indicated in italic.

	Labeling	B	DSC						ΔV [ml]					
			HGG/MYO_{ED}			LGG/MYO_{ES}			HGG/MYO_{ED}			LGG/MYO_{ES}		
			$\mathbb{I}$	$\mathbb{BI}$	$\mathbb{I}_\epsilon$	$\mathbb{I}$	$\mathbb{BI}$	$\mathbb{I}_\epsilon$	$\mathbb{I}$	$\mathbb{BI}$	$\mathbb{I}_\epsilon$	$\mathbb{I}$	$\mathbb{BI}$	$\mathbb{I}_\epsilon$
BRATS	Full	1	0.89	0.89	0.89	0.89	0.89	0.89	−4	−4	−7	−8	−7	−9
		2	0.89	0.89	0.89	0.89	0.88	0.88	−5	−5	−7	−9	−10	−12
		4	0.89	0.89	0.89	0.88	0.90	0.89	−6	−5	−7	−11	−7	−11
		8	0.89	0.89	0.89	0.89	0.88	0.89	−6	−4	−6	−12	−10	−9
	Partial	1	0.89	0.89	*0.83*	*0.88*	0.88	*0.23*	*−5*	*−5*	*−12*	*−11*	*−12*	*−88*
		2	0.89	*0.83*	*0.82*	0.88	*0.24*	*0.16*	−6	*−12*	*−13*	*−12*	*−89*	*−96*
		4	0.89	*0.82*	*0.83*	0.88	*0.20*	*0.20*	*−6*	*−12*	*−12*	*−15*	*−93*	*−94*
		8	0.89	*0.82*	*0.83*	*0.88*	*0.20*	*0.23*	−6	*−12*	*−12*	−14	*−94*	*−90*
ACDC	Full	2	0.88	0.88	0.87	0.89	0.89	0.89	−1	0	−2	−3	0	−3
	Partial	2	*0.88*	*0.80*	*0.80*	*0.88*	*0.08*	*0.06*	0	*−11*	*−14*	*−5*	*−129*	*−131*

$\mathbb{I}$ and $\mathbb{BI}$ ($B = 1$) delivered a comparable segmentation performance (in terms of both DSC and ΔV) compared to using a fully labeled training dataset. Second, using configurations $\mathbb{BI}$ ($B > 1$) and $\mathbb{I}_\epsilon$ the performance was consistently inferior. In this case, the CNN starts to learn when it needs to output empty maps. As a result, when calculating the DSC and ΔV with respect to a fully labeled validation dataset, we expect both metrics to remain similar for HGG and MYO_{ES}. On the other hand, we expect a mean DSC of 0 and a $|\Delta V|$ close to the mean volume of LGG or MYO_{ES}. Note that this is not the case due to the incorrect classification of LGG or MYO_{ES} as HGG or MYO_{ED}, respectively. Figure 2 shows the Receiver Operating Characteristic (ROC) curves when using a partially labeled training dataset with the goal to detect HGG or MYO_{ED} based on a threshold on the predicted volume at test time. For both tasks, we achieved an Area Under the Curve (AUC) of around 0.9. Figure 3 shows an example segmentation.

When comparing $\mathbb{I}$ with the marginal Dice loss [15] and the leaf Dice loss [7], no significant differences between any method for myocardium ($\text{MYO}_{ED} = 0.88$, $\text{MYO}_{ES} = 0.88$), LV ($\text{LV}_{ED} = 0.96$, $\text{LV}_{ES} = 0.92$) and RV ($\text{RV}_{ED} = 0.93$, $\text{RV}_{ES} = 0.86 - 0.87$) were found in both ED and ES.

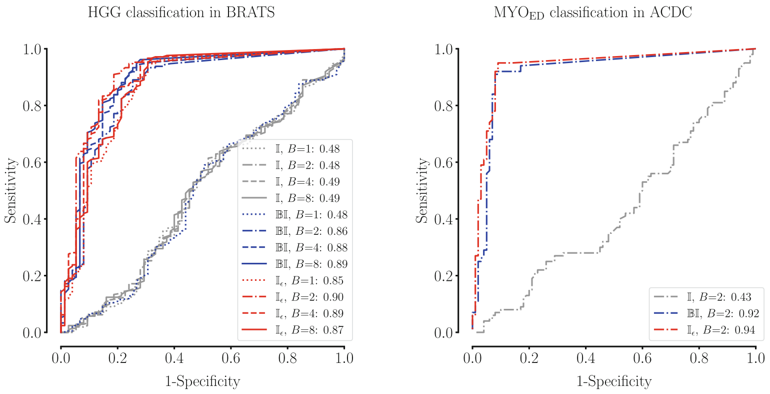

Fig. 2. ROC analysis if we want to detect the label that was always present during training by using different thresholds on the predicted volume. In the legend we also report the AUC for each setting.

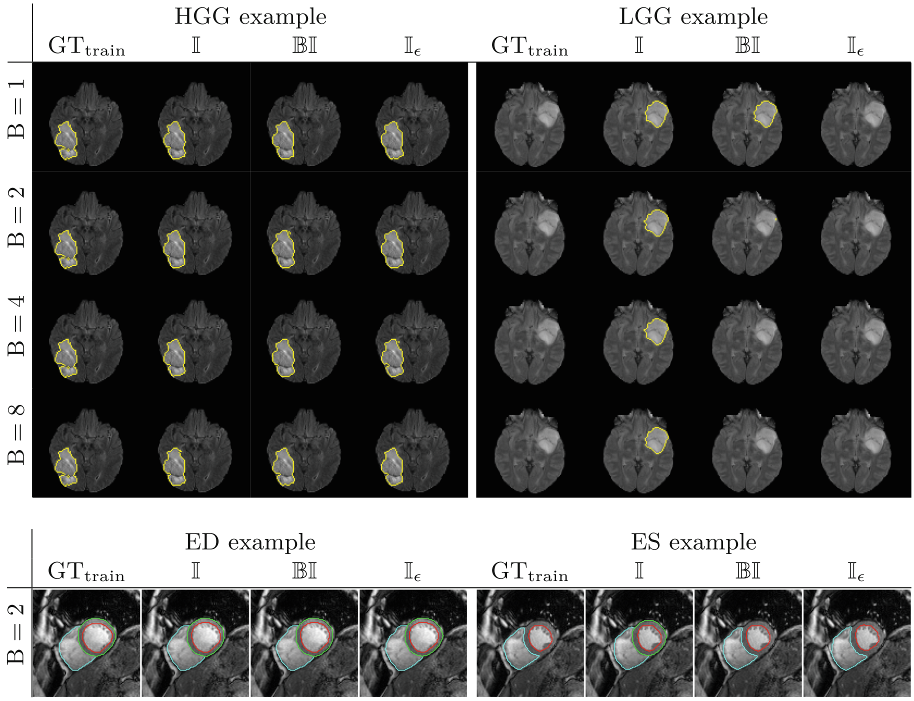

Fig. 3. Segmentation examples for BRATS (top) and ACDC (bottom). The ground truths for LGG and MYO_{ES} were replaced with empty maps during training (GT_{train}).

5 Discussion

The experiments confirmed the analysis from Sect. 2.2 that $\mathrm{DL}_{\mathbb{I}}$ (equal to $\mathrm{DL}_{\mathbb{BI}}$ when $\mathrm{B} = 1$) ignores missing labels during training and that it can be used in the context of missing labels naively. On the other hand, we confirmed that $\mathrm{DL}_{\mathbb{BI}}$ (with $\mathrm{B} > 1$) and $\mathrm{DL}_{\mathbb{I}}$ (with a heuristic choice of ϵ) can effectively learn to predict empty labels, e.g. for classification purposes or to be used with small patch sizes.

When heuristically determining ϵ for configuring $\mathbb{I}_\epsilon$ (Eq. 5), we only focused on the derivative for $y_\omega = 0$. Of course, by adapting ϵ, the derivative for $y_\omega = 1$ will also change. Nonetheless, our experiments showed that $\mathbb{I}_\epsilon$ can achieve the expected behavior, indicating that the effect on the derivative for $y_\omega = 1$ is only minor compared to $y_\omega = 0$. We wish to derive a more exact formulation of the optimal value of ϵ in future work. We expect this optimal ϵ to depend on the distribution between the classes, object size and other labels that might be present. Furthermore, it would be interesting to study the transition between the near-perfect prediction for the missing class ($\mathrm{DL}_{\mathbb{I}}$ with small ϵ) and the prediction of empty labels for the missing class ($\mathrm{DL}_{\mathbb{I}}$ with large ϵ).

All the code necessary for exact replication of the results including preprocessing, training scripts, statistical analysis, etc. was released to encourage further analysis on this topic (https://github.com/JeroenBertels/dicegrad).

6 Conclusion

We showed that the choice of the reduction dimensions Φ and the smoothing term ϵ for the Dice loss is non-trivial and greatly influences its behavior in the context of missing or empty labels. We believe that this work highlights some essential perspectives and hope that it encourages researchers to better describe their exact implementation of the Dice loss in the future.

Acknowledgement. This research received funding from the Flemish Government under the “Onderzoeksprogramma Artificiële intelligentie (AI) Vlaanderen” programme and is also partially funded by KU Leuven Internal Funds C24/18/047 (F. Maes).

References

1. Bakas, S., et al.: Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data **4**(1), 170117 (2017). https://doi.org/10.1038/sdata.2017.117, http://www.nature.com/articles/sdata2017117
2. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv, November 2018. http://arxiv.org/abs/1811.02629

3. Bernard, O., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging **37**(11), 2514–2525 (2018). https://doi.org/10.1109/TMI.2018.2837502
4. 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). https://doi.org/10.1016/j.media.2020.101833, https://linkinghub.elsevier.com/retrieve/pii/S1361841520301973
5. Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., Pal, C.: The importance of skip connections in biomedical image segmentation. In: Carneiro, G., et al. (eds.) LABELS/DLMIA -2016. LNCS, vol. 10008, pp. 179–187. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46976-8_19
6. Eelbode, T., et al.: 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). https://doi.org/10.1109/TMI.2020.3002417, https://ieeexplore.ieee.org/document/9116807/
7. Fidon, L., et al.: Label-set loss functions for partial supervision: application to fetal brain 3D MRI parcellation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 647–657. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_60
8. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016). http://www.deeplearningbook.org
9. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H.: No New-Net. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11384, pp. 234–244. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_21
10. Jadon, S.: A survey of loss functions for semantic segmentation. In: 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2020 (2020). https://doi.org/10.1109/CIBCB48159.2020.9277638
11. Kodym, O., Španěl, M., Herout, A.: Segmentation of head and neck organs at risk using CNN with batch dice loss. In: Brox, T., Bruhn, A., Fritz, M. (eds.) GCPR 2018. LNCS, vol. 11269, pp. 105–114. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-12939-2_8
12. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2015). https://doi.org/10.1109/TMI.2014.2377694, http://ieeexplore.ieee.org/document/6975210/
13. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016, pp. 565–571 (2016). https://doi.org/10.1109/3DV.2016.79
14. Nowozin, S.: Optimal decisions from probabilistic models: the intersection-over-union case. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 548–555. IEEE, June 2014. https://doi.org/10.1109/CVPR.2014.77, http://ieeexplore.ieee.org/document/6909471/
15. Shi, G., Xiao, L., Chen, Y., Zhou, S.K.: Marginal loss and exclusion loss for partially supervised multi-organ segmentation. Med. Image Anal. **70**, 101979 (2021). https://doi.org/10.1016/j.media.2021.101979
16. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 240–248. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_28

17. Tarlow, D., Adams, R.P.: Revisiting uncertainty in graph cut solutions. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2440–2447. IEEE, June 2012. https://doi.org/10.1109/CVPR.2012.6247958, http://ieeexplore.ieee.org/document/6247958/
18. 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 Med. Imaging Graph. **95**, 102026 (2021, 2022). https://doi.org/10.1016/j.compmedimag.2021.102026