Dice Semimetric Losses: Optimizing the Dice Score with Soft Labels

Zifu Wang[1(✉)], Teodora Popordanoska[1], Jeroen Bertels[1], Robin Lemmens[2,3], and Matthew B. Blaschko[1]

[1] ESAT-PSI, KU Leuven, Leuven, Belgium
zifu.wang@kuleuven.be, teodora.popordanoska@kuleuven.be, jeroen.bertels@kuleuven.be, matthewb.blaschko@kuleuven.be
[2] Department of Neurosciences, KU Leuven, Leuven, Belgium
robin.lemmens@kuleuven.be
[3] Department of Neurology, UZ Leuven, Leuven, Belgium

Abstract. The soft Dice loss (SDL) has taken a pivotal role in numerous automated segmentation pipelines in the medical imaging community. Over the last years, some reasons behind its superior functioning have been uncovered and further optimizations have been explored. However, there is currently no implementation that supports its direct utilization in scenarios involving soft labels. Hence, a synergy between the use of SDL and research leveraging the use of soft labels, also in the context of model calibration, is still missing. In this work, we introduce Dice semimetric losses (DMLs), which (i) are by design identical to SDL in a standard setting with hard labels, but (ii) can be employed in settings with soft labels. Our experiments on the public QUBIQ, LiTS and KiTS benchmarks confirm the potential synergy of DMLs with soft labels (e.g. averaging, label smoothing, and knowledge distillation) over hard labels (e.g. majority voting and random selection). As a result, we obtain superior Dice scores and model calibration, which supports the wider adoption of DMLs in practice. The code is available at https://github.com/zifuwanggg/JDTLosses.

Keywords: Dice Score · Dice Loss · Soft Labels · Model Calibration

1 Introduction

Image segmentation is a fundamental task in medical image analysis. One of the key design choices in many segmentation pipelines that are based on neural networks lies in the selection of the loss function. In fact, the choice of loss function goes hand in hand with the metrics chosen to assess the quality of the predicted segmentation [46]. The intersection-over-union (IoU) and the Dice

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43898-1_46.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14222, pp. 475–485, 2023.
https://doi.org/10.1007/978-3-031-43898-1_46

score are commonly used metrics because they reflect both size and localization agreement, and they are more in line with perceptual quality compared to, e.g., pixel-wise accuracy [9,27]. Consequently, directly optimizing the IoU or the Dice score using differentiable surrogates as (a part of) the loss function has become prevalent in semantic segmentation [2,9,20,24,47]. In medical imaging in particular, the Dice score and the soft Dice loss (SDL) [30,42] have become the standard practice, and some reasons behind its superior functioning have been uncovered and further optimizations have been explored [3,9,45].

Another mechanism to further improve the predicted segmentation that has gained significant interest in recent years, is the use of soft labels during training. Soft labels can be the result of data augmentation techniques such as label smoothing (LS) [21,43] and are integral to regularization methods such as knowledge distillation (KD) [17,36]. Their role is to provide additional regularization so as to make the model less prone to overfitting [17,43] and to combat overconfidence [14], e.g., providing superior model calibration [31]. In medical imaging, soft labels emerge not only from LS or KD, but are also present inherently due to considerable intra- and inter-rater variability. For example, multiple annotators often disagree on organ and lesion boundaries, and one can average their annotations to obtain soft label maps [12,23,25,41].

This work investigates how the medical imaging community can combine the use of SDL with soft labels to reach a state of synergy. While the original SDL surrogate was posed as a relaxed form of the Dice score, naively inputting soft labels to SDL is possible (e.g. in open-source segmentation libraries [6,19,20,51]), but it tends to push predictions towards 0–1 outputs rather than make them resemble the soft labels [3,32,47]. Consequently, the use of SDL when dealing with soft labels might not align with a user's expectations, with potential adverse effects on the Dice score, model calibration and volume estimation [3].

Motivated by this observation, we first (in Sect. 2) propose two probabilistic extensions of SDL, namely, Dice semimetric losses (DMLs). These losses satisfy the conditions of a semimetric and are fully compatible with soft labels. In a standard setting with hard labels, DMLs are identical to SDL and can safely replace SDL in existing implementations. Secondly (in Sect. 3), we perform extensive experiments on the public QUBIQ, LiTS and KiTS benchmarks to empirically confirm the potential synergy of DMLs with soft labels (e.g. averaging, LS, KD) over hard labels (e.g. majority voting, random selection).

2 Methods

We adopt the notation from [47]. In particular, we denote the predicted segmentation as $\dot{x} \in \{1, ..., C\}^p$ and the ground-truth segmentation as $\dot{y} \in \{1, ..., C\}^p$, where C is the number of classes and p the number of pixels. For a class c, we define the set of predictions as $x^c = \{\dot{x} = c\}$, the set of ground-truth as $y^c = \{\dot{y} = c\}$, the union as $u^c = x^c \cup y^c$, the intersection as $v^c = x^c \cap y^c$, the symmetric difference (i.e., the set of mispredictions) as $m^c = (x^c \setminus y^c) \cup (y^c \setminus x^c)$, the Jaccard index as $\text{IoU}^c = \frac{|v^c|}{|u^c|}$, and the Dice score as $\text{Dice}^c = \frac{2\text{IoU}^c}{1+\text{IoU}^c} = \frac{2|v^c|}{|x^c|+|y^c|}$. In what follows, we will represent sets as binary vectors $x^c, y^c, u^c, v^c, m^c \in \{0,1\}^p$ and denote $|x^c| = \sum_{i=1}^p x_i^c$ the cardinality of the relevant set. Moreover, when the context is clear, we will drop the superscript c.

2.1 Existing Extensions

If we want to optimize the Dice score, hence, minimize the Dice loss $\Delta_{\text{Dice}} = 1 - \text{Dice}$ in a continuous setting, we need to extend Δ_{Dice} with $\overline{\Delta}_{\text{Dice}}$ such that it can take any predicted segmentation $\tilde{x} \in [0,1]^p$ as input. Hereinafter, when there is no ambiguity, we will use x and $\tilde{x}$ interchangeably.

The soft Dice loss (SDL) [42] extends Δ_{Dice} by realizing that when $x, y \in \{0,1\}^p$, $|v| = \langle x, y\rangle$, $|x| = \|x\|_1$ and $|y| = \|y\|_1$. Therefore, SDL replaces the set notation with vector functions:

$$\overline{\Delta}_{\text{SDL}} : x \in [0,1]^p, y \in \{0,1\}^p \mapsto 1 - \frac{2\langle x, y\rangle}{\|x\|_1 + \|y\|_1}. \tag{1}$$

The soft Jaccard loss (SJL) [33,37] can be defined in a similar way:

$$\overline{\Delta}_{\text{SJL}} : x \in [0,1]^p, y \in \{0,1\}^p \mapsto 1 - \frac{\langle x, y\rangle}{\|x\|_1 + \|y\|_1 - \langle x, y\rangle}. \tag{2}$$

A major limitation of loss functions based on L^1 relaxations, including SDL, SJL, the soft Tversky loss [39] and the focal Tversky loss [1], as well as those relying on the Lovasz extension, such as the Lovasz hinge loss [49], the Lovasz-Softmax loss [2] and the PixIoU loss [50], is that they cannot handle soft labels [47]. That is, when y is also in $[0,1]^p$. In particular, both SDL and SJL do not reach their minimum at $x = y$, but instead they drive x towards the vertices $\{0,1\}^p$ [3,32,47]. Take for example $y = 0.5$; it is straightforward to verify that SDL achieves its minimum at $x = 1$, which is clearly erroneous.

Loss functions that utilize L^2 relaxations [9,30] do not exhibit this problem [47], but they are less commonly employed in practice and are shown to be inferior to their L^1 counterparts [9,47]. To address this, Wang and Blaschko [47] proposed two variants of SJL termed as Jaccard Metric Losses (JMLs). These two variants, $\overline{\Delta}_{\text{JML},1}$ and $\overline{\Delta}_{\text{JML},2} : [0,1]^p \times [0,1]^p \to [0,1]$ are defined as

$$\overline{\Delta}_{\text{JML},1} = 1 - \frac{\|x+y\|_1 - \|x-y\|_1}{\|x+y\|_1 + \|x-y\|_1}, \quad \overline{\Delta}_{\text{JML},2} = 1 - \frac{\|x \odot y\|_1}{\|x \odot y\|_1 + \|x-y\|_1}. \tag{3}$$

JMLs are shown to be a metric on $[0,1]^p$, according to the definition below.

Definition 1 (Metric [8]). *A mapping $f : [0,1]^p \times [0,1]^p \to \mathbb{R}$ is called a metric if it satisfies the following conditions for all $a, b, c \in [0,1]^p$:*

(i) (Reflexivity). $f(a,a) = 0$.
(ii) (Positivity). If $a \neq b$, then $f(a,b) > 0$.
(iii) (Symmetry). $f(a,b) = f(b,a)$.
(iv) (Triangle inequality). $f(a,c) \leq f(a,b) + f(b,c)$.

Note that reflexivity and positivity jointly imply $x = y \Leftrightarrow f(x,y) = 0$, hence, a loss function that satisfies these conditions will be compatible with soft labels.

2.2 Dice Semimetric Losses

We focus here on the Dice loss. For the derivation of the Tversky loss and the focal Tversky loss, please refer to our full paper on arXiv.

Since $\text{Dice} = \frac{2\text{IoU}}{1+\text{IoU}} \Rightarrow 1 - \text{Dice} = \frac{1-\text{IoU}}{2-(1-\text{IoU})}$, we have $\overline{\Delta}_{\text{Dice}} = \frac{\overline{\Delta}_{\text{IoU}}}{2-\overline{\Delta}_{\text{IoU}}}$. There exist several alternatives to define $\overline{\Delta}_{\text{IoU}}$, but not all of them are feasible, e.g., SJL. Generally, it is easy to verify the following proposition:

Proposition 1. *$\overline{\Delta}_{Dice}$ satisfies reflexivity and positivity iff $\overline{\Delta}_{IoU}$ does.*

Among the definitions of $\overline{\Delta}_{\text{IoU}}$, Wang and Blaschko [47] found only two candidates as defined in Eq. (3) satisfy reflexivity and positivity. Following Proposition 1, we transform these two IoU losses and define Dice semimetric losses (DMLs) $\overline{\Delta}_{\text{DML},1}, \overline{\Delta}_{\text{DML},2} : [0,1]^p \times [0,1]^p \rightarrow [0,1]$ as

$$\overline{\Delta}_{\text{DML},1} = 1 - \frac{\|x+y\|_1 - \|x-y\|_1}{\|x+y\|_1}, \quad \overline{\Delta}_{\text{DML},2} = 1 - \frac{2\|xy\|_1}{2\|xy\|_1 + \|x-y\|_1}. \tag{4}$$

Δ_{Dice} that is defined over integers does not satisfy the triangle inequality [11], which is shown to be helpful in KD [47]. Nonetheless, we can consider a weaker form of the triangle inequality:

$$f(a,c) \leq \rho(f(a,b) + f(b,c)). \tag{5}$$

Functions that satisfy the relaxed triangle inequality for some fixed scalar ρ and conditions (i)-(iii) of a metric are called semimetrics. Δ_{Dice} is a semimetric on $\{0,1\}^p$ [11]. $\overline{\Delta}_{\text{DML},1}$ and $\overline{\Delta}_{\text{DML},2}$, which extend Δ_{Dice} to $[0,1]^p$, remain semimetrics in the continuous space:

Theorem 1. *$\overline{\Delta}_{DML,1}$ and $\overline{\Delta}_{DML,2}$ are semimetrics on $[0,1]^p$.*

The proof can be found in Appendix A. Moreover, DMLs have properties that are similar to JMLs and they are presented as follows:

Theorem 2. *$\forall x \in [0,1]^p,\ y \in \{0,1\}^p$ and $x \in \{0,1\}^p,\ y \in [0,1]^p$, $\overline{\Delta}_{SDL} = \overline{\Delta}_{DML,1} = \overline{\Delta}_{DML,2}$. $\exists x, y \in [0,1]^p, \overline{\Delta}_{SDL} \neq \overline{\Delta}_{DML,1} \neq \overline{\Delta}_{DML,2}$.*

Theorem 3. *$\forall x, y \in [0,1]^p$, $\overline{\Delta}_{DML,1} \leq \overline{\Delta}_{DML,2}$.*

The proofs are similar to those given in [47]. Importantly, Theorem 2 indicates that we can safely substitute the existing implementation of SDL with DMLs and no change will be incurred, as they are identical when only hard labels are presented.

3 Experiments

In this section, we provide empirical evidence of the benefits of using soft labels. In particular, using QUBIQ [29], which contains multi-rater information, we show that models trained with averaged annotation maps can significantly surpass those trained with majority votes and random selections. Leveraging LiTS [4] and KiTS [16], we illustrate the synergistic effects of integrating LS and KD with DMLs.

3.1 Datasets

QUBIQ is a recent challenge held at MICCAI 2020 and 2021, specifically designed to evaluate the inter-rater variability in medical imaging. Following [23, 41], we use QUBIQ 2020, which contains 7 segmentation tasks in 4 different CT and MR datasets: Prostate (55 cases, 2 tasks, 6 raters), Brain Growth (39 cases, 1 task, 7 raters), Brain Tumor (32 cases, 3 tasks, 3 raters), and Kidney (24 cases, 1 task, 3 raters). For each dataset, we calculate the average Dice score between each rater and the majority votes in Table 1. In some datasets, such as Brain Tumor T2, the inter-rater disagreement can be quite substantial. In line with [23], we resize all images to 256×256.

LiTS contains 201 high-quality CT scans of liver tumors. Out of these, 131 cases are designated for training and 70 for testing. As the ground-truth labels for the test set are not publicly accessible, we only use the training set. Following [36], all images are resized to 512×512 and the HU values of CT images are windowed to the range of [−60, 140].

KiTS includes 210 annotated CT scans of kidney tumors from different patients. In accordance with [36], all images are resized to 512×512 and the HU values of CT images are windowed to the range of [−200, 300].

3.2 Implementation Details

We adopt a variety of backbones including ResNet50/18 [15], EfficientNetB0 [44] and MobileNetV2 [40]. All these models that have been pretrained on ImageNet [7] are provided by timm library [48]. We consider both UNet [38] and DeepLabV3+ [5] as the segmentation method.

We train the models using SGD with an initial learning rate of 0.01, momentum of 0.9, and weight decay of 0.0005. The learning rate is decayed in a poly policy with an exponent of 0.9. The batch size is set to 8 and the number of epochs is 150 for QUBIQ, 60 for both LiTS and KiTS. We leverage a mixture of CE and DMLs weighted by 0.25 and 0.75, respectively. Unless otherwise specified, we use $\overline{\Delta}_{\mathrm{DML},1}$ by default.

In this work, we are mainly interested in how models can benefit from the use of soft labels. The superiority of SDL over CE has been well established in the medical imaging community [9, 20], and our preliminary experiments also confirm this, as shown in Table 5 (Appendix C). Therefore, we do not include any further comparison with CE in this paper.

Table 1. The number of raters and the averaged Dice score between each rater and the majority votes for each QUBIQ dataset. D1: Prostate T1, D2: Prostate T2, D3: Brain Growth T1, D4: Brain Tumor T1, D5: Brain Tumor T2, D6: Brain Tumor T3, D7: Kidney T1.

Dataset	D1	D2	D3	D4	D5	D6	D7
# Raters	6	6	7	3	3	3	3
Dice (%)	96.49	92.17	91.20	95.44	68.73	92.71	97.41

3.3 Evaluation

We report both the Dice score and the expected calibration error (ECE) [14]. For QUBIQ experiments, we additionally present the binarized Dice score (BDice), which is the official evaluation metrics used in the QUBIQ challenge. To compute BDice, both predictions and soft labels are thresholded at different probability levels (0.1, 0.2, ..., 0.8, 0.9). We then compute the Dice score at each level and average these scores with all thresholds.

For all experiments, we conduct 5-fold cross validation, making sure that each case is presented in exactly one validation set, and report the mean values in the aggregated validation set. We perform statistical tests according to the procedure detailed in [9] and highlight results that are significantly superior (with a significance level of 0.05) in red.

3.4 Results on QUBIQ

In Table 2, we compare different training methods on QUBIQ using UNet-ResNet50. This comparison includes both hard labels, obtained through (i) majority votes [25] and (ii) random sampling each rater's annotation [22], as well as soft labels derived from (i) averaging across all annotations [12,25,41] and (ii) label smoothing [43].

In the literature [12,25,41], annotations are usually averaged with uniform weights. We additionally consider weighting each rater's annotation by its Dice score with respect to the majority votes, so that a rater who deviates far from the majority votes receives a low weight. Note that for all methods, the Dice score and ECE are computed with respect to the majority votes, while BDice is calculated as illustrated in Sect. 3.3.

Generally, models trained with soft labels exhibit improved accuracy and calibration. In particular, averaging annotations with uniform weights obtains the highest BDice, while a weighted average achieves the highest Dice score. It is worth noting that the weighted average significantly outperforms the majority votes in terms of the Dice score which is evaluated based on the majority votes themselves. We hypothesize that this is because soft labels contain extra inter-rater information, which can ease the network optimization at those ambiguous regions. Overall, we find the weighted average outperforms other methods, with the exception of Brain Tumor T2, where there is a high degree of disagreement among raters.

We compare our method with state-of-the-art (SOTA) methods using UNet-ResNet50 in Table 3. In our method, we average annotations with uniform weights for Brain Tumor T2 and with each rater's Dice score for all other datasets. Our method, which simply averages annotations to produce soft labels obtains superior results compared to methods that adopt complex architectures or training techniques.

Table 2. Comparing hard labels with soft labels on QUBIQ using UNet-ResNet50.

Dataset	Metric	Majority	Random	Uniform	Weighted	LS
Prostate T1	Dice (%)	95.65	95.80	95.74	95.99	95.71
	BDice (%)	94.72	95.15	95.19	95.37	94.91
	ECE (%)	0.51	0.39	0.22	0.20	0.36
Prostate T2	Dice (%)	89.39	88.87	89.57	89.79	89.82
	BDice (%)	88.31	88.23	89.35	89.66	88.85
	ECE (%)	0.52	0.47	0.26	0.25	0.41
Brain Growth	Dice (%)	91.09	90.65	90.94	91.46	91.23
	BDice (%)	88.72	88.81	89.89	90.40	89.88
	ECE (%)	1.07	0.85	0.27	0.34	0.41
Brain Tumor T1	Dice (%)	86.46	87.24	87.74	87.78	87.84
	BDice (%)	85.74	86.59	86.67	86.92	86.91
	ECE (%)	0.62	0.55	0.38	0.36	0.37
Brain Tumor T2	Dice (%)	58.58	48.86	52.42	61.01	61.23
	BDice (%)	38.68	49.19	55.11	44.23	40.61
	ECE (%)	0.25	0.81	0.74	0.26	0.22
Brain Tumor T3	Dice (%)	53.54	54.64	53.45	56.75	57.01
	BDice (%)	52.33	53.53	51.98	53.90	55.26
	ECE (%)	0.17	0.17	0.14	0.09	0.11
Kidney	Dice (%)	62.96	68.10	71.33	76.18	71.21
	BDice (%)	62.47	67.69	70.82	75.67	70.41
	ECE (%)	0.88	0.78	0.67	0.53	0.62
All	Dice (%)	76.80	76.30	77.31	79.85	79.15
	BDice (%)	72.99	75.59	77.00	76.59	75.26
	ECE (%)	0.57	0.57	0.38	0.29	0.35

Table 3. Comparing SOTA methods with ours on QUBIQ using UNet-ResNet50. All results are BDice (%).

Dataset	Dropout [10]	Multi-head [13]	MRNet [23]	SoftSeg [12,25]	Ours
Prostate T1	94.91	95.18	95.21	95.02	95.37
Prostate T2	88.43	88.32	88.65	88.81	89.66
Brain Growth	88.86	89.01	89.24	89.36	90.40
Brain Tumor T1	85.98	86.45	86.33	86.41	86.92
Brain Tumor T2	48.04	51.17	51.82	52.56	55.11
Brain Tumor T3	52.49	53.68	54.22	52.43	53.90
Kidney	66.53	68.00	68.56	69.83	75.67
All	75.03	75.97	76.18	76.34	78.14

3.5 Results on LiTS and KiTS

Wang and Blaschko [47] empirically found that a well-calibrated teacher can distill a more accurate student. Concurrently, Menon et al. [28] argued that the effectiveness of KD arises from the teacher providing an estimation of the Bayes class-probabilities $p^*(y|x)$ and this can lower the variance of the student's empirical loss.

Table 4. Comparing hard labels with LS and KD on LiTS and KiTS.

Method	Backbone	~~Metric~~	LiTS			KiTS		
			Hard	LS	KD	Hard	LS	KD
UNet	ResNet50	Dice (%)	59.79	60.59	-	72.66	73.92	-
		ECE (%)	0.51	0.49	-	0.39	0.33	-
UNet	ResNet18	Dice (%)	57.92	58.60	60.30	67.96	69.09	71.34
		ECE (%)	0.52	0.48	0.50	0.44	0.38	0.44
UNet	EfficientNetB0	Dice (%)	56.90	57.66	60.11	70.31	71.12	71.73
		ECE (%)	0.56	0.47	0.52	0.39	0.35	0.39
UNet	MobileNetV2	Dice (%)	56.16	57.20	58.92	67.46	68.19	68.85
		ECE (%)	0.54	0.48	0.50	0.42	0.38	0.41
DeepLabV3+	ResNet18	Dice (%)	56.10	57.07	59.12	69.95	70.61	70.80
		ECE (%)	0.53	0.50	0.52	0.40	0.38	0.40

In line with these findings, in Appendix B, we prove $|\mathbb{E}[p^*(y|x) - f(x)]| \leq \mathbb{E}[|\mathbb{E}[y|f(x)] - f(x)|]$. That is, the bias of the estimation is bounded above by the calibration error and this explains why the calibration of the teacher would be important for the student. Inspired by this, we apply a recent kernel density estimator (KDE) [35] that provides consistent estimation of $\mathbb{E}[y|f(x)]$. We then adopt it as a post-hoc calibration method to replace the temperature scaling to calibrate the teacher in order to improve the performance of the student. For more details of KDE, please refer to our full paper on arXiv.

In Table 4, we compare models trained with hard labels, LS [43] and KD [17] on LiTS and KiTS, respectively. For all KD experiments, we use UNet-ResNet50 as the teacher. Again, we obtain noticeable improvements in both the Dice score and ECE. It is worth noting that for UNet-ResNet18 and UNet-EfficientNetB0 on LiTS, the student's Dice score exceeds that of the teacher.

3.6 Ablation Studies

In Table 6 (Appendix C), we compare SDL with DMLs. For QUBIQ, we train UNet-ResNet50 with soft labels obtained from weighted average and report BDice. For LiTS and KiTS, we train UNet-ResNet18 with KD and present the Dice score. For a fair comparison, we disable KDE in all KD experiments.

We find models trained with SDL can still benefit from soft labels to a certain extent because (i) models are trained with a mixture of CE and SDL, and CE is compatible with soft labels; (ii) although SDL pushes predictions towards vertices, it can still add some regularization effects in a binary segmentation setting. However, SDL is notably outperformed by DMLs. As for DMLs, we find $\overline{\Delta}_{\mathrm{DML},1}$ is slightly superior to $\overline{\Delta}_{\mathrm{DML},2}$ and recommend using $\overline{\Delta}_{\mathrm{DML},1}$ in practice.

In Table 7 (Appendix C), we ablate the contribution of each KD term on LiTS and KiTS with a UNet-ResNet18 student. In the table, CE and DML represent adding the CE and DML term between the teacher and the student, respectively. In Table 8 (Appendix C), we illustrate the effect of bandwidth that controls the smoothness of KDE. Results shown in the tables verify the effectiveness of the proposed loss and the KDE method.

4 Future Works

In this study, our focus is on extending the Dice loss within the realm of medical image segmentation. It may be intriguing to apply DMLs in the context of long-tailed classification [26]. Additionally, while we employ DMLs in the label space, it holds potential for measuring the similarity of two feature vectors [18], for instance, as an alternative to cosine similarity.

5 Conclusion

In this work, we introduce the Dice semimetrics losses (DMLs), which are identical to the soft Dice loss (SDL) in a standard setting with hard labels, but are fully compatible with soft labels. Our extensive experiments on the public QUBIQ, LiTS and KiTS benchmarks validate that incorporating soft labels leads to higher Dice score and lower calibration error, indicating that these losses can find wide application in diverse medical image segmentation problems. Hence, we suggest to replace the existing implementation of SDL with DMLs.

Acknowledgements. We acknowledge support from the Research Foundation - Flanders (FWO) through project numbers G0A1319N and S001421N, and funding from the Flemish Government under the Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen programme. The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government.

References

1. Abraham, N., Khan, N.M.: A novel focal Tversky loss function with improved attention U-Net for lesion segmentation. In: ISBI (2019)
2. Berman, M., Triki, A.R., Blaschko, M.B.: The Lovasz-softmax loss: a tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In: CVPR (2018)
3. 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. MIA **67**, 101833 (2021)
4. Bilic, P., et al.: The liver tumor segmentation benchmark (LiTS). MIA **84**, 102680 (2023)
5. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 833–851. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_49
6. Contributors, M.: MMSegmentation: OpenMMLab semantic segmentation toolbox and benchmark (2020). https://github.com/open-mmlab/mmsegmentation
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR (2009)
8. Deza, M.M., Deza, E.: Encyclopedia of Distances. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-00234-2

9. Eelbode, T., et al.: Optimization for medical image segmentation: theory and practice when evaluating with dice score or Jaccard index. TMI **39**, 3679–3690 (2020)
10. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: ICML (2016)
11. Gragera, A., Suppakitpaisarn, V.: Relaxed triangle inequality ratio of the Sørensen-Dice and Tversky indexes. TCS **718**, 37–45 (2018)
12. Gros, C., Lemay, A., Cohen-Adad, J.: SoftSeg: advantages of soft versus binary training for image segmentation. MIA **71**, 102038 (2021)
13. Guan, M.Y., Gulshan, V., Dai, A.M., Hinton, G.E.: Who said what: modeling individual labelers improves classification. In: AAAI (2018)
14. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML (2017)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
16. Heller, N., et al.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: results of the KiTS19 challenge. MIA **67**, 101821 (2021)
17. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NeurIPS Workshop (2015)
18. Huang, T., et al.: Masked distillation with receptive tokens. In: ICLR (2023)
19. Iakubovskii, P.: Segmentation models PyTorch (2019). https://github.com/qubvel/segmentation_models.pytorch
20. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Meth. **18**, 203–211 (2021)
21. Islam, M., Glocker, B.: Spatially varying label smoothing: capturing uncertainty from expert annotations. In: IPMI (2021)
22. Jensen, M.H., Jørgensen, D.R., Jalaboi, R., Hansen, M.E., Olsen, M.A.: Improving uncertainty estimation in convolutional neural networks using inter-rater agreement. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 540–548. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_59
23. Ji, W., et al.: Learning calibrated medical image segmentation via multi-rater agreement modeling. In: CVPR (2021)
24. Kirillov, A., et al.: Segment anything. In: ICCV (2023)
25. Lemay, A., Gros, C., Karthik, E.N., Cohen-Adad, J.: Label fusion and training methods for reliable representation of inter-rater uncertainty. MELBA **031**, 1–29 (2023)
26. Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., Li, J.: Dice loss for data-imbalanced NLP tasks. In: ACL (2020)
27. Maier-Hein, L., et al.: Metrics reloaded: recommendations for image analysis validation. arXiv (2023)
28. Menon, A.K., Rawat, A.S., Reddi, S.J., Kim, S., Kumar, S.: A Statistical Perspective on Distillation. In: ICML (2021)
29. Menze, B., Joskowicz, L., Bakas, S., Jakab, A., Konukoglu, E., Becker, A.: Quantification of uncertainties in biomedical image quantification challenge. In: MICCAI (2020). https://qubiq.grand-challenge.org
30. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 3DV (2016)
31. Müller, R., Kornblith, S., Hinton, G.: When does label smoothing help? In: NeurIPS (2019)

32. Nordström, M., Hult, H., Maki, A., Löfman, F.: Noisy image segmentation with soft-dice. arXiv (2023)
33. Nowozin, S.: Optimal decisions from probabilistic models: the intersection-over-union case. In: CVPR (2014)
34. Popordanoska, T., Bertels, J., Vandermeulen, D., Maes, F., Blaschko, M.B.: On the relationship between calibrated predictors and unbiased volume estimation. In: MICCAI (2021)
35. Popordanoska, T., Sayer, R., Blaschko, M.B.: A consistent and differentiable Lp canonical calibration error estimator. In: NeurIPS (2022)
36. Qin, D., et al.: Efficient medical image segmentation based on knowledge distillation. TMI (2021)
37. Rahman, M.A., Wang, Y.: Optimizing intersection-over-union in deep neural networks for image segmentation. In: Bebis, G., et al. (eds.) ISVC 2016. LNCS, vol. 10072, pp. 234–244. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-50835-1_22
38. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: MICCAI (2015)
39. Salehi, S.S.M., Erdogmus, D., Gholipour, A.: Tversky loss function for image segmentation using 3D fully convolutional deep networks. In: MICCAI Workshop (2017)
40. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: inverted residuals and linear bottlenecks. In: CVPR (2018)
41. Silva, J.L., Oliveira, A.L.: Using soft labels to model uncertainty in medical image segmentation. In: MICCAI Workshop (2021)
42. 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: MICCAI Workshop (2017)
43. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR (2016)
44. Tan, M., Le, Q.V.: EfficientNet: rethinking model scaling for convolutional neural networks. In: ICML (2019)
45. Tilborghs, S., Bertels, J., Robben, D., Vandermeulen, D., Maes, F.: The dice loss in the context of missing or empty labels: introducing Φ and ϵ. In: MICCAI (2022)
46. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, New York (1995). https://doi.org/10.1007/978-1-4757-3264-1
47. Wang, Z., Blaschko, M.B.: Jaccard metric losses: optimizing the Jaccard index with soft labels. arXiv (2023)
48. Wightman, R.: Pytorch image models (2019). https://github.com/rwightman/pytorch-image-models
49. Yu, J., Blaschko, M.B.: The Lovász hinge: a novel convex surrogate for submodular losses. TPAMI **42**, 735–748 (2018)
50. Yu, J., et al.: Learning generalized intersection over union for dense pixelwise prediction. In: ICML (2021)
51. Zhang, D., et al.: Deep learning for medical image segmentation: tricks, challenges and future directions. arXiv (2022)