Calibrating Segmentation Networks with Margin-based Label Smoothing

Balamurali Murugesan[a,b,*], Bingyuan Liu[a,b], Adrian Galdran[c], Ismail Ben Ayed[a,b,d], Jose Dolz[a,b,d]

[a]LIVIA, ÃL'TS MontrÃl'al, Canada
[b]International Laboratory on Learning Systems (ILLS),
McGill - ETS - MILA - CNRS - UniversitÃl' Paris-Saclay - CentraleSupÃl'lec, Canada
[c]Universitat Pompeu Fabra, Barcelona, Spain
[d]Centre de Recherche du Centre Hospitalier de l'UniversitÃl' de MontrÃl'al (CRCHUM), Canada

ABSTRACT

Despite the undeniable progress in visual recognition tasks fueled by deep neural networks, there exists recent evidence showing that these models are poorly calibrated, resulting in over-confident predictions. The standard practices of minimizing the cross-entropy loss during training promote the predicted softmax probabilities to match the one-hot label assignments. Nevertheless, this yields a pre-softmax activation of the correct class that is significantly larger than the remaining activations, which exacerbates the miscalibration problem. Recent observations from the classification literature suggest that loss functions that embed implicit or explicit maximization of the entropy of predictions yield state-of-the-art calibration performances. Despite these findings, the impact of these losses in the relevant task of calibrating medical image segmentation networks remains unexplored. In this work, we provide a unifying constrained-optimization perspective of current state-of-the-art calibration losses. Specifically, these losses could be viewed as approximations of a linear penalty (or a Lagrangian term) imposing equality constraints on logit distances. This points to an important limitation of such underlying equality constraints, whose ensuing gradients constantly push towards a non-informative solution, which might prevent from reaching the best compromise between the discriminative performance and calibration of the model during gradient-based optimization. Following our observations, we propose a simple and flexible generalization based on inequality constraints, which imposes a controllable margin on logit distances. Comprehensive experiments on a variety of public medical image segmentation benchmarks demonstrate that our method sets novel state-of-the-art results on these tasks in terms of network calibration, whereas the discriminative performance is also improved. The code is available at
https://github.com/Bala93/MarginLoss

1. Introduction

Deep neural networks (DNNs) are driving progress in a variety of computer vision tasks across different domains and applications. In particular, these high-capacity models have become the *de-facto* solution in critical tasks, such as medical image segmentation. Despite their superior performance, there exists recent evidence (Guo et al., 2017; Mukhoti et al., 2020; Müller et al., 2019) which demonstrates that these models are poorly calibrated, often resulting in over-confident predictions. As

*Corresponding author: balamurali.murugesan.1@ens.etsmtl.ca

a result, the predicted probability values associated with each class overestimate the actual likelihood of correctness.

Quantifying the predictive uncertainty of modern DNNs has gained popularity recently, with several alternatives to train better calibrated models. A simple yet effective approach consists in integrating a post-processing step that modifies the predicted probabilities of a trained neural network (Guo et al., 2017; Zhang et al., 2020; Tomani et al., 2021; Ding et al., 2021). This strategy, however, presents several limitations. First, the choice of the transformation parameters, such as temperature scaling, is highly dependent on the dataset and network. And second, under domain drift, post-hoc calibration performance largely degrades (Ovadia et al., 2019), resulting in unreliable predictions. A more principled alternative is to explicitly maximize the Shannon entropy of the model predictions during training, which can be achieved by augmenting the learning objective with a term that penalizes confident output distributions (Pereyra et al., 2017). Furthermore, recent efforts to quantify the quality of predictive uncertainties have focused on investigating the effect of the entropy on the training labels (Xie et al., 2016; Müller et al., 2019; Mukhoti et al., 2020). Findings from these works evidence that, popular losses, which modify the hard-label assignments, such as label smoothing (Szegedy et al., 2016) and focal loss (Lin et al., 2017), implicitly integrate an entropy maximization objective and have a favourable effect on model calibration. As shown comprehensively in the recent study in (Mukhoti et al., 2020), these losses, with implicit or explicit maximization of the entropy, represent the state-of-the-art in model calibration in visual and non-visual recognition tasks.

Despite this progress, the benefit of these calibration losses remains unclear in medical image segmentation. Indeed, only a handful of works have addressed this important problem, mostly focusing on the calibration assessment of standard segmentation losses (Mehrtash et al., 2020), i.e., cross-entropy and Dice. Thus, we believe that it is of great significance and interest to study methods for confidence calibration of segmentation models in the context of medical image segmentation.

The contributions of this work are summarized as follows:

- We provide a unifying constrained-optimization perspective of current state-of-the-art calibration losses. Specifically, these losses could be viewed as approximations of a linear penalty (or a Lagrangian term) imposing equality constraints on logit distances. This points to an important limitation of such underlying hard equality constraints, whose ensuing gradients constantly push towards a non-informative solution, which might prevent from reaching the best compromise between the discriminative performance and calibration of the model during gradient-based optimization.

- Following our observations, we propose a simple and flexible generalization based on inequality constraints, which imposes a controllable margin on logit distances.

- We provide comprehensive experiments and ablation studies on five different public segmentation benchmarks

that focus on diverse targets and modalities, highlighting the generalization capabilities of the proposed approach. Our empirical results demonstrate the superiority of our method compared to state-of-the-art calibration losses in both calibration and discriminative performance.

This journal version provides a substantial extension of the conference work presented in (Liu et al., 2022). In particular, we provide a thorough literature review on calibration of segmentation models, with a main focus on the medical field. Second, we perform a comprehensive empirical validation, including *i)* multiple public benchmarks covering diverse modalities and targets, *ii)* adding recent approaches which specifically target calibration of segmentation models (i.e., (Islam and Glocker, 2021) and (Ding et al., 2021)), and *iii)* substantial in-depth analysis of the behaviour of the analyzed models. We believe that, to date, this work represents the most comprehensive evaluation of calibration models in the task of medical image segmentation, not only in terms of the amount of benchmarks employed, but also in regards of models compared.

2. Related work

Post-processing approaches. Including a post-processing step that transforms the probability predictions of a deep network (Guo et al., 2017; Zhang et al., 2020; Tomani et al., 2021; Ding et al., 2021) is a straightforward yet efficient strategy to mitigate miscalibrated predictions. Among these methods, *temperature scaling* (Guo et al., 2017), a variant of Platt scaling (Platt et al., 1999), employs a single scalar parameter over all the pre-softmax activations, which results in softened class predictions. Despite its good performance on in-domain samples, (Ovadia et al., 2019) demonstrated that temperature scaling does not work well under data distributional shift. (Tomani et al., 2021) mitigated this limitation by transforming the validation set before performing the post-hoc calibration step, whereas (Ma and Blaschko, 2021) introduced a ranking model to improve the post-processing model calibration.

Probabilistic and non-probabilistic approaches have been also investigated to measure the uncertainty of the predictions in modern deep neural networks. For example, prior literature has employed Bayesian neural networks to approximate inference by learning a posterior distribution over the network parameters, as obtaining the exact Bayesian inference is computationally intractable in deep networks. These Bayesian-based models include variational inference (Blundell et al., 2015; Louizos and Welling, 2016), stochastic expectation propagation (Hernández-Lobato and Adams, 2015) or dropout variational inference (Gal and Ghahramani, 2016). A popular non-parametric alternative is ensemble learning, where the empirical variance of the network predictions is used as an approximate measure of uncertainty. This yields improved discriminative performance, as well as meaningful predictive uncertainty with reduced miscalibration. Common strategies to generate ensembles include differences in model hyperparameters (Wenzel et al., 2020), random initialization of the network parameters and random shuffling of the data points (Lakshmi-

narayanan et al., 2017), Monte-Carlo Dropout (Gal and Ghahramani, 2016; Zhang et al., 2019), dataset shift (Ovadia et al., 2019) or model orthogonality constraints (Larrazabal et al., 2021). However, a main drawback of this strategy stems from its high computational cost, particularly for complex models and large datasets.

Explicit and implicit penalties. Modern classification networks trained under the fully supervised learning paradigm resort to training labels provided as binary one-hot encoded vectors. Therefore, all the probability mass is assigned to a single class, resulting in minimum-entropy supervisory signals (i.e., entropy equal to zero). As the network is trained to follow this distribution, we are implicitly forcing it to be overconfident (i.e., to achieve a minimum entropy), thereby penalizing uncertainty in the predictions. While temperature scaling artificially increases the entropy of the predictions, (Pereyra et al., 2017) included into the learning objective a term to penalize confident output distributions by explicitly maximizing the entropy. In contrast to tackling overconfidence directly on the predicted probability distributions, recent works have investigated the effect of the entropy on the training labels. The authors of (Xie et al., 2016) explored adding label noise as a regularization, where the disturbed label vector was generated by following a generalized Bernoulli distribution. Label smoothing (Szegedy et al., 2016), which successfully improves the accuracy of deep learning models, has been shown to implicitly calibrate the learned models, as it prevents the network from assigning the full probability mass to a single class, while maintaining a reasonable distance between the logits of the ground-truth class and the other classes (Müller et al., 2019). More recently, (Mukhoti et al., 2020) demonstrated that focal loss (Lin et al., 2017) implicitly minimizes a Kullback-Leibler (KL) divergence between the uniform distribution and the softmax predictions, thereby increasing the entropy of the predictions. Indeed, as shown in (Müller et al., 2019; Mukhoti et al., 2020), both label smoothing and focal loss implicitly regularize the network output probabilities, encouraging their distribution to be close to the uniform distribution. To our knowledge, and as demonstrated experimentally in the recent studies in (Müller et al., 2019; Mukhoti et al., 2020), loss functions that embed implicit or explicit maximization of the entropy of the predictions yield state-of-the-art calibration performances.

Calibration in medical image segmentation. Recent literature has focused on either estimating the predictive uncertainty or on leveraging this uncertainty to improve the discriminative performance of segmentation models (Wang et al., 2019). Nevertheless, research to improve both the calibration and segmentation performance of CNN-based segmentation models is scarce. (Jena and Awate, 2019) proposed a novel deep segmentation framework rooted in generative modeling and Bayesian decision theory, which allowed to define a principled measure of uncertainty associated with label probabilities. Recent findings (Fort et al., 2019), however, suggest that current state-of-the-art Bayesian neural networks have tendency to find solutions around a single minimum of the loss landscape and, consequently, lack diversity. In contrast, ensembling deep neural networks typically results in more diverse predictions, and therefore obtain better uncertainty estimates. This observation aligns with the recent work in (Jungo et al., 2020; Mehrtash et al., 2020), which evaluates several uncertainty estimation approaches and concludes that ensembling outperforms other methods. To promote model diversity within the ensemble, (Larrazabal et al., 2021) integrate an orthogonality constraint in the learning objective, showing significant gains over the non-constrained set. More recently, (Karimi and Gholipour, 2022) argue that training a single model in a multi-task manner on several different datasets yields better calibration on the different tasks. Nevertheless, these methods incur in high computationally expensive steps as they involve training either multiple models or a single model on multiple datasets. In an orthogonal direction, several recent methods have overcome this limitation and proposed lighter alternatives. For example, (Ding et al., 2021) extends the naive temperature scaling by integrating a simple CNN to predict the pixel-wise temperature values in a post-processing step. In addition, (Islam and Glocker, 2021) apply a weight matrix with a Gaussian kernel across the one-hot encoded expert labels to obtain soft class probabilities, adding into the standard Label smoothing a spatial-awareness. However, despite these initial efforts, and to the best of our knowledge, a comprehensive evaluation of calibration methods in multiple medical image segmentation benchmarks has not been conducted yet.

3. Preliminaries

Let $\mathcal{D}(\mathcal{X}, \mathcal{Y}) = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^N$ be the training dataset, with $\mathbf{x}^{(i)} \in \mathcal{X} \subset \mathbb{R}^{\Omega_i}$ representing the ith image, Ω_i the spatial image domain, and $\mathbf{y} \in \mathcal{Y} \subset \mathbb{R}^K$ its corresponding ground-truth label with K classes, provided as one-hot encoding. Given an input image $\mathbf{x}^{(i)}$, a neural network parameterized by θ generates a logit vector, defined as $f_\theta(\mathbf{x}^{(i)}) = \mathbf{l}^{(i)} \in \mathbb{R}^K$. To simplify the notations, we omit sample indices, as this does not lead to ambiguity, and just use $\mathbf{l} = (l_k)_{1 \le k \le K} \in \mathbb{R}^K$ to denote logit vectors. Note that the logits are the inputs of the softmax probability predictions of the network, which are computed as:

$$\mathbf{s} = (s_k)_{1 \le k \le K} \in \mathbb{R}^K; \quad s_k = \frac{\exp^{l_k}}{\sum_j^K \exp^{l_j}}$$

The predicted class is computed as $\hat{y} = \arg\max_k s_k$, whereas the predicted confidence is given by $\hat{p} = \max_k s_k$.

Calibrated models. *Perfectly calibrated* models are those for which the predicted confidence for each sample is equal to the model accuracy : $\hat{p} = \mathbb{P}(\hat{y} = y | \hat{p})$, where y denotes the true labels. Therefore, an *over-confident model* tends to yield predicted confidences that are larger than its accuracy, whereas an *under-confident model* displays lower confidence than the model's accuracy.

Miscalibration of DNNs. To train fully supervised discriminative deep models, the standard cross-entropy (CE) loss is commonly used as the training objective. We argue that, from a calibration performance, the supervision of CE is suboptimal.

Indeed, CE reaches its minimum when the predictions for all the training samples match the hard (binary) ground-truth labels, i.e., $s_k = 1$ when k is the ground-truth class of the sample and $s_k = 0$ otherwise. Minimizing the CE implicitly pushes softmax vectors $\mathbf{s}$ towards the vertices of the simplex, thereby magnifying the distances between the largest logit $\max_k(l_k)$ and the rest of the logits, yielding over-confident predictions and miscalibrated models.

4. A constrained-optimization perspective of calibration

We present in this section a novel constrained-optimization perspective of current calibration methods for deep networks, showing that the existing strategies, including Label Smoothing (LS) (Müller et al., 2019; Szegedy et al., 2016), Focal Loss (FL) (Mukhoti et al., 2020; Lin et al., 2017) and Explicit Confidence Penalty (ECP) (Pereyra et al., 2017), impose *equality* constraints on logit distances. Specifically, they embed either explicit or implicit penalty functions, which push all the logit distances to zero.

4.1. Definition of logit distances

Let us first define the vector of logit distances between the winner class and the rest as:

$$\mathbf{d}(\mathbf{l}) = (\max_j(l_j) - l_k)_{1 \leq k \leq K} \in \mathbb{R}^K \qquad (1)$$

Note that each element in $\mathbf{d}(\mathbf{l})$ is non-negative. In the following, we show that LS, FL and ECP correspond to different *soft penalty* functions for imposing the same hard equality constraint $\mathbf{d}(\mathbf{l}) = \mathbf{0}$ or, equivalently, imposing inequality constraint $\mathbf{d}(\mathbf{l}) \leq \mathbf{0}$ (as $\mathbf{d}(\mathbf{l})$ is non-negative by definition). Clearly, enforcing this equality constraint in a hard manner would result in all K logits being equal for a given sample, which corresponds to non-informative softmax predictions $s_k = \frac{1}{K} \forall k$.

4.2. Penalty functions in constrained optimization

In the general context of constrained optimization (Bertsekas, 1995), *soft* penalty functions are widely used to tackle *hard* equality or inequality constraints. For the discussion in the sequel, consider specifically the following hard equality constraint:

$$\mathbf{d}(\mathbf{l}) = \mathbf{0} \qquad (2)$$

The general principle of a soft-penalty optimizer is to replace a hard constraint of the form in Eq. 2 by adding an additional term $\mathcal{P}(\mathbf{d}(\mathbf{l}))$ into the main objective function to be minimized. Soft penalty $\mathcal{P}$ should be a continuous and differentiable function, which reaches its global minimum when the constraint is satisfied, i.e., it verifies: $\mathcal{P}(\mathbf{d}(\mathbf{l})) \geq \mathcal{P}(\mathbf{0}) \forall \mathbf{l} \in \mathbb{R}^K$. Thus, when the constraint is violated, i.e., when $\mathbf{d}(\mathbf{l})$ deviates from $\mathbf{0}$, the penalty term $\mathcal{P}$ increases.

Label smoothing. Recent evidence (Lukasik et al., 2020; Müller et al., 2019) suggests that, in addition to improving the discriminative performance of deep neural networks, Label Smoothing (LS) (Szegedy et al., 2016) positively impacts model calibration. In particular, LS modifies the hard target labels with a smoothing parameter α, so that the original one-hot training labels $\mathbf{y} \in \{0,1\}^K$ become $\mathbf{y}^{LS} = (y_k^{LS})_{1 \leq k \leq K}$, with $y_k^{LS} = y_k(1-\alpha) + \frac{\alpha}{K}$. Then, we simply minimize the cross-entropy between the modified labels and the network outputs:

$$\mathcal{L}_{LS} = -\sum_k y_k^{LS} \log s_k = -\sum_k ((1-\alpha)y_k + \frac{\alpha}{K}) \log s_k \quad (3)$$

where $\alpha \in [0,1]$ is the smoothing hyper-parameter. It is straightforward to verify that cross-entropy with label smoothing in Eq. 3 can be decomposed into a standard cross-entropy term augmented with a Kullback-Leibler (KL) divergence between uniform distribution $\mathbf{u} = \frac{1}{K}$ and the softmax prediction:

$$\mathcal{L}_{LS} \overset{c}{=} \mathcal{L}_{CE} + \frac{\alpha}{1-\alpha} \mathcal{D}_{KL}(\mathbf{u}||\mathbf{s}) \qquad (4)$$

where $\overset{c}{=}$ stands for equality up to additive and/or non-negative multiplicative constants. Now, consider the following bounding relationships between a linear penalty (or a Lagrangian) for equality constraint $\mathbf{d}(\mathbf{l}) = \mathbf{0}$ and the KL divergence in Eq. 4.

Proposition 1. *A linear penalty (or a Lagrangian term) for constraint $\mathbf{d}(\mathbf{l}) = \mathbf{0}$ is bounded from above and below by $\mathcal{D}_{KL}(\mathbf{u}||\mathbf{s})$, up to additive constants:*

$$\mathcal{D}_{KL}(\mathbf{u}||\mathbf{s}) - \log(K) \overset{c}{\leq} \frac{1}{K}\sum_k (\max_j(l_j) - l_k) \overset{c}{\leq} \mathcal{D}_{KL}(\mathbf{u}||\mathbf{s})$$

where $\overset{c}{\leq}$ stands for inequality up to an additive constant.

These bounding relationships could be obtained directly from the softmax and $\mathcal{D}_{KL}$ expressions, along with the following well-known property of the LogSumExp function: $\max_k(l_k) \leq \log \sum_k^K e^{l_k} \leq \max_k(l_k) + \log(K)$. For the details of the proof, please refer to Appendix A of the conference version in (Liu et al., 2022).

Prop. 1 means that LS is (approximately) optimizing a linear penalty (or a Lagrangian) for logit-distance constraint $\mathbf{d}(\mathbf{l}) = \mathbf{0}$, which encourages equality of all logits; see the illustration in Figure 1, top-left.

Focal loss. Another popular alternative for calibration is focal loss (FL) (Lin et al., 2017), which attempts to alleviate the over-fitting issue in CE by directing the training attention towards samples with low confidence in each mini-batch. More concretely, the authors proposed to use a modulating factor to the CE, $(1 - s_k)^\gamma$, which controls the trade-off between easy and hard examples. Very recently, (Mukhoti et al., 2020) demonstrated that focal loss is, in fact, an upper bound on CE augmented with a term that implicitly serves as a maximum-entropy regularizer:

$$\mathcal{L}_{FL} = -\sum_k (1-s_k)^\gamma y_k \log s_k \geq \mathcal{L}_{CE} - \gamma \mathcal{H}(\mathbf{s}) \qquad (5)$$

where γ is a hyper-parameter and $\mathcal{H}$ denotes the Shannon entropy of the softmax prediction, given by

$$\mathcal{H}(\mathbf{s}) = -\sum_k s_k \log(s_k)$$

In this connection, FL is closely related to ECP (Pereyra et al., 2017), which explicitly added the negative entropy term, $-\mathcal{H}(\mathbf{s})$, to the training objective. It is worth noting that minimizing the negative entropy of the prediction is equivalent to minimizing the KL divergence between the prediction and the uniform distribution, up to an additive constant, i.e.,

$$-\mathcal{H}(\mathbf{s}) \stackrel{c}{=} \mathcal{D}_{\mathrm{KL}}(\mathbf{s}||\mathbf{u})$$

which is a reversed form of the KL term in Eq. 4.

Therefore, all in all, and following Prop. 1 and the discussions above, LS, FL and ECP could be viewed as different penalty functions for imposing the same logit-distance equality constraint $\mathbf{d}(\mathbf{l}) = \mathbf{0}$. This motivates our margin-based generalization of logit-distance constraints, which we introduce in the following section, along with discussions of its desirable properties (e.g., gradient dynamics) for calibrating neural networks.
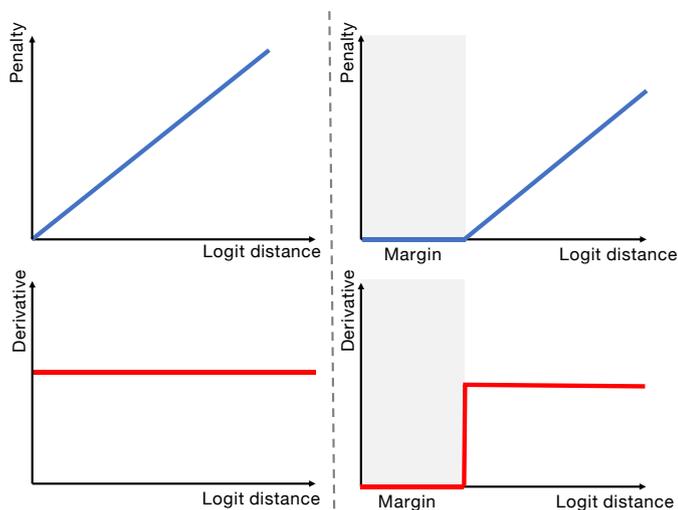

Figure 1: Illustration of the linear (left) and margin-based (right) penalties for imposing logit-distance constraints, along with the corresponding derivatives. Note that while the derivative of the linear penalty for constraint $\mathbf{d}(\mathbf{l}) = \mathbf{0}$ constantly pushes towards the trivial solution $s_k = \frac{1}{K} \forall K$ (i.e., LS, FL and EPC), the derivative of the proposed model only pushes towards zero those logits above the given margin.

4.3. Margin-based Label Smoothing (MbLS)

Our previous analysis shows that LS, FL and ECP are closely related from a constrained-optimization perspective, and they could be seen as approximations of a linear penalty for imposing constraint $\mathbf{d}(\mathbf{l}) = \mathbf{0}$, pushing all logit distances to zero; see Figure 1, top-left. Clearly, enforcing this constraint in a hard way yields a non-informative solution where all the classes have exactly the same logit and, hence, the same class prediction: $s_k = \frac{1}{K} \forall K$. While this trivial solution is not reached in practice when using soft penalties (as in LS, FL and ECP) jointly with CE, we argue that the underlying equality constraint $\mathbf{d}(\mathbf{l}) = \mathbf{0}$ has an important limitation, which might prevent from reaching the best compromise between the discriminative performance and calibration of the model during gradient-based optimization. Figure 1, left, illustrates this: With the linear penalty for constraint $\mathbf{d}(\mathbf{l}) = \mathbf{0}$ in the top-left of the Figure,

the derivative with respect to logit distances is a strictly positive constant (left-bottom), yielding during training *a gradient term that constantly pushes towards the trivial, non-informative solution* $\mathbf{d}(\mathbf{l}) = \mathbf{0}$ (or equivalently $s_k = \frac{1}{K} \forall K$). To alleviate this issue, we propose to replace the equality constraint $\mathbf{d}(\mathbf{l}) = \mathbf{0}$ with the more general inequality constraint $\mathbf{d}(\mathbf{l}) \leq \mathbf{m}$, where $\mathbf{m}$ denotes the K-dimensional vector with all elements equal to $m > 0$. Therefore, we include a margin m into the penalty, so that the logit distances in $\mathbf{d}(\mathbf{l})$ are allowed to be below m when optimizing the main learning objective:

$$\min \quad \mathcal{L}_{\mathrm{CE}} \quad \text{s.t.} \quad \mathbf{d}(\mathbf{l}) \leq \mathbf{m}, \quad \mathbf{m} > \mathbf{0} \qquad (6)$$

The intuition behind adding a strictly positive margin m is that, unlike the linear penalty for constraint $\mathbf{d}(\mathbf{l}) = \mathbf{0}$ (Figure 1, left), the gradient is back-propagated only on those logits where the distance is above the margin (Figure 1, right). This contrasts with the linear penalty, for which there exists always a gradient, and its value is the same across all the logits, regardless of their distance.

Even though the constrained problem in Eq. 6 could be solved by a Lagrangian-multiplier algorithm, we resort to a simpler unconstrained approximation by ReLU function:

$$\min \quad \mathcal{L}_{\mathrm{CE}} + \lambda \sum_k \max(0, \max_j(l_j) - l_k - m) \qquad (7)$$

Here, the non-linear ReLU penalty for inequality constraint $\mathbf{d}(\mathbf{l}) \leq \mathbf{m}$ discourages logit distances from surpassing a given margin m, and λ is a trade-off weight balancing the two terms. It is clear that, as discussed in Sec. 4, several competitive calibration methods could be viewed as approximations for imposing constraint $\mathbf{d}(\mathbf{l}) = \mathbf{0}$ and, therefore, correspond to the special case of our method when setting the margin to $m = 0$. Our comprehensive experiments in the next section demonstrate clearly the benefits of introducing a strictly positive margin m.

Note that our model in Eq. 7 has two hyper-parameters, m and λ. We fixed λ to 0.1 in our experiments for all the benchmarks, and tuned only the margin m over validation sets. In this way, when comparing with the existing calibration solutions, we use the same budget of hyper-parameter optimization (m in our method vs. α in LS or γ in FL).

5. Experiments

5.1. Experimental Setting

5.1.1. Datasets

To empirically validate our model, we employ five public multi-class segmentations benchmarks, whose detailes are specified below.

Automated Cardiac Diagnosis Challenge (ACDC) *(Bernard et al., 2018).* This dataset consists of 100 patient exams containing cardiac MR volumes and its respective multi-class segmentation masks for both diastolic and systolic phases. The

segmentation mask contains four classes, including the left ventricle (LV), right ventricle (RV), myocardium (Myo) and background. Following the standard practices on this dataset, 2D slices are extracted from the available volumes and resized to 224×224. Last, the dataset is randomly split into independent training (70), validation (10) and testing (20) sets.

Brain Tumor Segmentation (BRATS) Challenge. (Menze et al., 2015; Bakas et al., 2017, 2018) The dataset contains 484 multi-modal MR scans (FLAIR, T1, T1-contrast, and T2) with their corresponding Glioma segmentation masks. The classes representing the mask include tumor core (TC), enhancing tumor (ET) and whole tumor (WT). Each volume of dimension 155×240×240 is resampled and slices containing only background are removed from the training. The patient volumes are randomly split to 327, 54, 94 for training, validation, and testing respectively.

MRBrainS18 *(Mendrik et al., 2015a)*. The dataset contains paired T1, T2, and T1-IR volumes of 7 subjects and their segmentation masks, which correspond to brain tissue including Gray Matter (GM), White Matter (WM), and Cerebralspinal fluid (CSF). The dimensions of the volumes are 240×240×48. We utilize 5 subjects for training and 2 subjects for testing.

Fast and Low GPU memory Abdominal oRgan sEgmentation (FLARE) Challenge (Ma et al., 2021). The dataset contains 360 volumes of multi-organ abdomen CT including liver, kidneys, spleen and pancreas and their corresponding pixel-wise masks. The different resolutions are resampled to a common space and cropped to 192×192×30. The volumes are then randomly split to 240 for training, 40 for validation, 80 for testing.

PROMISE. (Litjens et al., 2014) The dataset was made available at the MICCAI 2012 prostate MR segmentation challenge. It contains the transversal T2-weighted MR images acquired at different centers with multiple MRI vendors and different scanning protocols. It is comprised of various diseases, i.e., benign and prostate cancers. The images resolution ranges from 15×256×256 to 54×512×512 voxels with a spacing ranging from 2×0.27×0.27 to 4×0.75×0.75mm³. We employed 22 patients for training, 3 for validation and 7 for testing.

Note that in all datasets, images are normalized to be within the range [0-1]. Furthermore, for the datasets containing multiple image modalities (i.e., MRBrainS and BRATS), all available modalities are concatenated in a single tensor, which is fed to the input of the neural network. In addition, there exists one dataset for which the low amount of available images impeded us to generate a proper training, validation and testing split (MRBrainS). In this case, we repeat the training-testing procedure several times randomly selecting the train-test data and report the mean values.

5.1.2. Evaluation Metrics

To assess the discriminative performance of the evaluated models, we resort to standard segmentation metrics in the medical segmentation literature, which includes the DICE coefficient (DSC) and the Average Surface Distance (ASD). To evaluate the calibration performance, we employ both the expected calibration error (ECE) (Naeini et al., 2015) and classwise expected calibration error (CECE). The reason to include CECE is because ECE only considers the softmax probability of the predicted class, ignoring the other scores in the softmax distribution (Mukhoti et al., 2020). To compute the ECE given a finite number of samples, we group predictions into M equispaced bins. Let B_i denote the set of samples with confidences belonging to the ith bin. The accuracy A_i of this bin is computed as $A_i = \frac{1}{|B_i|} \sum_{j \in B_i} 1(\hat{y}_j = y_j)$, where 1 is the indicator function, and $\hat{y}_j$ and y_j are the predicted and ground-truth labels for the jth sample. Similarly, the confidence C_i of the ith bin is computed as $C_i = \frac{1}{|B_i|} \sum_{j \in B_i} \hat{p}_j$, i.e. C_i is the average confidence of all samples in the bin. The ECE can be approximated as a weighted average of the absolute difference between the accuracy and confidence of each bin:

$$ECE = \sum_{i=1}^{M} \frac{|B_i|}{N} |A_i - C_i| \tag{8}$$

The ECE metric only considers the probability of the predicted class, without considering the other scores in the softmax distribution. A stronger definition of calibration would require the probabilities of all the classes in the softmax distribution to be calibrated. This can be achieved with a simple classwise extension of the ECE metric: Classwise ECE, given by

$$CECE = \sum_{i=1}^{M} \sum_{j=1}^{K} \frac{|B_{i,j}|}{N} |A_{i,j} - C_{i,j}| \tag{9}$$

where K is the number of classes, B_{ij} denotes the set of samples from the jth class in the ith bin, $A_i = \frac{1}{|B_{i,j}|} \sum_{k \in B_{i,j}} 1(j = y_k)$ and $C_{i,j} = \frac{1}{|B_{ij}|} \sum_{k \in B_{i,j}} \hat{p}_{kj})$

Following the recent literature on calibration of segmentation networks (Islam and Glocker, 2021) both ECE and CECE are obtained by considering only the foreground regions. The reason behind this is that most of the correct –and certain– predictions are from the background. If we exclude these areas from the statistics, the obtained results will better highlight the differences among the different approaches. In our implementation, the number of bins to compute ECE and CECE is set to $M = 15$. Furthermore, we also employ reliability plots (Niculescu-Mizil and Caruana, 2005) in our evaluation, which plot the expected accuracy as a function of class probability (confidence), and for a perfectly calibrated model it represents the identity function.

5.1.3. Implementation Details

To empirically evaluate the proposed model, we conduct experiments comparing a state-of-the-art segmentation network on a multi-class scenario trained with different learning objectives. In particular, we first include standard loss functions employed in medical image segmentation, which include the common Cross-entropy (CE) and the duple composed by CE and

Table 1: The discriminative performance (DSC and ASD) obtained by the different models across five popular medical image segmentation benchmarks. Best method is highlighted in bold, whereas second best approach is underlined.

Dataset	Region	CE		CE + DICE		FL		ECP		LS		SVLS		Ours	
		DSC	ASD	DSC	ASD	DSC	ASD	DSC	ASD	DSC	ASD	DSC	ASD	DSC	ASD
ACDC	RV	0.813	0.77	0.798	0.75	0.714	1.27	0.754	0.87	_0.815_	_0.68_	0.642	1.86	**0.866**	**0.42**
	MYO	_0.816_	0.43	0.795	0.46	0.734	0.61	0.751	0.53	0.805	_0.42_	0.664	1.79	**0.845**	**0.37**
	LV	_0.894_	0.36	0.888	0.35	0.846	0.47	0.839	0.41	0.886	_0.32_	0.795	1.21	**0.913**	**0.29**
	Mean	_0.841_	0.52	0.827	0.52	0.764	0.78	0.782	0.60	0.835	_0.48_	0.701	1.62	**0.875**	**0.36**
MRBrainS	GM	0.780	_0.42_	0.757	0.48	0.773	0.53	_0.793_	0.47	0.745	0.51	0.753	0.49	**0.800**	**0.39**
	WM	_0.811_	0.62	0.761	0.66	0.804	0.60	0.810	_0.55_	0.727	0.97	0.670	1.06	**0.831**	**0.46**
	CSF	0.772	0.44	0.780	0.46	0.793	0.40	0.803	_0.39_	0.772	0.46	**0.810**	0.39	_0.807_	**0.38**
	Mean	0.788	0.50	0.766	0.54	0.790	0.51	_0.802_	_0.47_	0.748	0.64	0.744	0.65	**0.813**	**0.41**
FLARE	Liver	0.949	0.60	0.942	_0.43_	0.951	**0.37**	_0.952_	0.56	0.952	1.44	0.949	1.47	**0.953**	1.52
	Kidney	0.944	0.37	0.941	0.37	0.946	_0.32_	**0.950**	**0.31**	_0.947_	0.38	0.946	0.40	0.945	0.35
	Spleen	0.929	0.56	0.904	0.61	0.924	0.55	0.924	0.68	**0.942**	_0.38_	0.932	0.56	_0.940_	**0.38**
	Pancreas	0.635	1.55	0.634	**1.41**	0.625	1.65	**0.649**	1.47	0.636	1.56	0.636	1.53	_0.645_	_1.42_
	Mean	0.864	0.77	0.855	_0.71_	0.862	_0.72_	_0.869_	0.75	0.869	0.94	0.866	0.99	**0.871**	0.92
BRATS	TC	0.754	**1.01**	**0.768**	1.13	0.761	1.17	0.756	1.14	0.737	1.26	0.751	1.29	_0.763_	1.14
	ET	0.498	_1.41_	**0.524**	**1.40**	0.499	1.72	0.498	1.64	0.477	1.89	0.490	1.92	_0.505_	1.66
	WT	0.839	1.08	0.850	_0.98_	0.852	1.00	**0.860**	**0.98**	0.856	1.05	0.856	1.09	0.857	1.08
	Mean	0.697	_1.17_	**0.714**	**1.17**	0.704	1.30	0.705	1.25	0.690	1.40	0.699	1.43	_0.708_	1.29
PROMISE	Prostate	0.737	1.33	0.751	_1.17_	0.729	1.42	0.736	1.27	0.713	1.72	_0.766_	1.27	**0.770**	**0.95**
	Tumor	0.258	5.81	0.328	4.10	0.361	3.35	0.344	2.48	0.350	3.29	_0.396_	**2.16**	**0.397**	_2.34_
	Mean	0.498	3.57	0.540	2.63	0.545	2.39	0.540	1.88	0.532	2.50	_0.581_	_1.71_	**0.583**	**1.64**

Table 2: The calibration performance (ECE and CECE) obtained by the different models across five popular medical image segmentation benchmarks. Best method is highlighted in bold, whereas second best approach is underlined. ∇ indicates the difference between the best model and our approach.

Dataset	CE		CE + DICE		FL		ECP		LS		SVLS		Ours		∇ECE	∇CECE
	ECE	CECE	ECE	CECE	ECE	CECE	ECE	CECE	ECE	CECE	ECE	CECE	ECE	CECE		
ACDC	_0.079_	_0.073_	0.137	0.084	0.113	0.116	0.109	0.095	0.081	0.107	0.176	0.135	**0.061**	**0.069**	–	–
MRBrainS	0.089	0.070	0.172	0.102	**0.020**	_0.064_	0.048	0.068	_0.036_	0.085	0.060	0.080	0.050	**0.058**	0.030	–
FLARE	0.045	0.029	0.058	0.034	**0.033**	0.035	_0.037_	**0.027**	0.055	0.050	0.039	0.036	0.038	_0.028_	0.005	0.001
BRATS	0.128	_0.108_	0.206	0.116	0.139	0.139	**0.107**	**0.099**	0.125	0.146	0.128	0.109	_0.116_	0.109	0.009	0.010
PROMISE	0.411	0.334	0.430	0.304	_0.247_	0.298	0.306	_0.252_	0.280	0.299	0.344	0.271	**0.232**	**0.237**	–	–

DSC losses. Furthermore, we also include training objectives which have been proposed to calibrate neural networks, which represent nowadays the state-of-the-art for this task. This includes Focal loss (FL) (Lin et al., 2017), Label Smoothing (LS) (Szegedy et al., 2016) and ECP (Pereyra et al., 2017). Last, we also compare to the recent Spatially-Varying LS (SVLS) (Islam and Glocker, 2021), which demonstrated to outperform the simpler LS version in the task of medical image segmentation. Following the literature, we have chosen the commonly used hyper-parameters and considered the one which provided the best compromise between DSC and ECE. For FL, γ values of 1, 2, and 3 are considered. In case of ECP and LS, α and λ values of 0.1, 0.2, 0.3 are used. For our method, we considered the margins to be 5, 8, and 10. In the case of SVLS, the one-hot label smoothing is performed with a kernel size of 3. For the experiments, we fixed the batch size to 4, epochs to 100, and optimizer to ADAM. The learning rate of 1e-3 and 1e-4 are used for the first 50 epochs, and the next 50 epochs respectively.

Backbones. The main experiments are conducted on the popular UNet (Ronneberger et al., 2015). Nevertheless, to show the versatility of the proposed margin based label smoothing, we have evaluated our model on other popular architectures in medical image segmentation including AttUNet (Oktay et al., 2018), TransUNet (Chen et al., 2021), and UNet++ (Zhou et al., 2020).

5.2. Results

5.2.1. Main results

The discriminative quantitative results obtained by the proposed model, as well as prior literature, are reported in Table 1. We observe that across the different datasets, our model consistently achieves the best discriminative performance, typically ranking as the best or second-best model in both region-based (i.e., DSC) and distance-based (i.e., ASD) metrics. This demonstrates that our method yields not only a better identification of target regions, but also an improvement in the boundary regions, highlighted by lower ASD values. An interesting observation is that, while other learning objectives typically result in performance gains compared to the standard CE loss, their superiority over the others depends on the selected dataset.

Table 2 summarizes the calibration performance, in terms of ECE and CECE of all the analyzed models. We can observe that, similar to the discriminative performance reported earlier, the proposed model typically ranks as best or second best method. An interesting observation is that, according to the results, focal loss provides well-calibrated models, whereas their discriminative performance is typically far from best performing models. As exposed in our motivation, one of the reasons behind this behaviour might be the undesirable effect of pushing all logit distances to zero. Enforcing this constraint may alleviate the problem of overconfidence in deep networks, at
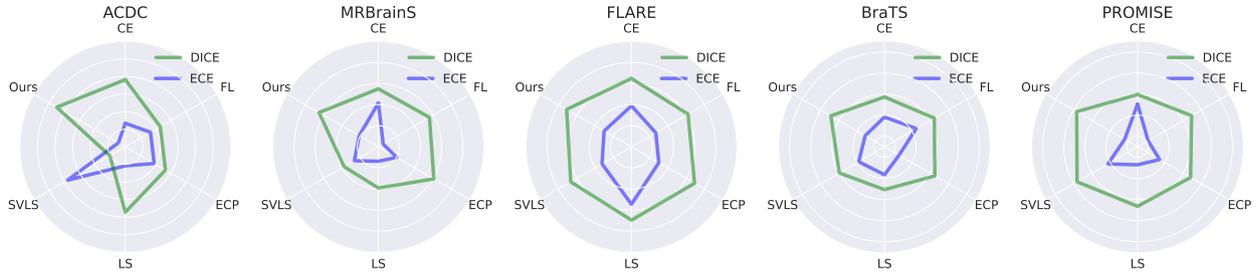
Figure 2: **Compromise between calibration and discriminative performance.** For each dataset, we show the discriminative (DICE) and calibration (ECE) results for each method. In order to get the best performance, we expect a model to achieve large DSC (*in green*) and small ECE (*in blue*) values.

the cost of providing non-informative solutions.

An interesting summary of these results is depicted in Figure 2, where we resort to radar plots to highlight the better compromise between discriminative and calibration performance shown by our model. In particular, a *well-calibrated* model should have a balanced compromise between a high discriminative power (*green line*) and low calibration metrics (*purple line*). This means that, following these plots, the larger the gap between green and purple lines, the better the compromise between discriminative and calibration performance.

Furthermore, to have a better overview of the general performance across different models, we follow the strategy followed in several MICCAI Challenges, e.g., MRBrainS (Mendrik et al., 2015b), where the final ranking is given as the sum of individual ranking metrics: $R_T = \sum_{m=0}^{|M|} r_m$, where r_m is the rank of the segmentation model for the metric m (mean)[1]. Thus, if a model ranks first in terms of DSC in the FLARE dataset, it will receive one point, whereas five points will be added in case the model ranks fifth. The final ranking is obtained after the overall scores R_T for each model are sorted in ascending order, and ranked from 1 to n. Figure 3 provides the rank comparison through heatmap visualization. It can be inferred that, for both discriminative and calibration metrics, our methods achieves the highest rank. Interestingly, the proposed loss term yields very competitive discriminative results, outperforming the popular compounded CE+DSC loss. It is noteworthy to highlight that the optimization goal of these two terms are different. Networks trained with CE tend to achieve a lower average negative log-likelihood over all the pixels, whereas using Dice as loss function should increase the discriminative performance, in terms of Dice. Thus, it is expected that the compounded loss brings the better of both worlds. Nevertheless, we can observe that this is not what happens in practice. On the one hand, the networks trained with CE+DSC loss rank among the best discriminative models (third in DSC and second in ASD). On the other hand, their calibration performance is substantially degraded, ranking last and second-last in ECE and CECE, respectively. These results align with recent findings (Mehrtash et al., 2020), which highlight the deficiencies of models trained with the DSC loss to deliver well-calibrated models. While adjusting the balancing hyperparameter could improve the performance on one task,

the results on the other task would likely degrade due to the different nature of both learning objectives. Thus, based on these observations, we argue that obtaining a good compromise between calibration and segmentation quality is hardly attainable with the popular CE+DSC loss, and promote our model as a better alternative.

	CE	CE+DSC	FL	ECP	LS	SVLS	Ours
DSC	24	21	22	18	24	25	6
ASD	20	15	23	19	21	30	12
ECE	24	33	13	15	17	25	11
CECE	18	25	22	10	30	24	9
Total	86	94	80	62	92	104	38

Figure 3: Ranking (*global* and *per-metric*) of the different methods based on the sum-rank approach.

5.2.2. Comparison to post-hoc calibration

The proposed approach is orthogonal to post-hoc calibration strategies, which can still be used after training, as long as there exists an independent validation set to find the optimal hyperparameters (for example T in temperature scaling). To demonstrate this, we now report the performance of pre-scaling and post-scaling for ACDC, FLARE, and PROMISE datasets across the different approaches. In particular, we have included two post-hoc calibration strategies. First, we use the standard Temperature scaling approach, referred to as TS, where a single value for the entire image is employed. Furthermore, we also include the Local Temperature Scaling (LTS) method in (Ding et al., 2021), which was recently proposed in the context of medical image segmentation and provides a temperature value at each image pixel. For both TS and LTS, the optimal temperature values are found by optimizing the network parameters to decrease the negative log likelihood on an independent validation set. From the quantitative comparison, which can be found in Table 3, it can be inferred that our method further benefits from scaling the raw softmax probability predictions. Interestingly, the calibration performance obtained by our method prior

---

[1]Note that the per-class scores are not used in the sum-rank computation.

Table 3: Calibration performance of post-hoc calibration methods: temperature scaling (TS) and Local Temperature Scaling (LTS) (Ding et al., 2021). Best method is highlighted in bold, whereas second best approach is underlined.

Dataset	Method	CE		CE + DICE		FL		ECP		LS		SVLS		Ours	
		ECE	CECE	ECE	CECE	ECE	CECE	ECE	CECE	ECE	CECE	ECE	CECE	ECE	CECE
ACDC	Pre	<u>0.079</u>	<u>0.073</u>	0.137	0.084	0.113	0.116	0.109	0.095	0.081	0.107	0.176	0.135	**0.061**	**0.069**
	TS	<u>0.077</u>	<u>0.073</u>	0.135	0.084	0.112	0.117	0.105	0.095	0.084	0.109	0.174	0.135	**0.055**	**0.065**
	LTS	0.067	<u>0.065</u>	0.070	0.076	0.127	0.125	0.080	0.094	<u>0.065</u>	0.072	0.118	0.116	**0.041**	**0.046**
FLARE	Pre	0.045	0.029	0.058	0.034	**0.033**	0.035	<u>0.037</u>	**0.027**	0.055	0.050	0.039	0.036	0.038	<u>0.028</u>
	TS	0.040	0.030	0.051	0.036	**0.030**	0.038	<u>0.032</u>	**0.028**	0.042	0.039	0.039	0.038	0.033	<u>0.029</u>
	LTS	0.033	0.030	0.044	0.038	0.065	0.048	**0.026**	0.028	0.031	<u>0.026</u>	0.040	0.036	<u>0.031</u>	**0.026**
PROMISE	Pre	0.411	0.334	0.430	0.304	<u>0.247</u>	0.298	0.306	<u>0.252</u>	0.280	0.299	0.344	0.271	**0.232**	**0.237**
	TS	0.408	0.334	0.429	0.304	<u>0.245</u>	0.299	0.303	<u>0.251</u>	0.279	0.298	0.342	0.271	**0.229**	**0.237**
	LTS	0.294	0.283	0.312	0.263	<u>0.209</u>	0.291	0.230	<u>0.235</u>	0.255	0.257	0.234	0.238	**0.189**	**0.217**

to temperature scaling still outperforms the results obtained by several other approaches even after applying LTS on their predictions. Another unexpected observation is that, under some settings, the use of temperature scaling (either TS or LTS) deteriorates the calibration performance. We argue that this phenomenon could be due to noticeable differences between the validation and testing datasets. As empirically demonstrated in (Ovadia et al., 2019), applying temperature scaling when differences between datasets exist might result in a negative impact. In addition, similar observations were reported in (Kock et al., 2021), where the calibration performance of segmentation models on several datasets was degraded after applying temperature scaling.

5.2.3. Effects of logit margin constraints

In our motivation, we hypothesized that the suboptimal supervision delivered by CE in multi-class scenarios might likely result in poorly calibrated models, as the posterior probability assigned to each of the non-true classes cannot be directly controlled. Indeed, it is expected that by minimizing the CE the softmax vectors are pushed towards the vertex of the probability simplex. This implies that the distances between the largest logit and the rest are magnified, resulting in overconfident models. To validate this hypothesis, and to empirically demonstrate that our proposed term can alleviate this issue, we plot the average logit distributions across classes on two datasets. In particular, we first separate all the voxels based on their ground truth labels. Then, for each category group, we average the per-voxel vector of logit predictions for both CE and the proposed model, whose results are depicted in Figure 4. First, we can observe that a model trained with CE indeed tends to provide large logit differences, which intensifies overconfidence predictions. Furthermore, while the mean logit value of the target class is considerably large and greatly differs from the largest value across other categories, the differences with the remaining logits –from non-target classes– remain uncontrolled. In contrast, we can clearly observe the impact on the logit distribution when we include the proposed term into the learning objective. In particular, our margin-based term *i)* promotes similar values of the true class logit across classes and *ii)* encourages more equidistant logits between this and the remaining classes, which implicitly constraints the logit values of untargeted classes to be very close (mimicking a uniform distribution). These re-

sults empirically validate our hypothesis in regards of the weaknesses of CE and the benefits brought by our approach.
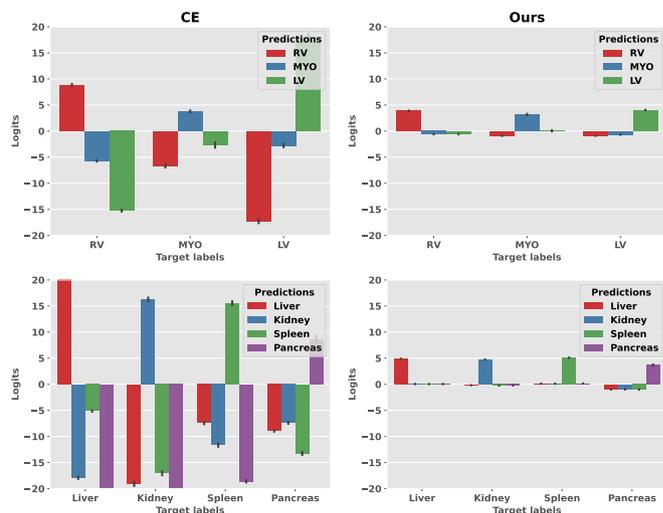

Figure 4: **Adopting the proposed term during training *substantially* reduces the logit distances, producing less overconfident predictions.** These plots depict the average predicted logit distributions for each target class –based on the ground truth– on ACDC (*top*) and FLARE (*bottom*) datasets when the model is trained with CE (*left*) and the proposed loss (*right*).
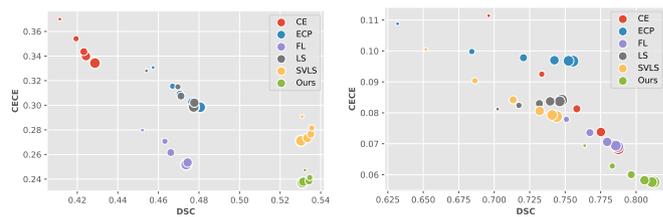

Figure 5: Robustness to distributional drift on PROMISE (*left*) and MRBrainS (*right*) datasets. Note that larger circles represent lower sigma values for the Gaussian noise corruptions.

5.2.4. Calibration and discriminative performance under distribution shift.

There have been recent empirical studies (Ovadia et al., 2019; Minderer et al., 2021) on the robustness of calibration models under distribution shift. In particular, (Minderer et al., 2021) explores out-of-distribution calibration by resorting to ImageNet-C (Hendrycks and Dietterich, 2018), a computer vision dataset
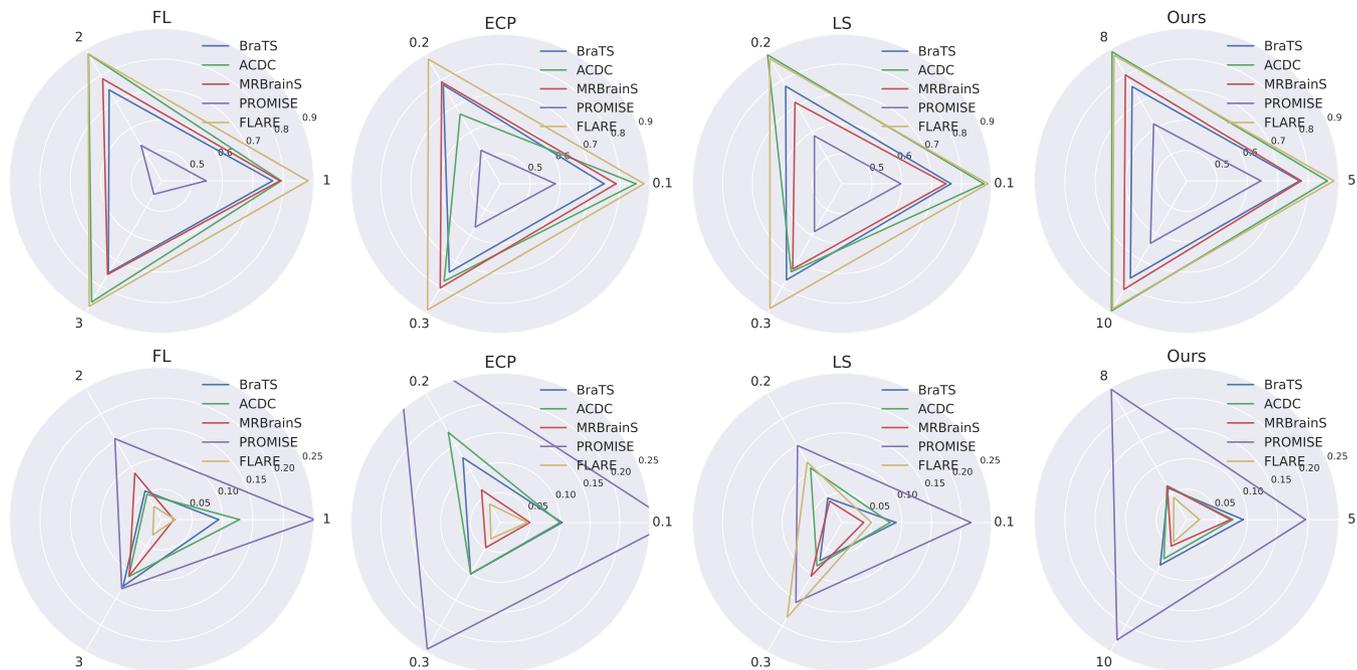
Figure 6: **Sensibility to hyperparameters across datasets.** For each method, we use the standard hyperparameters used in the literature and compare its variation across different datasets. The discriminative performance (DSC) is reported in the *top row*, whereas the calibration analysis (ECE) is depicted in the *bottom row*.

that contains images that have been synthetically corrupted, for example by including Gaussian noise. Inspired by these works, we now assess the robustness of our model in the presence of domain drift. To do this, we added Gaussian noise to the images on the testing set, with sigma values ranging from 0 to 0.05 with an increment of 0.01. From the plots in Figure 5 we can clearly observe that models trained with our objective function are less sensitive to noise, compared to prior state-of-the-art methods. More concretely, on the PROMISE dataset, the discriminative and calibration performance of our approach remains almost invariant to image perturbations with different levels of Gaussian noise. Furthermore, while the results obtained by our method in the MRBrainS data are affected by noise, its performance degradation is significantly lower than nearly all previous approaches, being on par with the focal loss. Nonetheless, it is noteworthy to mention that despite the relative decrease in performance is similar between the proposed method and FL, their global performance differences are substantially large (e.g., 6-8% difference in DSC). Based on these results we can argue that the proposed method delivers higher performing models that are, in addition, more robust to distributional drifts produced by Gaussian noise.

5.2.5. On the impact of hyperparameters

We now assess the sensitivity of each model to the choice of the hyperparameters on each dataset. We stress that, for each method, we have selected a range of common values used in the literature. In particular, γ is set to 1.0, 2.0 and 3.0 in Focal loss, λ is fixed to 0.1, 0.2 and 0.3 in ECP and Label smoothing, whereas the margin values in our method are set to 5.0, 8.0 and 10. The discriminative (DSC) and calibration (ECE) performances obtained across the different hyperparameter values are

depicted in Fig. 6. From this figure, it can be easily observed that, while prior approaches are very sensitive to the value of their balancing term, our method is significantly more robust to these changes. For example, the discriminative performance is drastically affected in both ECP and LS across several datasets when changing the value of the balancing term from 0.1 and 0.2 to 0.2 and 0.3, respectively. On the other hand, this phenomenon is more pronounced in the calibration metrics, where FL, ECP and LS show much higher variations than the proposed approach. A potential drawback that can be extracted from these findings is that, in order to get a well calibrated and high performance model, prior approaches might require multiple training iterations to find a satisfactory compromise. Furthermore, we believe that these large variations indicate that differences in the data –e.g., image contrast, target size and heterogeneity, or class distribution– might have a different impact on these losses, entangling the convergence of models trained with these terms.
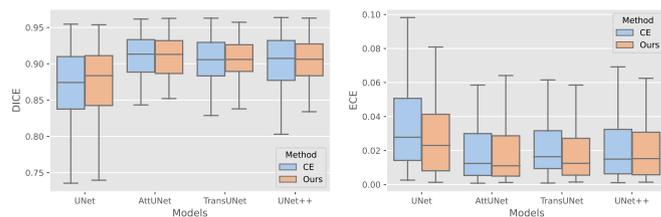

Figure 7: **Our loss is model agnostic.** Robustness to segmentation backbone, which evaluates the standard cross-entropy and the proposed model on the FLARE segmentation benchmark.
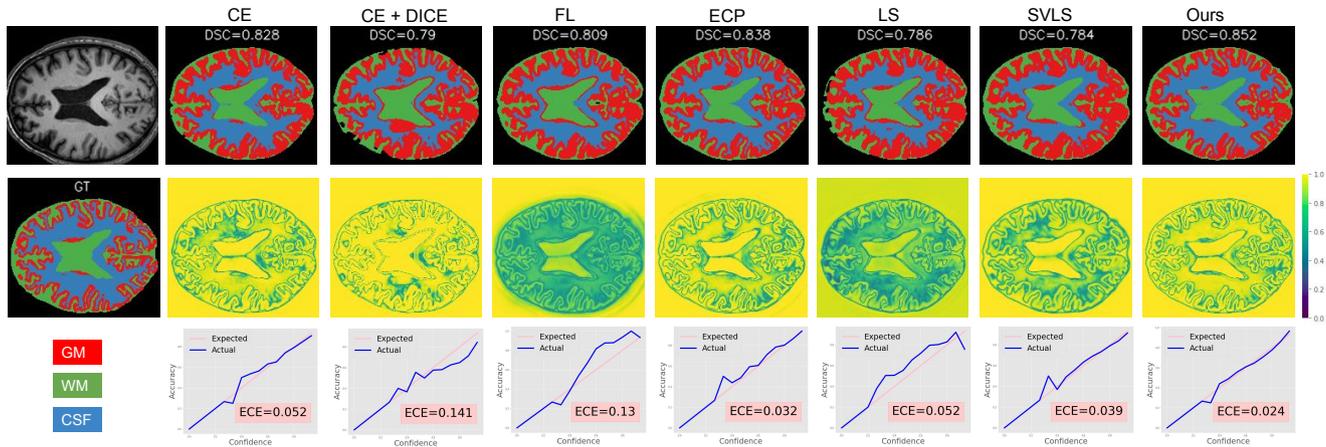
5.2.6. Robustness to backbone

Figure 8: Qualitative results on MRBrainS dataset for different methods. In particular, we show the original image and the corresponding segmentation masks provided by each method (*top row*), the ground-truth (GT) mask followed by maximum confidence score of each method (*middle row*) and the respective reliability plots (*bottom row*). Methods from left to right: CE, CE+DICE, FL, ECP, LS, SVLS, Ours.

In this experiment, we evaluate the proposed loss on several standard medical image segmentation networks, including: AttUNet (Oktay et al., 2018), TransUNet (Chen et al., 2021), and UNet++ (Zhou et al., 2020). For this study, we consider the FLARE dataset due to its larger number of classes. The quantitative comparison of CE and our method for these backbones is presented in Fig. 7, from which it can be inferred that, irrespective of the backbone used, our method is capable of consistently achieving better model calibration compared to the popular cross-entropy loss, while yielding at par performance in the discrimination task. We can therefore say that the proposed term is model agnostic, and the improvement observed is consistent across different models.

5.2.7. Qualitative results and reliability diagrams

Figure 8 depicts the predicted segmentation masks (*top*), confidence maps (*middle*) and their corresponding reliability plots (*bottom*) on one subject across the different methods. While the segmentation masks reveal several differences in terms of discriminative performance, the confidence maps present more interesting observations. Note that, as highlighted in prior works (Liu et al., 2022), better calibrated models should show better edge sharpness, matching the expected property that the model should be less confident at the boundaries, whereas yielding more confident predictions in inner target regions. First, we can observe that adding the DSC loss term substantially degrades the confidence map compared to its single CE loss counterpart. In particular, the CE+DSC compounded loss tends to produce smoother edges, in terms of confidence, which is derived from worst calibrated models. Furthermore, while it can increase the confidence of predictions in several inner object regions, it decreases this score in others. In addition, we can clearly observe that the remaining analyzed methods provide a diverse span of confidence estimates, with several models providing highly unconfident inner regions (e.g., FL (Mukhoti et al., 2020) and LS (Szegedy et al., 2016)). In contrast, our method yields confidence estimates that are sharp in the edges and low in within-region pixels, as expected in a well-calibrated model. These

visual findings are supported by the reliability diagrams. Indeed, our model yields the best reliability diagram, as the ECE curves are closer to the diagonal, indicating that the predicted probabilities serve as a good estimate of the correctness of the prediction.

6. Conclusion

Despite the popularity of network calibration in a broad span of applications, the connection between state-of-the-art calibration losses remains unexplored, and their impact on segmentation networks, particularly in the medical field, has largely been overlooked. In this work, we have demonstrated that these popular losses are closely related from a constrained optimization perspective, whose implicit or explicit constraints lead to non-informative solutions, preventing the model predictions to reach the best compromise between discriminative and calibration performance. To overcome this issue, we proposed a simple solution that integrates an inequality constraint into the main learning objective, which imposes a controlled margin on the logit distances. Through an extensive empirical evaluation, which contains multiple popular segmentation benchmarks, we have assessed the discriminative and calibration performance of state-of-the-art calibration losses in the important task of medical image segmentation. The results highlight several important benefits of the proposed loss. First, it achieves consistent improvements over state-of-the-art calibration and segmentation losses, both in terms of discriminative and calibration performance. Second, the proposed model is much less sensitive to hyperparameters changes compared to prior losses, which reduces the training time to find a satisfactory compromise between discrimination and calibration tasks. In addition, the empirical observations support our hypothesis that the suboptimal supervision delivered by the standard cross-entropy loss likely results in poorly calibrated models, as model trained with this loss tend to produce largest logit differences. Thus, we advocate that the proposed loss term should be preferred to train

models that provide higher discriminative performance, while yet delivering accurate uncertainty estimates.

Acknowledgements

References

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C., 2017. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data 4, 1–13.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al., 2018. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 .

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al., 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE TMI 37, 2514–2525.

Bertsekas, D., 1995. Nonlinear Programming. Athena Scientific, Belmont, MA.

Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D., 2015. Weight uncertainty in neural network, in: ICML.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 .

Ding, Z., Han, X., Liu, P., Niethammer, M., 2021. Local temperature scaling for probability calibration, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6889–6899.

Fort, S., Hu, H., Lakshminarayanan, B., 2019. Deep ensembles: A loss landscape perspective. arXiv preprint arXiv:1912.02757 .

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: ICML.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks, in: ICML.

Hendrycks, D., Dietterich, T., 2018. Benchmarking neural network robustness to common corruptions and perturbations, in: International Conference on Learning Representations.

Hernández-Lobato, J.M., Adams, R., 2015. Probabilistic backpropagation for scalable learning of bayesian neural networks, in: ICML.

Islam, M., Glocker, B., 2021. Spatially varying label smoothing: Capturing uncertainty from expert annotations, in: International Conference on Information Processing in Medical Imaging, pp. 677–688.

Jena, R., Awate, S.P., 2019. A bayesian neural net to segment images with uncertainty estimates and good calibration, in: International Conference on Information Processing in Medical Imaging, pp. 3–15.

Jungo, A., Balsiger, F., Reyes, M., 2020. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. Frontiers in neuroscience 14, 282.

Karimi, D., Gholipour, A., 2022. Improving calibration and out-of-distribution detection in deep models for medical image segmentation. IEEE Transactions on Artificial Intelligence .

Kock, F., Thielke, F., Chlebus, G., Meine, H., 2021. Confidence histograms for model reliability analysis and temperature calibration, in: Medical Imaging with Deep Learning.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles, in: NeurIPS.

Larrazabal, A.J., Martínez, C., Dolz, J., Ferrante, E., 2021. Orthogonal ensemble networks for biomedical image segmentation, in: MICCAI.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection, in: CVPR.

Litjens, G., Toth, R., van de Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., et al., 2014. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. Medical image analysis 18, 359–373.

Liu, B., Ben Ayed, I., Galdran, A., Dolz, J., 2022. The devil is in the margin: Margin-based label smoothing for network calibration, in: CVPR.

Louizos, C., Welling, M., 2016. Structured and efficient variational deep learning with matrix gaussian posteriors, in: ICML.

Lukasik, M., Bhojanapalli, S., Menon, A., Kumar, S., 2020. Does label smoothing mitigate label noise?, in: ICML.

Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., Cao, S., Zhang, Q., Liu, S., Wang, Y., Li, Y., He, J., Yang, X., 2021. Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence doi:10.1109/TPAMI.2021.3100536.

Ma, X., Blaschko, M.B., 2021. Meta-cal: Well-controlled post-hoc calibration by ranking, in: ICML.

Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T., 2020. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE transactions on medical imaging 39, 3868–3878.

Mendrik, A.M., Vincken, K.L., Kuijf, H.J., Breeuwer, M., Bouvy, W.H., De Bresser, J., Alansary, A., De Bruijne, M., Carass, A., El-Baz, A., et al., 2015a. Mrbrains challenge: online evaluation framework for brain image segmentation in 3t mri scans. Comput. Intell. Neurosci. 2015, 1.

Mendrik, A.M., Vincken, K.L., Kuijf, H.J., Breeuwer, M., Bouvy, W.H., De Bresser, J., Alansary, A., De Bruijne, M., Carass, A., El-Baz, A., et al., 2015b. Mrbrains challenge: online evaluation framework for brain image segmentation in 3t mri scans. Computational intelligence and neuroscience 2015.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.A., Arbel, T., Avants, B.B., Ayache, N., Buendia, P., Collins, D.L., Cordier, N., Corso, J.J., Criminisi, A., Das, T., Delingette, H., Demiralp, Ã., Durst, C.R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K.M., Jena, R., John, N.M., Konukoglu, E., Lashkari, D., Mariz, J.A., Meier, R., Pereira, S., Precup, D., Price, S.J., Raviv, T.R., Reza, S.M.S., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H.C., Shotton, J., Silva, C.A., Sousa, N., Subbanna, N.K., Szekely, G., Taylor, T.J., Thomas, O.M., Tustison, N.J., Unal, G., Vasseur, F., Wintermark, M., Ye, D.H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., Van Leemput, K., 2015. The multimodal brain tumor image segmentation benchmark (brats). IEEE Transactions on Medical Imaging 34, 1993–2024. doi:10.1109/TMI.2014.2377694.

Minderer, et al., 2021. Revisiting the calibration of modern neural networks, in: NeurIPS.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P.H., Dokania, P.K., 2020. Calibrating deep neural networks using focal loss, in: NeurIPS.

Müller, R., Kornblith, S., Hinton, G., 2019. When does label smoothing help?, in: NeurIPS.

Naeini, M.P., Cooper, G., Hauskrecht, M., 2015. Obtaining well calibrated probabilities using bayesian binning, in: Twenty-Ninth AAAI Conference on Artificial Intelligence.

Niculescu-Mizil, A., Caruana, R., 2005. Predicting good probabilities with supervised learning, in: Proceedings of the 22nd international conference on Machine learning, pp. 625–632.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 .

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J.V., Lakshminarayanan, B., Snoek, J., 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift, in: NeurIPS.

Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., Hinton, G., 2017. Regularizing neural networks by penalizing confident output distributions, in: ICLR.

Platt, J., et al., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers 10, 61–74.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Medical Image Computing and

Computer-Assisted Intervention – MICCAI 2015, pp. 234–241.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision, in: CVPR.

Tomani, C., Gruber, S., Erdem, M.E., Cremers, D., Buettner, F., 2021. Post-hoc uncertainty calibration for domain drift scenarios, in: CVPR.

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., 2019. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing 338, 34–45.

Wenzel, F., Snoek, J., Tran, D., Jenatton, R., 2020. Hyperparameter ensembles for robustness and uncertainty quantification, in: NeurIPS.

Xie, L., Wang, J., Wei, Z., Wang, M., Tian, Q., 2016. Disturblabel: Regularizing cnn on the loss layer, in: CVPR.

Zhang, J., Kailkhura, B., Han, T., 2020. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning, in: ICML.

Zhang, Z., Dalca, A.V., Sabuncu, M.R., 2019. Confidence calibration for convolutional neural networks using structured dropout. arXiv preprint arXiv:1906.09551 .

Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2020. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Transactions on Medical Imaging 39, 1856–1867. doi:10.1109/TMI.2019.2959609.