



Neighbor-Aware Calibration of Segmentation Networks with Penalty-Based Constraints

Balamurali Murugesan^{a,*}, Sukesh Adiga Vasudeva^a, Bingyuan Liu^b, Herve Lombaert^a, Ismail Ben Ayed^a, Jose Dolz^a

^aÉTS Montréal, Canada

^bAmazon, Canada

ARTICLE INFO

Article history:

Received 1 September 2022

ABSTRACT

Ensuring reliable confidence scores from deep neural networks is of paramount significance in critical decision-making systems, particularly in real-world domains such as healthcare. Recent literature on calibrating deep segmentation networks has resulted in substantial progress. Nevertheless, these approaches are strongly inspired by the advancements in classification tasks, and thus their uncertainty is usually modeled by leveraging the information of individual pixels, disregarding the local structure of the object of interest. Indeed, only the recent *Spatially Varying Label Smoothing (SVLS)* approach considers pixel spatial relationships across classes, by softening the pixel label assignments with a discrete spatial Gaussian kernel. In this work, we first present a constrained optimization perspective of SVLS and demonstrate that it enforces an implicit constraint on soft class proportions of surrounding pixels. Furthermore, our analysis shows that SVLS lacks a mechanism to balance the contribution of the constraint with the primary objective, potentially hindering the optimization process. Based on these observations, we propose NACL (Neighbor Aware CaLibration), a principled and simple solution based on equality constraints on the logit values, which enables to control explicitly both the enforced constraint and the weight of the penalty, offering more flexibility. Comprehensive experiments on a wide variety of well-known segmentation benchmarks demonstrate the superior calibration performance of the proposed approach, without affecting its discriminative power. Furthermore, ablation studies empirically show the model agnostic nature of our approach, which can be used to train a wide span of deep segmentation networks. The code is available at <https://github.com/Bala93/MarginLoss>

© 2024 Elsevier B. V. All rights reserved.

1. Introduction

Despite the remarkable progress made by deep neural networks (DNNs) in a wide span of recognition tasks, there exists growing evidence suggesting that these models are poorly calibrated, leading to overconfident predictions that may assign high confidence to incorrect predictions (Gal and Ghahramani,

2016; Guo et al., 2017). This represents a major problem, as inaccurate uncertainty estimates can carry serious implications in safety-critical applications such as medical diagnosis, whose outcomes are used in subsequent tasks of critical importance. The underlying cause of miscalibration in deep models is hypothesized to stem from their high capacity, which makes them prone to overfitting on the negative log-likelihood loss, commonly used during training (Guo et al., 2017). Indeed, modern classification networks trained under the fully supervised learning paradigm resort to binary one-hot encoded vectors as super-

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

visory signals of training data points. 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.

In light of the significance of this issue, there has been a surge in popularity for quantifying the predictive uncertainty in modern DNNs. A simple approach involves a post-processing step that modifies the softmax probability predictions of an already trained network (Guo *et al.*, 2017; Tomani *et al.*, 2021; Zhang *et al.*, 2020; Ding *et al.*, 2021). These methods, however, see their performance degrade under distributional drifts (Ovadia *et al.*, 2019). More principled alternatives incorporate a term that maximizes the Shannon entropy of the model predictions during training, penalizing confident output distributions. This regularization term is either implicitly derived from the original loss (Mukhoti *et al.*, 2020; Müller *et al.*, 2019) or explicitly integrated as additional learning objectives (Pereyra *et al.*, 2017; Liu *et al.*, 2022, 2023).

Due to the importance of correctly modeling the uncertainty estimates in deep segmentation models, just a few works have recently studied the impact of existing approaches in this problem (Jena and Awate, 2019; Larrazabal *et al.*, 2021; Ding *et al.*, 2021; Murugesan *et al.*, 2023b). Nevertheless, these approaches are directly borrowed from the classification literature, which presents important limitations in the segmentation scenario. In particular, dense prediction tasks, such as image segmentation, greatly benefit from capturing pixel relationships due to the ambiguity in the boundaries between neighboring organs or regions. Indeed, the nature of structured predictions in segmentation involves pixel-wise classification based on spatial dependencies, which limits the effectiveness of these strategies to yield performances similar to those observed in classification tasks (Mukhoti *et al.*, 2020; Müller *et al.*, 2019; Liu *et al.*, 2022). This potentially suboptimal performance can be attributed to the uniform (or near-to-uniform) distribution enforced on the softmax/logits distributions, which disregards the spatial context information. While modeling these pixel-wise relationships, for example, by modeling the class distribution around a given pixel, is extremely important, virtually none of existing methods explicitly considers these relationships.

To address this important issue, Spatially Varying Label Smoothing (SVLS) (Islam and Glocker, 2021) introduces a soft labeling approach that captures the structural uncertainty required in semantic segmentation. In practice, smoothing the hard-label assignment is achieved through a Gaussian kernel applied across the one-hot encoded ground truth, which results in soft class probabilities based on neighboring pixels. Nevertheless, while the reasoning behind this smoothing strategy relies on the intuition of giving an equal contribution to the central label and all surrounding labels combined, its impact on the training, from an optimization standpoint, has not been studied.

We can summarize our **contributions** as follows:

- In this work, we provide a constrained-optimization perspective of Spatially Varying Label Smoothing (SVLS)

(Islam and Glocker, 2021), demonstrating that it could be viewed as a standard cross-entropy loss coupled with an implicit constraint that enforces the softmax predictions to match a soft class proportion of surrounding pixels. Our formulation shows that SVLS lacks a mechanism to control explicitly the importance of the constraint, which may hinder the optimization process as it becomes challenging to balance the constraint with the primary objective effectively.

- Following these observations, we propose a simple and flexible solution based on equality constraints on the logit distributions. The proposed constraint is enforced with a simple linear penalty, which incorporates an explicit mechanism to control the weight of the penalty. Our approach not only offers a more efficient strategy to model the logit distributions but implicitly decreases the logit values, which results in less overconfident predictions.
- We conduct comprehensive experiments and ablation studies over multiple medical image segmentation benchmarks, including diverse targets and modalities, and show the superiority of our method compared to state-of-the-art calibration losses. Furthermore, several ablation studies empirically validate the design choices of our approach, as well as demonstrate its model agnostic nature.

This journal version provides a substantial extension of the preliminary work presented in (Murugesan *et al.*, 2023a). More concretely, we first provide a thorough literature review on calibration models, with an extensive overview of their use in medical image segmentation. Second, we perform a comprehensive empirical validation, including *i*) multiple additional public benchmarks covering diverse modalities and targets, *ii*) several ablation studies that motivate our choices, *iii*) showing the agnostic nature of NACL regarding the segmentation backbone, and *iv*) additional results that help us to understand the underlying benefits of the proposed approach.

2. Related work

Post-processing approaches. A straightforward and effective approach to mitigate the miscalibration issue involves implementing 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). In this scenario, a validation set, drawn from the generative distribution of the training data $\pi(X, Y)$ is leveraged to rescale the network outputs, resulting in well-calibrated in-domain predictions. Temperature scaling (TS) (Guo *et al.*, 2017), a simple generalization of Platt scaling (Platt *et al.*, 1999) to the multi-class setting, uses a single value overall logit (i.e., pre-softmax) predictions to control the shape of the class predicted distributions. (Tomani *et al.*, 2021) proposes to transform the validation set before transforming the softmax distributions, whereas (Zhang *et al.*, 2020) combines isotonic regression (IR) after performing temperature scaling. Despite its efficiency, most approaches within

this family present important limitations, including *i*) a dataset-dependency on the value of the transformation parameters and *ii*) a significant degradation observed on out-of-domain samples (Ovadia et al., 2019).

Penalizing low-entropy predictions. To alleviate the issue of overconfident predictions inherent in minimizing a negative log-likelihood loss, a natural strategy is to encourage high-entropy, i.e., uncertain, predictions. A straightforward solution to achieve this is to include into the learning objective a term to penalize confident output distributions by explicitly maximizing the entropy (Pereyra et al., 2017). More recently, several works (Müller et al., 2019; Mukhoti et al., 2020) have shed light into the implicit calibration properties of popular losses (label smoothing and focal loss) that modify the one-hot encoding labels used for training. More concretely, label smoothing (Szegedy et al., 2016) has been shown to implicitly calibrate the trained models, as it prevents the network from assigning the full probability mass to a single class, while encouraging the differences between the logits of the target class and the other categories to be a constant dependent on α^1 (Müller et al., 2019). In addition, (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 network predictions, thereby increasing the entropy of the predictions. Thus, we can see both label smoothing and focal loss as classification losses that implicitly regularize the network output probabilities, encouraging their distribution to be close to the uniform distribution. More recently, (Liu et al., 2022) presented a unified view of state-of-the-art calibration approaches (Pereyra et al., 2017; Szegedy et al., 2016; Lin et al., 2017) showing that these strategies can be viewed as approximations of a linear penalty enforcing equality constraints on logit distances, which are encouraged to be zero across all the logits. This view exposes important limitations of the ensuing gradients, which constantly push towards a non-informative solution, compromising an optimal trade-off between discriminative and calibration performance. To circumvent this limitation, authors proposed a simple and flexible generalization of label smoothing (MbLS) based on inequality constraints, which imposes a controllable margin on logit distances.

Calibration in medical image segmentation. Despite recent efforts to model the predictive uncertainty, or to leverage this uncertainty to improve the discriminative performance of segmentation models (Wang et al., 2019), little attention has been devoted to improving both the calibration and segmentation performance of deep models in the medical domain. (Jena and Awate, 2019) presented a Bayesian decision theoretic framework based on deep models for image segmentation. This framework produced analytical estimates of uncertainty, allowing to define a principled measure of uncertainty associated with label probabilities, which led to an improvement on both segmentation and calibration performances. Nevertheless, there exists recent evidence (Fort et al., 2019) that indicates that

¹In label smoothing, α controls the mass that is uniformly distributed across the different classes: $y_k^{L.S} = y_k(1 - \alpha) + \alpha/K$.

Bayesian neural networks tend to find solutions around a single minimum of the loss landscape, resulting in a lack of diversity. In contrast, ensembling multiple deep neural networks usually yields more diverse predictions, consequently leading to improved uncertainty estimates which outperform other methods (Jungo et al., 2020; Mehrtash et al., 2020). In the context of medical image segmentation, several strategies have been adopted to promote model diversity within the ensemble, such as imposing orthogonality constraints during training (Larrazabal et al., 2021) or training a single model in a multi-task manner on several different datasets (Karimi and Gholipour, 2022). A main drawback of these approaches, however, lies in their increased complexity cost, as they entail the training of either multiple models or a single model on multiple datasets.

Ding et al. (2021) present a lighter alternative that extends the simple temperature scaling approach by integrating a shallow neural network to predict the voxel-wise temperature values, which are used in a post-processing step. While this method outperforms the naive TS, it inherits the limitations of temperature scaling and related post-processing approaches. More recently, (Murugesan et al., 2023b) performed a comprehensive evaluation of existing calibration approaches in the task of medical image segmentation. The reported results suggested that methods integrating explicit penalties, and in particular MbLS (Liu et al., 2022), largely outperformed other existing techniques in both discrimination and calibration metrics. All these methods, however, are predominantly adopted from the classification literature, which ignores the underlying properties of dense prediction problems, such as semantic image segmentation. In these cases, the spatial relations between a given pixel and its neighbors play a crucial role in the predictions, and the surrounding class distributions in the pixel-wise annotations should be considered for modeling the uncertainty. Indeed, and as to the best of our knowledge, the work in (Islam and Glocker, 2021) is the only method that considers the pixel vicinity of the labeled mask to improve the calibration performance of deep segmentation models. More concretely, authors apply a Gaussian kernel across the one-hot encoded labels to obtain soft class probabilities, integrating spatial-awareness into the standard label smoothing process.

3. Methodology

3.1. Preliminaries

Notation. Let us denote the training dataset as $\mathcal{D} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^N$, where the set of N pairs are *i.i.d.* realizations of the random variables X, Y which follow a ground truth joint distribution $\pi(X, Y) = \pi(X|Y)\pi(X)$. In this setting, $\mathbf{x}^{(n)} \in \mathbb{R}^{\Omega_n}$ represents the n^{th} image, Ω_n the spatial image domain, and $\mathbf{y}^{(n)} \in \mathbb{R}^K$ its corresponding ground-truth label with K classes, provided as a one-hot encoding vector. For simplicity and clarity in the formulation, we will omit in what follows the superscript to indicate the sample used, and \mathbf{x} will denote any image in the training set. Now, given an input image \mathbf{x} , a neural network parameterized by θ generates the set of logit predictions $f_\theta(\mathbf{x}) = \mathbf{l} \in \mathbb{R}^{\Omega_n \times K}$. Last, we use the

softmax function, denoted as $\phi(\cdot)$ to obtain the predicted model probabilities $\hat{\mathbf{p}} = \phi(f_\theta(\mathbf{x})) \in \mathbb{R}^{\Omega_n \times K}$.

What is calibration? Calibration measures the correspondence between the predicted probabilities assigned by a model and the empirical likelihood of the associated events. A well-calibrated model ensures that its predicted probabilities align with the actual observed frequencies of outcomes. For instance, when the model assigns a probability of 0.7 to an event, it is expected that this event materializes approximately 70% of the time in the empirical data. In a classification scenario, we can formally define that a model presents *perfect calibration of confidence* if the following conditional probability holds:

$$\mathbb{P}(\hat{y} = y | \hat{p} = p) = p, \quad \forall p \in [0, 1], \quad (1)$$

where $\hat{y} = \arg \max(\hat{\mathbf{p}})$ is the predicted class of input image \mathbf{x} , and $\hat{p} = \max(\hat{\mathbf{p}})$ its associated confidence. Equation 1 tells us that, to be perfectly calibrated, when the model predicts the probability distribution $\phi(f_\theta(X))$ over the set of classes $[K] = \{1, 2, \dots, K\}$, the true probability distribution for these categories should be $\phi(f_\theta(X))$. Thus, any difference between the left and right terms is known as calibration error, or *miscalibration*.

3.2. A constrained optimization perspective of SVLS

Spatially Varying Label Smoothing (SVLS) (Islam and Glocker, 2021) considers the surrounding class distribution of a given pixel p in the ground truth \mathbf{y} to estimate the amount of smoothness over the one-hot label of that pixel. In particular, let us consider that we have a 2D patch \mathbf{x} of size $d_1 \times d_2$ and its corresponding ground truth \mathbf{y}^2 . Furthermore, the predicted softmax in a given pixel is denoted as $\mathbf{s} = [s_0, s_1, \dots, s_{k-1}]$. Let us now transform the surrounding patch of the segmentation mask around a given pixel into a unidimensional vector $\mathbf{y} \in \mathbb{R}^d$, where $d = d_1 \times d_2$. SVLS employs a discrete Gaussian kernel \mathbf{w} to obtain soft class probabilities from one-hot labels, which can also be reshaped into $\mathbf{w} \in \mathbb{R}^d$. Thus, for a given pixel j , and a class k , SVLS Islam and Glocker (2021) can be defined as:

$$\tilde{y}_j^k = \frac{1}{|\sum_i^d w_i|} \sum_{i=1}^d y_i^k w_i. \quad (2)$$

We can replace the smoothed labels \tilde{y}_p^k in the standard cross-entropy (CE) loss, resulting in the following learning objective:

$$\mathcal{L} = - \sum_k \left(\frac{1}{|\sum_i^d w_i|} \sum_{i=1}^d y_i^k w_i \right) \log \hat{p}_j^k, \quad (3)$$

where s_p^k is the softmax probability for the class k at pixel p (the pixel in the center of the patch). Now, we can decompose this loss into:

$$\mathcal{L} = - \frac{1}{|\sum_i^d w_i|} \sum_k y_j^k \log \hat{p}_j^k \quad (4)$$

$$- \frac{1}{|\sum_i^d w_i|} \sum_k \left(\sum_{\substack{i=1 \\ i \neq j}}^d y_i^k w_i \right) \log \hat{p}_j^k, \quad (5)$$

with p denoting the index of the pixel in the center of the patch. Note that the term in the left is the cross-entropy between the posterior softmax probability and the hard label assignment for pixel p . Furthermore, let us denote $\tau_k = \sum_{i \neq j}^d y_i^k w_i$ as the soft proportion of the class k inside the patch/mask \mathbf{y} , weighted by the filter values \mathbf{w} . By replacing τ_k into the Eq. 4, and removing $|\sum_i^d w_i|$ as it multiplies both terms, the loss becomes:

$$\mathcal{L} = - \underbrace{\sum_k y_j^k \log \hat{p}_j^k}_{CE} - \underbrace{\sum_k \tau_k \log \hat{p}_j^k}_{\text{Constraint on } \tau}. \quad (6)$$

As τ is constant, the second term in Eq. 6 can be replaced by a Kullback-Leibler (KL) divergence, leading to the following learning objective:

$$\mathcal{L} \stackrel{c}{=} \mathcal{L}_{CE} + \mathcal{D}_{KL}(\tau || \hat{\mathbf{p}}), \quad (7)$$

where $\stackrel{c}{=}$ stands for equality up to additive and/or non-negative multiplicative constant. Thus, optimizing the loss in SVLS results in minimizing the cross-entropy between the hard label and the softmax probability distribution on the pixel j , while imposing the equality constraint $\tau = \hat{\mathbf{p}}$, where τ depends on the class distribution of surrounding pixels. Indeed, this term implicitly enforces the softmax predictions to match the soft-class proportions computed around pixel j .

3.3. Proposed constrained calibration approach

Our previous analysis exposes two important limitations of SVLS: 1) the importance of the implicit constraint cannot be controlled explicitly, and 2) the prior τ is derived from the σ value in the Gaussian filter, making it difficult to model properly. To alleviate this issue, we propose a simple solution, which consists in minimizing the standard cross-entropy between the softmax predictions and the one-hot encoded masks coupled with an explicit and controllable constraint on the logits \mathbf{l} . In particular, we propose to minimize the following constrained objective:

$$\min_{\theta} \mathcal{L}_{CE} \quad \text{s.t.} \quad \tau = \mathbf{l}, \quad (8)$$

²For the sake of simplicity, we consider a patch as an image \mathbf{x} (or mask \mathbf{y}), whose spatial domain Ω is equal to the patch size, i.e., $d_1 \times d_2$.

where τ now represents a desirable prior, and $\tau = 1$ is a hard constraint. Note that the reasoning behind working directly on the logit space is two-fold. First, observations in Liu et al. (2022) suggest that directly imposing the constraints on the logits results in better performance than in the softmax predictions. And second, by imposing a bounded constraint on the logits values³, their magnitudes are further decreased, which has a favorable effect on model calibration Müller et al. (2019). We stress that despite both Liu et al. (2022) and our method enforce constraints on the predicted logits, Liu et al. (2022) is fundamentally different. In particular, Liu et al. (2022) imposes an *inequality* constraint on the logit distances so that it encourages uniform-like distributions up to a given margin, disregarding the importance of each class in a given patch. This can be important in the context of image segmentation, where the uncertainty of a given pixel may be strongly correlated with the labels assigned to its neighbors. In contrast, our solution enforces *equality* constraints on an adaptive prior, encouraging distributions close to class proportions in a given patch.

Even though the constrained optimization problem presented in Eq. 8 could be solved by a standard Lagrangian-multiplier algorithm, we replace the hard constraint by a soft penalty of the form $\mathcal{P}(|\tau - 1|)$, transforming our constrained problem into an unconstrained one, which is easier to solve. In particular, the soft penalty \mathcal{P} should be a continuous and differentiable function that reaches its minimum when it verifies $\mathcal{P}(|\tau - 1|) \geq \mathcal{P}(\mathbf{0})$, $\forall l \in \mathbb{R}^K$, i.e., when the constraint is satisfied. Following this, when the constraint $|\tau - 1|$ deviates from $\mathbf{0}$ the value of the penalty term increases. Thus, we can approximate the problem in Eq. 8 as the following simpler unconstrained problem:

$$\min_{\theta} \mathcal{L}_{CE} + \lambda \sum_k |\tau_k - l_k|, \quad (9)$$

where the hyperparameter λ controls the importance of the penalty.

4. Experiments

4.1. Experimental Setting

4.1.1. Datasets

To empirically validate our model, we resort to six public multi-class segmentation benchmarks, whose details are provided below.

Automated Cardiac Diagnosis Challenge (ACDC) (Bernard et al., 2018). This dataset comprises short-axis cardiac cine-MRI scans from 100 patients, in both diastolic and systolic phases with their respective segmentation annotations. The task of this challenge is to understand the cardiac function through segmenting key regions, including the left ventricle (LV), the right ventricle (RV), and the myocardium (Myo). Following

standard practices, we randomly split the dataset into 70 patients for training, 10 for validation, and the remaining 20 for testing. From each of these volumes, we extract 2D slices, which are resized to 224×224 .

Brain Tumor Segmentation (BRATS) 2019 Challenge (Menze et al., 2015; Bakas et al., 2017, 2018). The goal of this challenge is to identify glioma tumors in multi-channel MRI scans (FLAIR, T1, T1-contrast, and T2). The dataset consists of 335 volumes with their corresponding segmentation masks, which include tumor core (TC), enhancing tumor (ET), and whole tumor (WT). Following prior works, we randomly split the volumes into subsets of 235, 35, and 65 scans for training, validation, and testing, respectively. We also resample the volumes, extract the 2D slices and discard the empty slices.

Fast and Low GPU memory Abdominal oRgan sEgmentation (FLARE) Challenge (Ma et al., 2021). This dataset contains 360 abdominal CT scans obtained from diverse medical centers with pixel-wise masks of several organs, including liver, kidneys, spleen, and pancreas. Following standard protocols, we randomly split the scans into 240 for training, 40 for validation, and 80 for testing. Furthermore, CT scans with different resolutions are resampled to the same space and cropped to $192 \times 192 \times 30$, from which 2D slices are obtained.

PROSTATE (Antonelli et al., 2022) The dataset was acquired at Radboud University Medical Center and was released as a part of Medical Segmentation Decathlon (MSD) challenge. The dataset consists of 32 MRI volumes with target regions of prostate peripheral zone (PZ) and the transition zone (TZ). The dataset is challenging because of segmenting two adjoining regions large inter-subject variability. We split the dataset to 22 patients for training, 3 for validation and 7 for testing.

Kidney Tumor Segmentation (KiTS) challenge (Heller et al., 2019). This dataset consists of 210 CT scans with their respective segmentation masks, including the kidney and tumor classes. Following (Islam and Glocker, 2021), we resampled cases with varying resolutions and image sizes to a common resolution of $3.22 \times 1.62 \times 1.62$ mm and center crop to image size $80 \times 160 \times 160$. The dataset is randomly split into 150 cases for training, 25 for validation, and 40 for testing.

MRBrainS18 (Mendrik et al., 2015). The purpose of this challenge is to segment the brain MRI scans into Gray Matter (GM), White Matter (WM), and Cerebrospinal fluid (CSF). The dataset contains paired T1, T2, and T1-IR sequences of 3D volumes ($240 \times 240 \times 48$) of 7 subjects and their associated pixel-wise masks. For the experiments, we consider 5 subjects for training and 2 for testing.

Note that in all the 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 performed leave-one-out-cross-validation in our experiments, whereas the other datasets followed standard training, validation, and testing procedures,

³Note that the proportion priors are generally normalized.

using a single split in the experiments.

4.1.2. Evaluation Metrics

We assess the discriminative performance of the model using standard segmentation metrics in the medical imaging community, including the overlap-based metric DICE (DSC) coefficient, and spatial distance metric Hausdorff distance (HD). For understanding the calibration performance, we resort to Expected Calibration Error (ECE) and Classwise Expected Calibration Error (CECE) (Naeini *et al.*, 2015). ECE concentrates only on maximum confidence score of the prediction, while CECE considers the confidence distribution of all the classes, including the winner class (Mukhoti *et al.*, 2020). Importantly, we obtain the calibration metrics only for the foreground regions following the recent literature (Islam and Glocker, 2021; Murugesan *et al.*, 2023b). The notion behind this is because the class distribution is skewed towards background, particularly in most cases of medical image segmentation. Hence, excluding background allows us to better compare the performance of different methods. We further understand the calibration performance through reliability plots (Niculescu-Mizil and Caruana, 2005), wherein accuracy is expected to be directly correlated to class probability. In both the cases, we set the number of bins to 15.

To compute ECE and CECE for N samples with K classes, we group predictions into M equispaced bins. Let B_i denote the set of samples with maximum confidences belonging to the i^{th} bin, and $B_{i,j}$ denotes the set of samples from the j^{th} class in the i^{th} bin. The accuracy A_i of i -th 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. Similarly, for class-wise, the accuracy is given by $A_{i,j} = \frac{1}{|B_{i,j}|} \sum_{k \in B_{i,j}} 1(j = y_k)$. The confidence C_i of the i^{th} bin and $C_{i,j}$ of i^{th} bin, j^{th} class is given by $C_i = \frac{1}{|B_i|} \sum_{j \in B_i} \hat{p}_j$ and $C_{i,j} = \frac{1}{|B_{i,j}|} \sum_{k \in B_{i,j}} \hat{p}_{kj}$ respectively. Hence, ECE and CECE is given by:

$$ECE = \sum_{i=1}^M \frac{|B_i|}{N} |A_i - C_i| \quad (10)$$

$$CECE = \sum_{i=1}^M \sum_{j=1}^K \frac{|B_{i,j}|}{N} |A_{i,j} - C_{i,j}| \quad (11)$$

4.1.3. 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 employ standard loss functions employed in medical image segmentation, which include the popular Cross-entropy (CE) combined with DSC loss. Furthermore, we also include training objectives that have been proposed to calibrate deep neural networks for both classification and segmentation problems, 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), ECP (Pereyra *et al.*, 2017), SVLS (Islam and Glocker, 2021), and MbLS (Liu *et al.*, 2022). Following the literature, we have chosen the following hyper-parameters for the different approaches: FL ($\gamma=3.0$), ECP ($\alpha=0.1$), LS ($\lambda=0.1$), SVLS ($\sigma=2.0$) and MbLS ($m=5$). Note that in the main experiments, these hyperparameters remain fixed across the different datasets for all the models, to better highlight the generability of each approach. For the experiments, we fixed the batch size to 16, 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. The models are trained on 2D slices, while the evaluation is done over 3D volumes. The best model is selected based on the mean DSC score on the validation dataset.

Backbones. The experiments are predominantly conducted on the standard UNet (Ronneberger *et al.*, 2015) architecture. Nevertheless, to demonstrate the model-agnostic nature of our approach we also evaluate the effect of our method on other common architectures in medical image segmentation, including convolutional neural networks (AttUNet (Oktay *et al.*, 2018), UNet++ (Zhou *et al.*, 2020) and nnUNet (Isensee *et al.*, 2021)) and Vision Transformer based architectures (TransUNet (Chen *et al.*, 2021)).

4.2. Results

4.2.1. Main results

We present the quantitative results across a diverse set of segmentation datasets, which include multiple organs, pathologies, as well as several imaging protocols, from a segmentation and a calibration standpoint.

Segmentation results. First, in Table 1, we compare the discriminative performance of our Neighbor Aware CaLibration method, which we refer to as NACL, to relevant calibration approaches. Notably, we can observe that our approach consistently outperforms existing literature across nearly all the datasets and metrics, yielding improvements which range between 3.4% and 10% (DSC), compared to the second and last performing method, respectively. Indeed, if we consider the mean DSC and HD values for each dataset, the proposed approach achieves the best performance in 10 out of the 12 settings, being the second and third best performance method in the remaining 2 scenarios. An important observation is that, whereas our method typically ranks first and second for all targets and metrics, there is no other approach that presents a consistent trend on performance across datasets. For example, Focal loss yields the second best average DSC performance in BraTS, while it ranks last in ACDC or MRBrains.

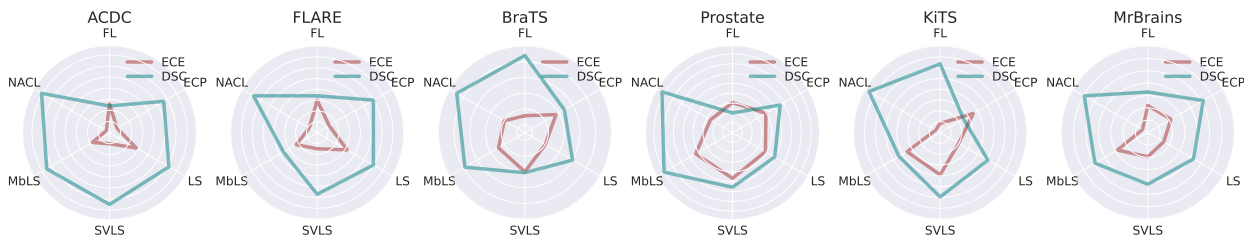
Calibration performance. Similarly to the segmentation scenario, the results in terms of calibration (Table 2) reveal that our approach consistently yields the best, and second best, uncertainty estimates across datasets and target objects. Furthermore, and as observed in Table 1, there is no a clear trend on the prior literature, as methods performing competitively in one dataset considerably fail in another, whose discrepancies can also be

Table 1: Discriminative performance obtained by the different evaluated models across six popular medical image segmentation benchmarks. Best method is highlighted in bold, whereas second best approach is underlined.

Dataset	Region	CE+DSC		FL		ECP		LS		SVLS		MbLS		NACL	
		DSC \uparrow	HD \downarrow	DSC \uparrow	HD \downarrow	DSC \uparrow	HD \downarrow	DSC \uparrow	HD \downarrow	DSC \uparrow	HD \downarrow	DSC \uparrow	HD \downarrow	DSC \uparrow	HD \downarrow
ACDC	RV	0.799	3.10	0.580	9.37	0.751	4.93	0.796	3.34	0.791	<u>2.89</u>	<u>0.812</u>	2.59	0.837	3.02
	MYO	0.795	<u>2.57</u>	0.557	5.55	0.757	3.54	0.772	3.07	0.798	2.66	0.795	2.86	0.820	2.04
	LV	<u>0.889</u>	3.75	0.724	6.97	0.839	4.85	0.858	3.49	0.882	<u>2.89</u>	<u>0.875</u>	3.53	0.905	2.59
	Mean	<u>0.828</u>	3.14	0.620	7.30	0.782	4.44	0.809	3.30	0.824	<u>2.81</u>	0.827	2.99	0.854	2.55
FLARE	Liver	0.950	<u>6.09</u>	0.952	7.54	<u>0.953</u>	7.41	0.952	8.50	0.951	7.72	0.941	7.18	0.954	6.04
	Kidney	0.945	2.07	0.947	2.16	<u>0.950</u>	2.05	0.947	1.76	0.947	1.84	0.937	2.49	0.952	1.84
	Spleen	0.892	9.49	0.887	9.09	0.887	3.98	0.905	4.62	0.879	6.40	0.868	4.73	<u>0.900</u>	4.26
	Pancreas	0.636	7.95	0.626	7.80	0.649	7.77	0.637	6.45	<u>0.650</u>	6.91	0.596	8.61	0.664	7.37
	Mean	0.855	6.40	0.853	6.65	<u>0.860</u>	5.30	0.860	5.33	0.857	5.72	0.836	5.75	0.867	4.88
BraTS	TC	0.731	5.73	0.799	7.80	0.749	7.53	0.773	5.16	0.744	7.56	<u>0.803</u>	4.88	0.804	3.98
	ET	0.766	8.27	<u>0.854</u>	10.02	0.790	11.31	0.807	10.23	0.783	9.22	0.821	10.85	0.854	6.58
	WT	0.872	6.88	0.889	9.19	0.884	7.28	0.879	7.94	0.877	8.55	<u>0.889</u>	8.09	0.893	6.78
	Mean	0.789	6.96	<u>0.848</u>	9.00	0.808	8.71	0.820	7.78	0.801	8.44	0.838	7.94	0.850	5.78
PROSTATE	CG	0.329	16.00	0.223	23.45	0.344	19.97	0.292	13.51	0.341	15.24	0.427	10.93	0.418	<u>12.73</u>
	PZ	0.752	7.13	0.677	12.57	0.736	6.19	0.756	<u>5.12</u>	0.737	9.28	<u>0.774</u>	5.65	0.796	4.02
	Mean	0.540	11.56	0.450	18.01	0.540	13.08	0.524	9.31	0.539	12.26	<u>0.601</u>	8.29	0.607	<u>8.37</u>
KiTS	Kidney	0.786	9.11	<u>0.784</u>	8.74	0.735	10.27	0.759	<u>9.06</u>	0.770	9.86	0.749	10.56	0.780	9.08
	Tumor	0.447	13.09	<u>0.470</u>	<u>13.57</u>	0.365	15.49	0.446	16.61	0.468	15.96	0.426	16.85	0.525	15.77
	Mean	0.616	11.10	<u>0.627</u>	<u>11.15</u>	0.550	12.88	0.602	12.83	0.619	12.91	0.588	13.71	0.652	12.42
MRBrainS	GM	<u>0.754</u>	1.73	0.672	2.81	0.747	2.23	0.707	2.12	0.725	<u>1.71</u>	0.741	2.09	0.781	1.41
	WM	0.759	2.91	0.598	5.60	<u>0.783</u>	<u>2.73</u>	0.702	4.98	0.603	6.24	0.729	3.08	0.791	2.64
	CSF	0.776	2.00	0.722	4.18	0.746	3.10	0.730	2.34	<u>0.800</u>	<u>1.41</u>	0.769	1.71	0.820	1.21
	Mean	<u>0.763</u>	<u>2.22</u>	0.664	4.20	0.759	2.68	0.713	3.15	0.709	3.12	0.747	2.29	0.797	1.75

Table 2: Calibration performance obtained by the different evaluated models across six popular medical image segmentation benchmarks. Best method is highlighted in bold, whereas second best approach is underlined. In this case, the calibration metrics are averaged across the different target objects.

Dataset	CE+DSC		FL		ECP		LS		SVLS		MbLS		NACL	
	ECE \downarrow	CECE \downarrow	ECE \downarrow	CECE \downarrow	ECE \downarrow	CECE \downarrow	ECE \downarrow	CECE \downarrow	ECE \downarrow	CECE \downarrow	ECE \downarrow	CECE \downarrow	ECE \downarrow	CECE \downarrow
ACDC	0.137	0.084	0.153	0.179	0.130	0.094	<u>0.083</u>	0.093	0.091	0.083	0.103	<u>0.081</u>	0.048	0.061
FLARE	0.058	0.034	0.053	0.059	<u>0.037</u>	0.027	0.055	0.049	0.039	0.036	0.046	0.041	0.033	<u>0.031</u>
BraTS	0.178	0.122	0.097	0.119	0.139	0.100	0.112	0.108	0.146	0.111	0.127	0.095	<u>0.112</u>	<u>0.097</u>
PROSTATE	0.430	0.304	<u>0.271</u>	0.381	0.306	<u>0.252</u>	0.304	0.301	0.335	0.272	0.322	0.250	0.253	0.254
KiTS	0.188	0.144	<u>0.098</u>	<u>0.133</u>	0.155	<u>0.151</u>	0.122	0.141	0.163	0.144	0.155	0.147	0.090	0.124
MRBrainS	0.177	0.105	0.085	0.123	0.084	0.082	<u>0.061</u>	0.101	0.077	<u>0.080</u>	0.107	0.093	0.027	0.056

Figure 1: **Compromise between calibration and discriminative performance.** For each dataset, we show the discriminative (DSC) and calibration (ECE) results obtained by each method. We expect a *well-calibrated* model to achieve simultaneously large DSC (*in blue*) and small ECE (*in brown*) values.

observed across metrics. For instance, Focal loss yields the best calibrated model, in terms of ECE, for the BraTS dataset, but its ECE value in ACDC is three times higher than the ECE obtained by our approach. This phenomenon is also observed in other approaches, such as ECP (best CECE in FLARE and worst in KiTS) or MbLS (best CECE in BraTS and PROSTATE, but among the worst in MRBrainS). It is important to note that these methods contain different hyperparameters that remained fixed across datasets (e.g., α in LS, γ in FL, or λ and margin m in MbLS). Thus, even though a specific per-dataset fine-tuning of these hyperparameters may lead to a performance increase (both in terms of segmentation and calibration), results in Ta-

ble 1 and 2 demonstrate empirically that our approach presents a robust alternative to existing methods, as it yields the overall best performance across diverse target objects and datasets.

For a more comprehensive understanding of the overall performance across various approaches and datasets, we now introduce two studies that expand upon the quantitative values provided in Table 1 and 2. First, we resort to radar plots in Figure 1 to better highlight the trade-off between discriminative and calibration performance achieved by different methods. For a model to be *well-calibrated*, it should present high discriminative performance (*blue line*), while yielding low calibration values (*brown line*). In the case of these radar plots, this im-



Figure 2: Ranking *global* and *per-metric* of the different methods based on the sum-rank and mean of case-specific approach.

plies that a greater distance between the blue and brown lines indicates a more favorable balance between discriminative and calibration performance. Looking at the plots, we can easily observe that the proposed method consistently yields the best trade-offs across datasets, offering high discriminative power without degrading its calibration performance. Other methods, however, must sometimes compromise their discriminative performance to produce calibrated models, or vice-versa. The second study considers the evaluation strategies adopted in several MICCAI Challenges, i.e., sum-rank (Mendrik *et al.*, 2015) and mean-case-rank (Maier *et al.*, 2017). As we can observe in the heatmaps provided in Fig. 2, our approach yields the best rank across all the metrics in both strategies, clearly outperforming any other method. Interestingly, some methods such as FL or ECP typically provide well-calibrated predictions, but at the cost of degrading their discriminative performance.

4.2.2. On the impact of constraining the logit space

Constraint over logits vs softmax. Recent evidence (Liu *et al.*, 2022; Murugesan *et al.*, 2023b) have suggested that imposing constraints on the logits offers a better alternative than its softmax counterpart. To demonstrate that this observation holds in our model, we further present the results of our formulation when the constraint is enforced on the softmax distributions, i.e., replacing l by \hat{p} in Equation 9. From these results, reported in Figure 3, it is evident that working on the logit space substantially increases both the segmentation and calibration performance across the datasets. This could be attributed to the range of logits being larger than softmax, allowing for a better control.

Effect on the logit distributions. In order to demonstrate the benefits of our method over existing approaches, in terms of controlling the logits, we have plotted the average logit distribution across classes on ACDC and FLARE test sets in Figure 4. In particular, we first separate all the voxels based on their ground truth labels. Then, for each category, we average the per-voxel vector of logit predictions across each category (in absolute value). From the figure, it can be inferred that the popular CE+DSC loss provides higher logit values for the winner class, and the distance between the winner logits and rest are large, typical characteristics of an overconfident model (Murugesan *et al.*, 2023b). Interestingly, SVLS seems to follow the logit distribution of CE+DSC, up to a given extent, even though it was designed to emulate LS, but integrating class spa-

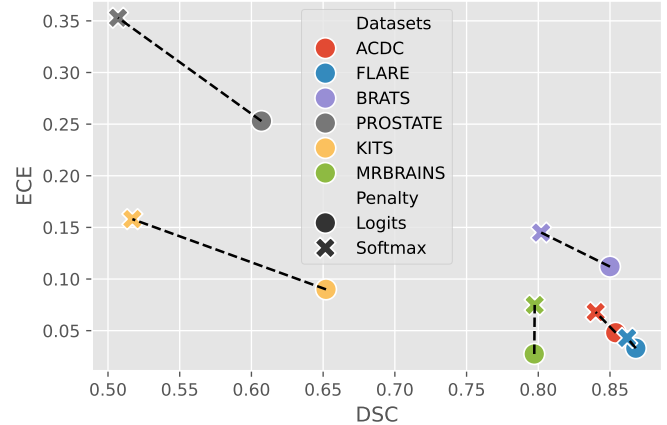


Figure 3: Impact of applying the penalty over softmax (cross) vs logits (circle) predictions across the different datasets.

tial information. In contrast, whereas LS and MbLS have a desired logit distribution for calibration, particularly for the winner class, the distance with the remaining categories is shorter. This may have an undesirable effect, as predictions where the distance between the winner and remaining logits are very small may lack semantic information needed for maintaining the discriminative performance. Finally, our approach brings the best of both worlds, i.e., it keeps the magnitude of the winner logit low, which facilitates the training of a well-calibrated model, effectively pushes the remaining logit values to a considerable distance, thereby preserving robust discriminative power.

To further understand how the different methods control the logit predictions, we plot the maximum logit distribution over epochs during training, which is depicted in Fig. 5. It is well known that, calibrated methods show a better regularization, restricting the range of logits to a particular range (Müller *et al.*, 2019). From the figure, it could be observed that, during initial epochs, most of the methods show similar distribution. However, as the number of epochs increases, several methods focusing on improving the calibration performance have a narrower range. Indeed, only LS, MbLS and the proposed NACL approach present the narrowest logit distribution when the network has been trained during a large number of epochs. Based on the findings in (Müller *et al.*, 2019), we can therefore say that our method presents very strong regularization capabilities compared to other approaches, as the range of logits provided by the trained model is very restricted, with most of the logits encountered between a value of 4 and 5.

4.2.3. On the impact of hyperparameters

In this experiment, we assess the sensitivity of the hyperparameters in the different methods, and possibly find a setting which works best across datasets. For FL, γ values of 1, 2, and 3 are considered. In the case of ECP and LS, α and λ are set to of 0.1, 0.2 and 0.3. For MbLS, we considered the margins to be 5, 8, and 10, while λ was fixed to 0.1. In the case of SVLS, we fixed the kernel size to 3 and used 0.5, 1, and 2 as sigma values. Finally, we fixed λ in our method to 0.1, 0.2 and 0.3. We compared the discriminative (DSC) and calibration (ECE)

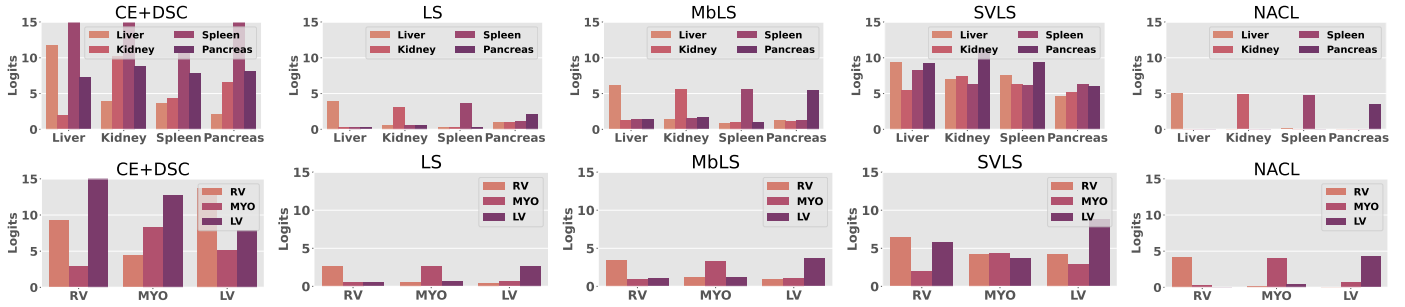


Figure 4: Distribution of logit predictions provided by a model trained with CE+DSC, LS, MbLS, SVLS and our approach (from left to right) on FLARE (top) and ACDC (bottom).

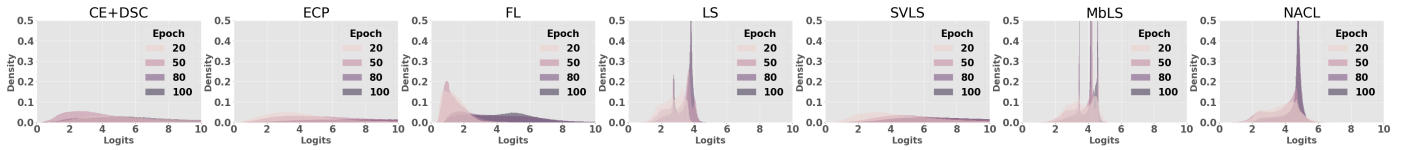


Figure 5: Histogram of global logit distribution over epochs obtained by the different approaches.

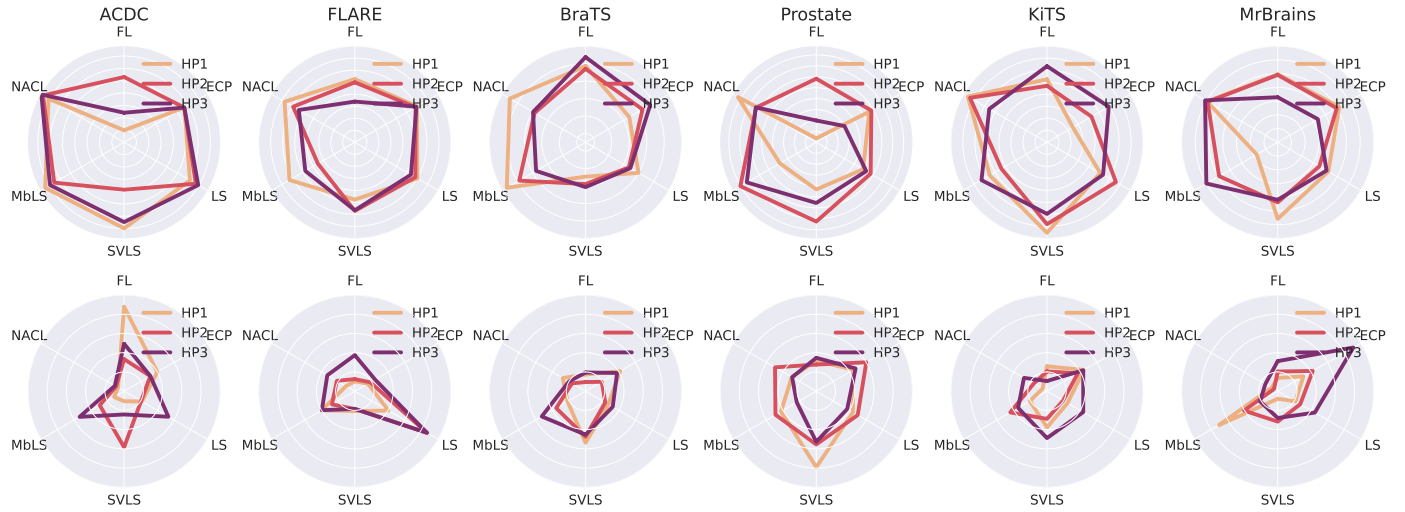


Figure 6: Radar plots displaying hyperparameter-dependence performance (DSC on top and ECE in the bottom). HP1, HP2 and HP3 denote the respective hyper-parameter set: FL ($\gamma=[1,2,3]$), ECP ($\lambda=[0.1,0.2,0.3]$), LS ($\alpha=[0.1,0.2,0.3]$), MbLS ($m=[3,5,10]$) and SVLS ($\sigma=[0.5,1,2]$, and ours ($\lambda=[0.1,0.2,0.3]$). Our method consistency provides best performance for 0.1 across datasets.

performances using these hyper-parameters across the different datasets and depicted the results in Figure 6. From this figure, it can be observed that, our method is fairly consistent with a particular hyper-parameter (HP1). Moreover, varying λ does not typically result in drastic performance degradation, which demonstrates the robustness of our approach to different hyper-parameter values. In contrast, other methods presented larger performance variations, as discrimination and calibration metrics were highly sensitive to the hyper-parameter choice. For example, in ACDC, focal loss and SVLS suffer large performance degradation for different values of their respective hyper-parameters, whereas in MrBrainS, ECP and MbLS results considerably decrease across different values of λ and m , respectively.

4.2.4. Effect of the prior

Ablation on different priors. A benefit of the proposed formulation, particularly compared to SVLS (Islam and Glocker, 2021), is that diverse priors can be enforced on the logit distributions. Thus, we now assess the impact of different priors, τ in our formulation, that can distribute the label distribution. The results presented in Table 3 reveal that selecting a suitable prior can further improve the performance of our model.

Varying sigma with a Gaussian prior. One of the advantages of the proposed approach compared to SVLS is its flexibility to include any prior in the constraint, as well as the integration of a blending parameter that controls the influence of the constraint during training. We now compare the impact of employing different sigma values in both SVLS and our approach. In particular, we use the following values in the Gaussian filter ($\sigma = \{1, 2, 3\}$) used in SVLS, as well to define a Gaussian

Table 3: **Impact of using different priors.** We compare the discriminative and calibration performance of our approach across the six datasets when using different priors τ in Equation 9.

	FLARE		ACDC		BraTS		PROSTATE		KiTS		MRBrainS		Mean	
Prior τ	DSC	ECE	DSC	ECE	DSC	ECE	DSC	ECE	DSC	ECE	DSC	ECE	DSC	ECE
Mean	0.868	0.033	0.854	0.048	0.850	0.112	0.607	0.253	0.652	0.090	0.797	0.027	0.771	0.094
Gaussian	0.860	0.033	0.876	0.042	0.813	0.140	0.559	0.293	0.615	0.134	0.779	0.045	0.750	0.115

prior in our formulation, whose results are depicted in Fig. 7. In this figure, the x-axis represents the relative difference in performance between our method and SVLS. More precisely, if we look at the top row for $\sigma = 1$, we can observe that in the ACDC dataset, the proposed approach outperforms SVLS by nearly 10%, whereas in PROSTATE, SVLS obtains nearly 2% improvement over our method. Taking this information into account, one can clearly see that, using the same prior, the proposed approach typically outperforms SVLS, and sometimes by a large margin, in both DSC and ECE metrics. Importantly, our approach achieves these results even without changing the weighing factor (λ), as it fixed to 0.1 to have a fair comparison to SVLS, since SVLS cannot control the importance of the penalty, as exposed in Section 3.2. These results show empirically that our method is able to better leverage the neighboring class information compared to SVLS.

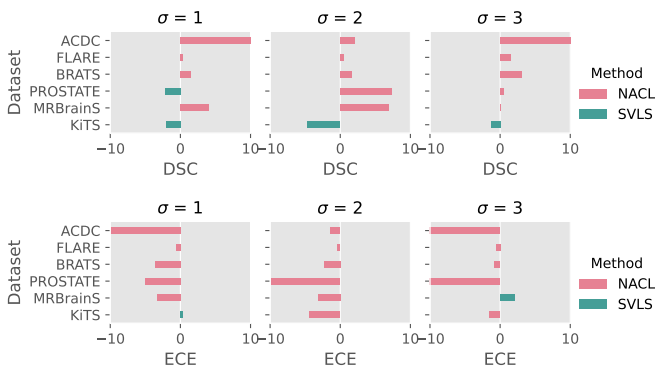


Figure 7: **Direct comparison of SVLS (Islam and Glocker, 2021) vs. NACL (Ours).** Relative error differences (%) between SVLS and our method when using the same Gaussian prior (with $\sigma = \{1, 2, 3\}$).

4.2.5. Robustness to backbone

We study the impact of our proposed loss when using other recent state-of-the-art segmentation networks including: AtUNet (Oktay et al., 2018), TransUNet (Chen et al., 2021), UNet++ (Zhou et al., 2020), and nnUNet (Isensee et al., 2021). We considered the FLARE dataset for this study, whose quantitative results, compared to MbLS and SVLS (our closest competitors in terms of methodology) are presented in Fig. 8. From the figure, it can be inferred that, regardless of the backbone choice, our method is able to consistently improve both segmentation and calibration performance. This can be attributed to the ability of our method to control the logit distribution, enabling it to be directly plugged into any standard segmentation architecture.

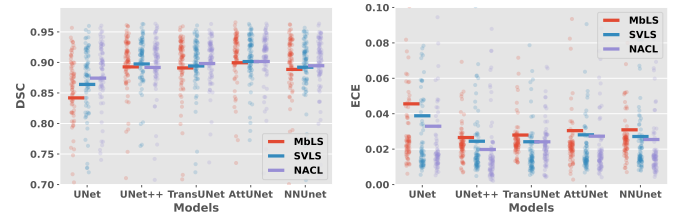


Figure 8: **Robustness to the segmentation backbone.** We evaluate the performance of competing approaches (i.e., MbLS and SVLS) on the FLARE dataset when using different architectures as segmentation backbones.

4.2.6. Sensitivity to the number of training samples

In this experiment, we investigate whether varying the number of training samples impacts the performance of the calibration methods. Indeed, one source of uncertainty in machine learning models is the lack of enough data, which is referred to as *epistemic* uncertainty, or knowledge uncertainty. While this kind of uncertainty can be addressed by adding more knowledge, for example in the form of additional labeled training samples, we want to evaluate how different calibration models behave under different labeled data scenarios. To do so, instead of considering all the samples for training, we only employ 25%, 50% and 75% of the available images. Note that, we use the same validation and test data as we did in our main experiments. Fig. 9 depicts the obtained results for ACDC and FLARE datasets. From these experiments, it is expected that decreasing the number of samples potentially impacts both the discriminative and calibration performance across all the methods. Nevertheless, this trend is not followed by several methods, particularly in terms of correctly modeling the uncertainty. For instance, ECP and SVLS present worst calibration performances for the 50% and 75% settings in ACDC, which is also observed in the DSC metrics. Last, across all the labeled scenarios, our approach yields typically the best performance, indicating that it can better handle the epistemic uncertainty derived from lack of enough knowledge during training.

4.2.7. Choice of the penalty

In this work, we have shown that regularizing the logits based on their neighboring class distribution coupled with the conventional cross entropy is helpful in improving both segmentation and calibration performance. For all the experiments, we have considered a linear penalty to enforce the spatial information. In this section, we now try to control the logits through a quadratic penalty instead. Table 4 presents the comparison of our method with L_1 and L_2 penalties. From these results, we can observe L_2 provides better segmentation results over L_1 in more cases, even though in some cases the improvement gains are marginal.

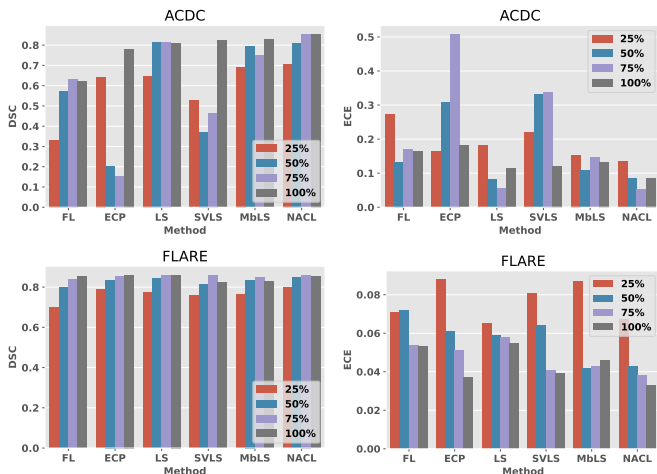


Figure 9: **Performance variation with number of labeled images.** These plots depict the performance of different approaches under several data labeled scenarios, going from 100% (i.e., original provided data) to 25% of images from the original dataset.

Nevertheless, when it underperforms its linear counterpart, the performance gap is significant (e.g., -6% in PROSTATE). In terms of calibration, L_1 yields the best performance in multiple cases. This could be due the nature of L_2 , which is more aggressive in forcing the logits to follow the prior class distribution compared to L_1 . It is important to note that, increasing the weighing factor (λ) of the penalty could mitigate the aggressiveness of L_2 to enforce the constraint, potentially leading to the improvement of the segmentation and calibration quality over L_1 . However, the goal of this work is to provide a unique solution that generalizes across multiple diverse datasets, and that does not require fine-tuning multiple hyper-parameters in each scenario. Thus, we did not explore individual configurations that lead to the best performance for each dataset.

	DSC		ECE	
	L_1	L_2	L_1	L_2
ACDC	0.854	0.871	0.048	0.059
FLARE	0.868	0.851	0.033	0.065
BraTS	0.850	0.851	0.112	0.078
PROSTATE	0.607	0.541	0.253	0.320
KiTS	0.652	0.673	0.090	0.106
MRBrainS	0.797	0.803	0.027	0.023
Mean	0.771	0.765	0.094	0.109

Table 4: **Impact of different penalties.** Comparison of using a L_1 vs a L_2 penalty to impose the constraint in Equation 9.

4.2.8. Calibration metrics over prediction and target foregrounds

Through all the experiments, the calibration metrics have been obtained by using only the foreground regions of the ground truth. Nevertheless, there is a possibility that a model prediction may be discarded, as it might not overlap with the target ground truth due to an over-segmentation. In this experiment, we recompute the calibration metrics over the union of

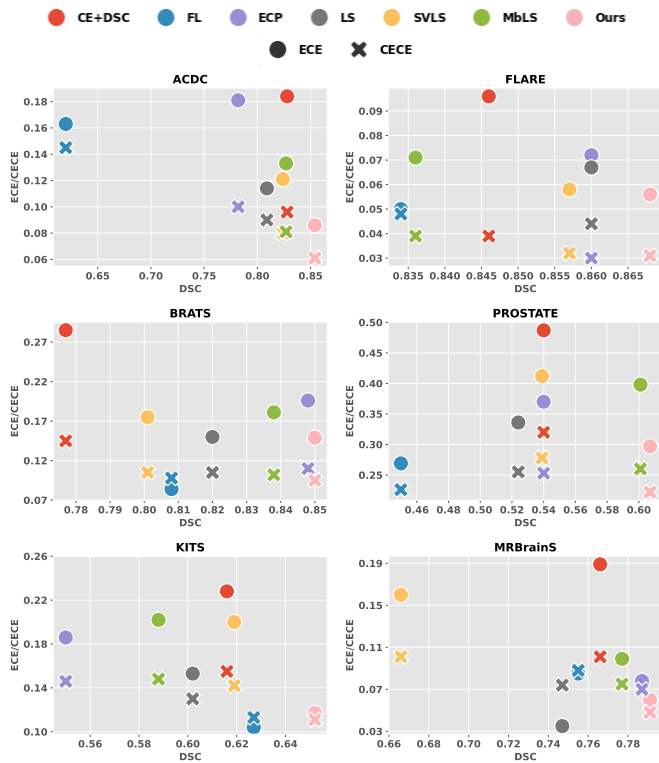


Figure 10: Scatter plots comparing DSC vs ECE/CECE when considering the foreground (prediction \cup target) to compute the calibration metrics.

target and predicted foregrounds, whose ECE and CECE values, against the DSC metric, are depicted in Figure 10. We can observe that, even after including the prediction regions in obtaining the calibration metrics, our method still yields the best performance trade-off between DSC and both ECE and CECE across all the datasets. Hence, the strategy for assessing the calibration performance does not change the message that the proposed approach offers a better alternative to existing calibration methods.

4.2.9. Qualitative results and reliability diagrams

Last, we show in Figure 11 the predicted segmentation masks (*top*), uncertainty maps (*middle*) and their corresponding reliability plots (*bottom*) on one subject across the different methods. From the predicted segmentation outputs, it is evident that our method generates segmentations closer to the target, which is supported quantitatively by the reported DSC metric. Methods such as MbLS, LS, FL tend to oversegment several categories, whereas ECP and SVLS have difficulties in differentiating challenging regions. The uncertainty maps given by the maximum confidence scores provide more interesting observations on the dynamics of the different methods. Note that, as highlighted in prior works (Liu et al., 2022), the model should be less confident at the boundaries, while providing more confident predictions in the inner regions. First, we can observe that the CE+DSC compound loss provides the worst calibrated models, as there are no remarkable edges to demarcate between regions. Second, methods such as FL and LS achieve better uncertainty by reducing the overall confidence scores across many

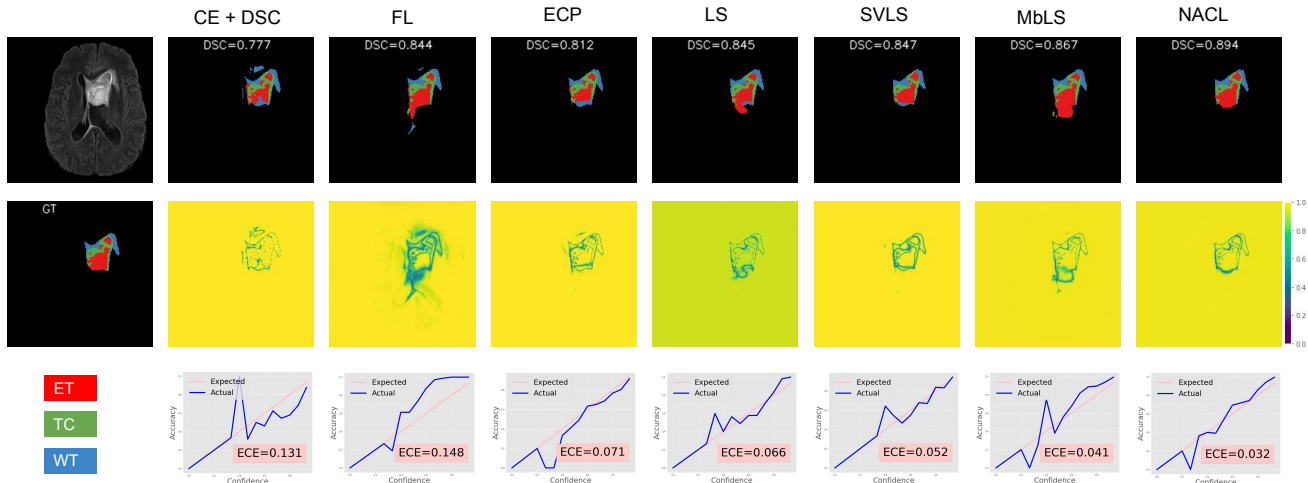


Figure 11: Qualitative results on BraTS dataset for different methods. In particular, we show the original image and the corresponding segmentation masks provided by each method (*top row*), the ground-truth (GT) mask followed by maximum confidence score of each method (*middle row*) and the respective reliability plots (*bottom row*). Methods from left to right: CE+DSC, FL, ECP, LS, SVLS, MbLS, and Ours

regions, which might impact the discriminative performance (as supported by quantitative results reported in previous sections). Third, SVLS provides a distinct edge map, but not particularly sharp because of the smoothing effect of the Gaussian filter. Finally, we could observe that MbLS, as well as our approach, provide confidence estimates that are sharp in the edges and low in within-region pixels, as expected in a well-calibrated model. However, it should be noted that MbLS uses a margin to control the magnitude of the logits, and lacks spatial awareness, as this value is chosen empirically and is equal for all the pixels. This contrasts with our method, where the prior is dynamically chosen depending on the neighboring class distribution for each pixel. Furthermore, we show the our model yields the best reliability diagram, i.e., ECE curves are closer to the diagonal, indicating that the predicted probabilities serve as a good estimate of the correctness of the prediction.

5. Conclusion

While network calibration has emerged as a mainstay problem in machine learning, most state-of-the-art calibration losses are specifically designed for classification problems, ignoring the spatial information, crucial in dense prediction tasks. Indeed, only the recent SVLS integrates spatial awareness to transform the hard one-hot encoding labels into a smoother version, capturing the class distribution surrounding each pixel. Inspired by the need of leveraging neighboring information to improve the calibration performance of deep segmentation models, in this work we delve into the details of SVLS, and present a constrained optimization perspective of this approach. Our analysis demonstrates that SVLS enforces an implicit constraint on soft class proportions of surrounding pixels. Our formulation exposed two weaknesses of SVLS. First, it lacks a mechanism to control explicitly the importance of the constraint, which may hinder the optimization process as it becomes challenging to balance the constraint with the primary objective ef-

fectively. And second, the *a priori* knowledge enforced in the constrained is directly derived from the Gaussian distribution of a pixel neighborhood, which may be difficult to define (as it depends on σ), and did not always provide the best performance, as shown empirically in our results.

To overcome the limitations of SVLS, we proposed a principled and simple approach based on equality constraints on the logit values, which allows us to control explicitly both the prior to be enforced in the constraint, as well as the weight of the penalty, offering more flexibility. We conducted a comprehensive evaluation, incorporating diverse well-known segmentation benchmarks, to evaluate the performance of the proposed approach, and compared it to state-of-the-art calibration losses in the crucial task of medical image segmentation. The empirical findings demonstrate that our approach outperforms existing approaches in both discriminative and calibration metrics. Furthermore, the proposed formulation yields stable results across multiple segmentation backbones, hyper-parameter values, and several labeled data scenarios, establishing itself as a robust alternative within the current literature.

While the proposed solution offers superior performance to existing approaches, there exist multiple avenues which are worth to explore. For example, a limitation of our approach is that it disregards image intensity information, which sometimes emerges as the source of annotation uncertainty. Thus, incorporating surrounding image intensity in the constraint could potentially lead to better results. Furthermore, simple penalties (i.e., linear and quadratic) have been explored to enforce the proposed constraint. Integrating more powerful strategies, for example based on log-barrier methods, have shown interesting performance gains in medical imaging problems (Kervadec *et al.*, 2022). Therefore, the exploration of these strategies to enforce the imposed constraints could shed light into more powerful alternatives in our formulation.

Acknowledgements

This work is supported by the National Science and Engineering Research Council of Canada (NSERC), via its Discovery Grant program and FRQNT through the Research Support for New Academics program. We also thank Calcul Quebec and Compute Canada.

References

- Antonelli, M., Reinke, A., Bakas, S., et al., 2022. The medical segmentation decathlon. *Nature communications* 13, 4128.
- Bakas, S., Akbari, H., Sotiras, A., et al., 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., 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., Lalonde, A., Zotti, C., et al., 2018. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging* 37, 2514–2525.
- Chen, J., Lu, Y., Yu, Q., et al., 2021. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.
- Ding, Z., Han, X., Liu, P., and 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., and Lakshminarayanan, B., 2019. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*.
- Gal, Y. and Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: *international conference on machine learning*, PMLR. pp. 1050–1059.
- Guo, C., Pleiss, G., Sun, Y., and Weinberger, K.Q., 2017. On calibration of modern neural networks, in: *International conference on machine learning*, PMLR. pp. 1321–1330.
- Heller, N., Sathianathan, N., Kalapara, A., et al., 2019. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*.
- Isensee, F., Jaeger, P.F., Kohl, S.A., et al., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* 18, 203–211.
- Islam, M. and 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. and 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., and Reyes, M., 2020. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. *Frontiers in neuroscience* 14, 282.
- Karimi, D. and Gholipour, A., 2022. Improving Calibration and Out-of-Distribution Detection in Deep Models for Medical Image Segmentation. *IEEE Transactions on Artificial Intelligence*.
- Kervade, H., Dolz, J., Yuan, J., et al., 2022. Constrained deep networks: Lagrangian optimization via log-barrier extensions, in: *2022 30th European Signal Processing Conference (EUSIPCO)*, IEEE. pp. 962–966.
- Larrazabal, A.J., Martínez, C., Dolz, J., and Ferrante, E., 2021. Orthogonal ensemble networks for biomedical image segmentation, in: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021*, pp. 594–603.
- Lin, T.Y., Goyal, P., Girshick, R., et al., 2017. Focal loss for dense object detection, in: *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988.
- Liu, B., Ben Ayed, I., Galdran, A., and Dolz, J., 2022. The devil is in the margin: Margin-based label smoothing for network calibration, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 80–88.
- Liu, B., Rony, J., Galdran, A., et al., 2023. Class Adaptive Network Calibration, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16070–16079.
- Ma, J., Zhang, Y., Gu, S., et al., 2021. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 6695–6714.
- Maier, O., Menze, B.H., von der Gabelntz, J., et al., 2017. Isles 2015 - a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri. *Medical Image Analysis* 35, 250–269.
- Mehrtash, A., Wells, W.M., Tempany, C.M., et al., 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., et al., 2015. 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., et al., 2015. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging* 34, 1993–2024.
- Mukhoti, J., Kulharia, V., Sanyal, A., et al., 2020. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems* 33, 15288–15299.
- Müller, R., Kornblith, S., and Hinton, G.E., 2019. When does label smoothing help? *Advances in neural information processing systems* 32.
- Murugesan, B., Adiga Vasudeva, S., Liu, B., et al., 2023a. Trust your neighbours: Penalty-based constraints for model calibration, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 572–581.
- Murugesan, B., Liu, B., Galdran, A., et al., 2023b. Calibrating segmentation networks with margin-based label smoothing. *Medical Image Analysis* 87, 102826.
- Naeini, M.P., Cooper, G., and Hauskrecht, M., 2015. Obtaining well calibrated probabilities using bayesian binning, in: *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- Niculescu-Mizil, A. and 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., et al., 2018. Attention u-net: Learning where to look for the pancreas, in: *Medical Imaging with Deep Learning*.
- Ovadia, Y., Fertig, E., Ren, J., et al., 2019. Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems* 32.
- Pereyra, G., Tucker, G., Chorowski, J., et al., 2017. Regularizing neural networks by penalizing confident output distributions, in: *International Conference on Learning Representations (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., and 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., et al., 2016. Rethinking the inception architecture for computer vision, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826.
- Tomani, C., Gruber, S., Erdem, M.E., et al., 2021. Post-hoc uncertainty calibration for domain drift scenarios, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10124–10132.
- Wang, G., Li, W., Aertsen, M., et al., 2019. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing* 338, 34–45.
- Zhang, J., Kailkhura, B., and Han, T.Y.J., 2020. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning, in: *International conference on machine learning*, PMLR. pp. 11117–11128.
- Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., and Liang, J., 2020. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Transactions on Medical Imaging* 39, 1856–1867.