Trust your neighbours:
Penalty-based constraints for model calibration

Balamurali Murugesan
ÉTS Montreal

Sukesh Adiga V
ÉTS Montreal

Bingyuan Liu
ÉTS Montreal

Hervé Lombaert
ÉTS Montreal

Ismail Ben Ayed
ÉTS Montreal

Jose Dolz
ÉTS Montreal

Abstract

Ensuring reliable confidence scores from deep networks is of pivotal importance in critical decision-making systems, notably in the medical domain. While recent literature on calibrating deep segmentation networks has led to significant progress, their uncertainty is usually modeled by leveraging the information of individual pixels, which disregards the local structure of the object of interest. In particular, only the recent Spatially Varying Label Smoothing (SVLS) approach addresses this issue 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 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 variety of well-known segmentation benchmarks demonstrate the superior performance of the proposed approach.

1 Introduction

Deep neural networks (DNNs) have achieved remarkable success in important areas of various domains, such as computer vision, machine learning and natural language processing. Nevertheless, there exists growing evidence that suggests that these models are poorly calibrated, leading to overconfident predictions that may assign high confidence to incorrect predictions [5, 6]. This represents a major problem, as inaccurate uncertainty estimates can have severe consequences in safety-critical applications such as medical diagnosis. The underlying cause of network miscalibration is hypothesized to be the high capacity of these models, which makes them susceptible to overfitting on the negative log-likelihood loss that is conventionally used during training [6].

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 [4, 6, 23, 24]. Despite its efficiency, this family of approaches presents important limitations, which include *i)* a dataset-dependency on the value of the transformation parameters and *ii)* a large degradation observed under distributional drifts [20]. A more principled solution integrates a term that penalizes confident output distributions into the learning objective, which explicitly maximizes the Shannon entropy of the model predictions during training [21]. Furthermore, findings from recent works on calibration [16, 17] have demonstrated that popular classification losses, such as Label Smoothing (LS) [22]

and Focal Loss (FL) [10], have a favorable effect on model calibration, as they implicitly integrate an entropy maximization objective. Following these works, [11, 18] presented a unified view of state-of-the-art calibration approaches [21, 22, 10] showing that these strategies can be viewed as approximations of a linear penalty imposing equality constraints on logit distances. The associated equality constraint results in gradients that continually push towards a non-informative solution, potentially hindering the ability to achieve the optimal balance between discriminative performance and model calibration. To alleviate this limitation, [11, 18] proposed a simple and flexible alternative based on inequality constraints, which imposes a controllable margin on logit distances. Despite the progress brought by these methods, none of them explicitly considers pixel relationships, which is fundamental in the context of image segmentation.

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. In particular, 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. To address this important issue, Spatially Varying Label Smoothing (SVLS) [7] 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.

The **contributions** of this work can be summarized as follows:

- We provide a constrained-optimization perspective of Spatially Varying Label Smoothing (SVLS) [7], demonstrating that it imposes an implicit constraint on 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 our 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.

- Comprehensive experiments over multiple medical image segmentation benchmarks, including diverse targets and modalities, show the superiority of our method compared to state-of-the-art calibration losses.

2 Methodology

Formulation. Let us denote the training dataset as $\mathcal{D}(\mathcal{X}, \mathcal{Y}) = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$, with $\mathbf{x}^{(n)} \in \mathcal{X} \subset \mathbb{R}^{\Omega_n}$ representing the nth image, Ω_n the spatial image domain, and $\mathbf{y}^{(n)} \in \mathcal{Y} \subset \mathbb{R}^{K}$ its corresponding ground-truth label with K classes, provided as a one-hot encoding vector. Given an input image $\mathbf{x}^{(n)}$, a neural network parameterized by θ generates a softmax probability vector, defined as $f_\theta(\mathbf{x}^{(n)}) = \mathbf{s}^{(n)} \in \mathbb{R}^{\Omega_n \times K}$, where $\mathbf{s}$ is obtained after applying the softmax function over the logits $\mathbf{l}^{(n)} \in \mathbb{R}^{\Omega_n \times K}$. To simplify the notations, we omit sample indices, as this does not lead to any ambiguity.

2.1 A constrained optimization perspective of SVLS

Spatially Varying Label Smoothing (SVLS) [7] 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 $\mathbf{y}^1$. Furthermore, the predicted softmax in a given pixel is denoted as $\mathbf{s} = [s_0, s_1, ..., s_{k-1}]$.

---

[1]For 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., $d_1 \times d_2$.

Let us now transform the surrounding patch of the segmentation mask around a given pixel into a unidimensional vector $\mathbf{y} \in \mathbb{R}^{\mathbf{d}}$, 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 $\mathbf{w} \in \mathbb{R}^{\mathbf{d}}$. Following this, for a given pixel p, and a class k, SVLS [7] can be defined as:

$$\tilde{y}_p^k = \frac{1}{|\sum_i^d w_i|} \sum_{i=1}^d y_i^k w_i. \tag{1}$$

Thus, once we replace the smoothed labels $\tilde{y}_p^k$ in the standard cross-entropy (CE) loss, the new learning objective becomes:

$$\mathcal{L} = -\sum_k \left(\frac{1}{|\sum_i^d w_i|} \sum_{i=1}^d y_i^k w_i \right) \log s_p^k, \tag{2}$$

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

$$\mathcal{L} = -\frac{1}{|\sum_i^d w_i|} \sum_k y_p^k \log s_p^k - \frac{1}{|\sum_i^d w_i|} \sum_k \left(\sum_{\substack{i=1 \\ i \neq p}}^d y_i^k w_i \right) \log s_p^k, \tag{3}$$

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 $\tau_k = \sum_{\substack{i=1 \\ i \neq p}}^d y_i^k w_i$ 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. 3, and removing $|\sum_i^d w_i|$ as it multiplies both terms, the loss becomes:

$$\mathcal{L} = \underbrace{-\sum_k y_p^k \log s_p^k}_{CE} \underbrace{-\sum_k \tau_k \log s_p^k}_{\text{Constraint on } \boldsymbol{\tau}}. \tag{4}$$

As $\boldsymbol{\tau}$ is constant, the second term in Eq. 4 can be replaced by a Kullback-Leibler (KL) divergence, leading to the following learning objective:

$$\mathcal{L} \stackrel{c}{=} \mathcal{L}_{CE} + \mathcal{D}_{KL}(\boldsymbol{\tau} || \mathbf{s}), \tag{5}$$

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 p, while imposing the equality constraint $\boldsymbol{\tau} = \mathbf{s}$, where $\boldsymbol{\tau}$ depends on the class distribution of surrounding pixels. Indeed, this term implicitly enforces the softmax predictions to match the soft-class proportions computed around p.

2.2 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 $\boldsymbol{\tau}$ 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:

$$\min \quad \mathcal{L}_{CE} \quad \text{s.t.} \quad \boldsymbol{\tau} = \mathbf{1}, \tag{6}$$

where $\boldsymbol{\tau}$ now represents a desirable prior, and $\boldsymbol{\tau} = \mathbf{1}$ is a hard constraint. Note that the reasoning behind working directly on the logit space is two-fold. First, observations in [11] 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 values[2], their magnitudes are further decreased, which has a favorable effect on model calibration [17]. We stress that despite both [11] and our method enforce constraints on the predicted logits, [11] is fundamentally different. In particular, [11] imposes an *inequality* constraint on the logit distances so that it encourages uniform-alike 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. 6 could be solved by a standard Lagrangian-multiplier algorithm, we replace the hard constraint by a soft penalty of the form $\mathcal{P}(|\boldsymbol{\tau} - \mathbf{1}|)$, 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 $\mathcal{P}(|\boldsymbol{\tau} - \mathbf{1}|) \geq \mathcal{P}(\mathbf{0}), \forall \boldsymbol{l} \in \mathbb{R}^K$, i.e., when the constraint is satisfied. Following this, when the constraint $|\boldsymbol{\tau} - \mathbf{1}|$ deviates from $\mathbf{0}$ the value of the penalty term increases. Thus, we can approximate the problem in Eq. 6 as the following simpler unconstrained problem:

$$\min \quad \mathcal{L}_{CE} + \lambda \sum_k \max(0, |\tau_k - l_k|), \tag{7}$$

where the penalty is modeled here as a ReLU function, whose importance is controlled by the hyperparameter λ.

3 Experiments

3.1 Setup

Datasets. FLARE Challenge [12] contains 360 volumes of multi-organ abdomen CT including liver, kidneys, spleen and pancreas with their corresponding pixel-wise masks. Volumes are resampled to a common space and cropped to $192{\times}192{\times}30$, and then randomly split into training (240), validation (40) and testing (80). **ACDC Challenge** [3] consists of 100 patient exams containing cardiac MR volumes and their respective segmentation masks, which include the left and right ventricles and the myocardium. Following the standard practices on this dataset, 2D slices are extracted from the volumes and resized to $224{\times}224$. The dataset is randomly split into training (70), validation (10) and testing (20) sets. **BraTS-19 Challenge** [15, 1, 2] contains 335 multi-modal MR scans (FLAIR, T1, T1-contrast, and T2) with their corresponding segmentation masks, which include tumor core, enhancing tumor and whole tumor. Each volume has a dimension of $155{\times}240{\times}240$ voxels and is resampled $128{\times}192{\times}192$. Furthermore, slices containing only background are removed from the training. Last, the volumes are randomly split into training (235), validation (35), and testing (65).

Evaluation metrics. To assess the discriminative performance of the evaluated models, we resort to standard segmentation metrics in medical segmentation, which includes the DICE coefficient (DSC) and the 95% Hausdorff Distance (HD). To evaluate the calibration performance, we employ the expected calibration error (ECE) [19] on foreground classes, as in [7], and classwise expected calibration error (CECE) [9], following [16, 18]. The reason behind including the CECE is because ECE only considers the softmax probability of the predicted class, ignoring the other scores in the softmax distribution [16]. The ECE can be approximated as a weighted average of the absolute difference between the accuracy and confidence of each bin:

---

[2]Note that the proportion priors are generally normalized.

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

where M denotes the number of equispaced bins, B_i denote the set of samples with confidences belonging to the ith bin, A_i is the accuracy of the i-th bin, and it 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, $\hat{y}_j$, 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 simple classwise extension of the ECE (referred to as CECE) metric is defined as:

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

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

Implementation details. We benchmark the proposed model against several losses, including state-of-the-art calibration losses. These models include the compounded CE + Dice loss (CE+DSC), FL [10], Entropy penalty (ECP) [21], LS [22], SVLS [7] and MbLS [11]. Following the literature, we consider the hyperparameters values typically employed and select the value which provided the best average DSC on the validation set across all the datasets. More concretely, for FL, γ values of 1, 2, and 3 are considered, whereas 0.1, 0.2, and 0.3 are used for α and λ in LS and ECP, respectively. We consider the margins of MbLS to be 3, 5, and 10, while fixing λ to 0.1, as in [18]. In the case of SVLS, the one-hot label smoothing is performed with a kernel size of 3 and $\sigma = [0.5, 1, 2]$. For training, we fixed the batch size to 16, epochs to 100, and used ADAM [8], with a learning rate of 10^{-3} for the first 50 epochs, and reduced to 10^{-4} afterwards. Following [18], the models are trained on 2D slices, and the evaluation is performed over 3D volumes. Last, we use the following prior $\tau_k = \sum_{i=1}^{d} y_i^k$, which is computed over a 3×3 patch, similarly to SVLS.

3.2 Results

Comparison to state-of-the-art. Table 1 reports the discriminative and calibration results achieved by the different methods. We can observe that, across all the datasets, the proposed method consistently outperforms existing approaches, always ranking first and second in all the metrics. Furthermore, while other methods may obtain better performance than the proposed approach in a single metric, their superiority strongly depends on the selected dataset. For example, ECP [21] yields very competitive performance on the FLARE dataset, whereas it ranks among the worst models in ACDC or BraTS.

	FLARE				ACDC				BraTS			
	DSC	HD	ECE	CECE	DSC	HD	ECE	CECE	DSC	HD	ECE	CECE
CE+DSC ($\lambda = 1$)	0.846	5.54	0.058	0.034	<u>0.828</u>	3.14	0.137	0.084	0.777	**<u>6.96</u>**	0.178	0.122
FL [10] ($\gamma = 3$)	0.834	6.65	0.053	0.059	0.620	7.30	0.153	0.179	**<u>0.848</u>**	9.00	**0.097**	0.119
ECP [21] ($\lambda = 0.1$)	**<u>0.860</u>**	**<u>5.30</u>**	**<u>0.037</u>**	**<u>0.027</u>**	0.782	4.44	0.130	0.094	0.808	8.71	0.138	0.099
LS [22] ($\alpha = 0.1$)	**<u>0.860</u>**	5.33	0.055	0.049	0.809	3.30	**<u>0.083</u>**	0.093	0.820	7.78	**<u>0.112</u>**	0.108
SVLS [7] ($\sigma = 2$)	0.857	5.72	0.039	0.036	0.824	**<u>2.81</u>**	0.091	0.083	0.801	8.44	0.146	0.111
MbLS [11] ($m=5$)	0.836	5.75	0.046	0.041	0.827	2.99	0.103	**<u>0.081</u>**	0.838	7.94	0.127	**0.095**
Ours ($\lambda = 0.1$)	**0.868**	**4.88**	**0.033**	**<u>0.031</u>**	**0.854**	**2.55**	**0.048**	**0.061**	**0.850**	**5.78**	**<u>0.112</u>**	**<u>0.097</u>**

Table 1: **Comparison to state-of-the-art.** Discriminative (DSC ↑, HD ↓) and calibration (ECE ↓, CECE ↓) performance obtained by the different models (best method in bold, and second best in bold and underlined).

To have a better overview of the performance of the different methods, we follow the evaluation strategies adopted in several MICCAI Challenges, i.e., *sum-rank* [14] and *mean-case-rank* [13]. In the *sum-rank* strategy, the final ranking is given as the sum of individual ranking metrics: r_{mc}, where r_{mc} $R_T = \sum_{m=0}^{|M|} r_m$, where r_m is the rank of the segmentation model for the metric m (mean). On
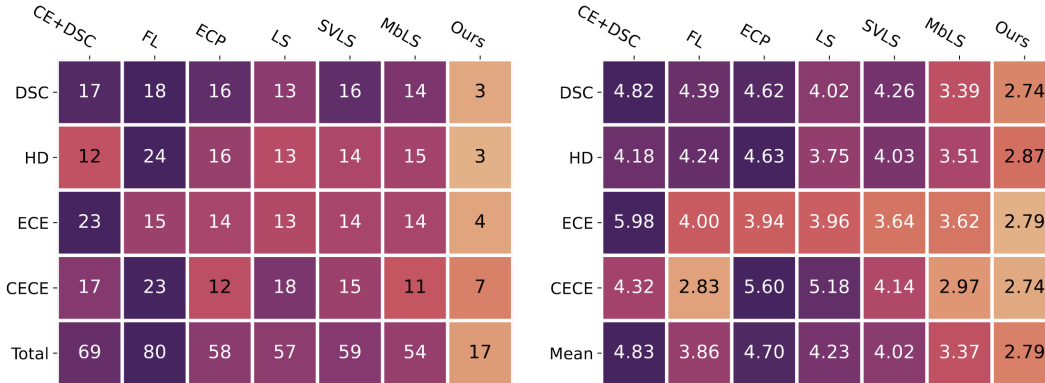
Figure 1: **Sum-rank and mean-rank evaluation**. Ranking of the different methods based on the sum-rank (*left*) and mean of case-specific (*right*) approaches. The lower the value, the better the performance.

the other hand, to evaluate the different methods following the *mean-case-rank* alternative, we first compute the DSC, HD, ECE, and CECE values for each sample, and compute the rank for each method based on these metrics, separately for each case. Then, we compute the mean rank over all the evaluation metrics, in a per-case manner, in order to obtain the method's rank for that given sample. Last, the mean over all case-specific ranks is computed to obtain the final rank of each method. As we can observe in the heatmaps provided in Fig. 1, 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.

Ablation studies. 1-Constraint over logits *vs* softmax. Recent evidence [11] suggests that imposing constraints on the logits presents a better alternative than its softmax counterpart. To demonstrate that this observation holds in our model, we present the results of our formulation when the constraint is enforced on the softmax distributions, i.e., replacing $\mathbf{l}$ by $\mathbf{s}$ (Table 2, *top*), which yields inferior results. **2-Choice of the penalty.** To solve the unconstrained problem in Eq. 7, we can approximate the second term with a liner penalty, modeled as a ReLU function. Nevertheless, we can resort to other polynomial penalties, e.g., quadratic penalties, whose main difference stems from the more aggressive behavior of quadratic penalties over larger constraint violations. The results obtained when the linear penalty is replaced by a quadratic penalty are reported in Table 2 (*middle*). From these results, we can observe that, while a quadratic penalty could achieve better results in a particular dataset (e.g., ACDC or calibration performance on BraTS), a linear penalty yields more consistent results across datasets. **3-Patch size.** For a fair comparison with SVLS, we used a patch of size 3×3 in our model. Nevertheless, we now investigate the impact of employing a larger patch to define the prior $\boldsymbol{\tau}$, whose results are presented in Table 2 (*bottom*). Even though a larger patch seems to bring comparable results in one dataset, the performance on the other two datasets is largely degraded, which potentially hinders its scalability to other applications. We believe that this is due to the higher degree of noise in the class distribution, particularly when multiple organs overlap, as the employed patch covers a wider region.

	FLARE				ACDC				BraTS			
	DSC	HD	ECE	CECE	DSC	HD	ECE	CECE	DSC	HD	ECE	CECE
Constraint on $\mathbf{s}$	0.862	5.14	0.043	0.030	0.840	2.66	0.068	0.071	0.802	8.28	0.145	0.104
L2-penalty	0.851	5.48	0.065	0.054	0.871	1.78	0.059	0.080	0.851	7.90	0.078	0.091
Patch size: 5×5	0.875	5.96	0.032	0.031	0.813	3.50	0.078	0.077	0.735	7.45	0.119	0.092

Table 2: Empirical results to motivate our methodological and technical choices.

Impact of the prior. A benefit of the proposed formulation , particularly compared to SVLS [7], is that diverse priors can be enforced on the logit distributions. Thus, we now assess the impact of different priors $\boldsymbol{\tau}$ in our formulation (See Supplemental Material for a detailed explanation of each alternative to compute the prior). The results presented in Table 3 reveal that selecting a suitable prior can further improve the performance of our model.

Prior τ	FLARE				ACDC				BraTS			
	DSC	HD	ECE	CECE	DSC	HD	ECE	CECE	DSC	HD	ECE	CECE
Mean	**0.868**	**4.88**	**0.033**	**0.031**	0.854	2.55	0.048	0.061	**0.850**	**5.78**	0.112	0.097
Gaussian	0.860	5.40	**0.033**	0.032	0.876	2.92	0.042	0.053	0.813	7.01	0.140	0.106
Max	0.859	4.95	0.038	0.036	0.876	1.74	0.046	0.054	0.833	8.25	0.114	0.094
Min	0.854	5.42	0.034	0.033	0.881	1.80	0.040	0.053	0.836	7.23	0.104	0.092
Median	0.867	5.90	**0.033**	0.032	0.835	3.29	0.075	0.075	0.837	7.53	**0.095**	**0.089**
Mode	0.854	5.41	0.035	0.034	0.876	**1.62**	0.045	0.056	0.808	8.21	0.135	0.113

Table 3: Impact of using different priors (τ) in Eq. 7.

Magnitude of the logits. To empirically demonstrate that the proposed solution decreases the logit values, we plot average logit distributions across classes on the FLARE test set (Fig. 2). 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 (in absolute value). We can observe that, compared to SVLS and MbLS, –which also imposes constraints on the logits–, our approach leads to much lower logit values, particularly compared to SVLS.
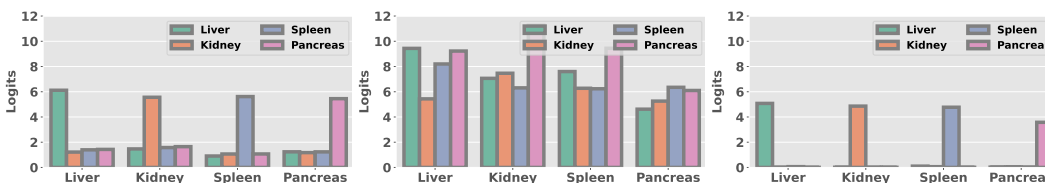

Figure 2: **Distribution of logit values.** From left to right: MbLS, SVLS and ours.

4 Conclusion

We have presented a constrained-optimization perspective of SVLS, which has revealed two important limitations of this method. First, the implicit constraint enforced by SVLS cannot be controlled explicitly. And second, the prior imposed in the constraint is directly derived from the Gaussian kernel used, which makes it hard to model. In light of these observations, we have proposed a simple alternative based on equality constraints on the logits, which allows to control the importance of the penalty explicitly, and the inclusion of any desirable prior in the constraint. Our results suggest that the proposed method improves the quality of the uncertainty estimates, while enhancing the segmentation performance.

References

[1] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.

[2] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. 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*, 2018.

[3] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE TMI*, 37(11):2514–2525, 2018.

[4] Zhipeng Ding, Xu Han, Peirong Liu, and Marc Niethammer. Local temperature scaling for probability calibration. In *ICCV*, 2021.

[5] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.

[6] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

[7] Mobarakol Islam and Ben Glocker. Spatially varying label smoothing: Capturing uncertainty from expert annotations. In *International Conference on Information Processing in Medical Imaging*, pages 677–688. Springer, 2021.

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

[9] Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *NeurIPS*, 32, 2019.

[10] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *CVPR*, 2017.

[11] Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. The devil is in the margin: Margin-based label smoothing for network calibration. In *CVPR*, 2022.

[12] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, Shucheng Cao, Qi Zhang, Shangqing Liu, Yunpeng Wang, Yuhui Li, Jian He, and Xiaoping Yang. Abdomenct-1K: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[13] Oskar Maier et al. ISLES 2015 - a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. *Medical Image Analysis*, 35:250–269, 2017.

[14] Adriënne M Mendrik, Koen L Vincken, Hugo J Kuijf, Marcel Breeuwer, Willem H Bouvy, Jeroen De Bresser, Amir Alansary, Marleen De Bruijne, Aaron Carass, Ayman El-Baz, et al. MRBrainS challenge: online evaluation framework for brain image segmentation in 3T MRI scans. *Computational intelligence and neuroscience*, 2015.

[15] Bjoern H. Menze et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015.

[16] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip HS Torr, and Puneet K Dokania. Calibrating deep neural networks using focal loss. In *NeurIPS*, 2020.

[17] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? In *NeurIPS*, 2019.

[18] Balamurali Murugesan, Bingyuan Liu, Adrian Galdran, Ismail Ben Ayed, and Jose Dolz. Calibrating segmentation networks with margin-based label smoothing. *arXiv preprint arXiv:2209.09641*, 2022.

[19] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[20] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *NeurIPS*, 2019.

[21] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. In *ICLR*, 2017.

[22] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[23] Christian Tomani, Sebastian Gruber, Muhammed Ebrar Erdem, Daniel Cremers, and Florian Buettner. Post-hoc uncertainty calibration for domain drift scenarios. In *CVPR*, 2021.

[24] Jize Zhang, Bhavya Kailkhura, and T Han. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *ICML*, 2020.

A Further details on the different priors

For a given pixel p with, and a surrounding patch of size $d = d_1 \times d_2$ (note that in this case $\mathbf{y} \in \{0,1\}^{d \times K}$ becomes the label assignment on the patch), the following equations can be used to obtain the different priors employed in Table 3 of the main manuscript:

$$\text{Mean:} \quad \tau_k = \frac{1}{|d|} \sum_{i=1}^{d} y_i^k \tag{10}$$

$$\text{Gaussian:} \quad \tau_k = \frac{1}{|\sum_i^d w_i|} \sum_{i=1}^{d} y_i^k w_i, \tag{11}$$

where w_i is the Gaussian kernel with $\sigma = 2.0$ in our experiments, similarly to the value we employed for SVLS [7].

Now, let $\mathbf{y}^* \in \{0, ..., K-1\}^d$ be the label encoding derived from the one-hot labels $\mathbf{y}$. The following priors can be therefore be defined as:

$$\text{Max:} \quad \tau_k = \begin{cases} 1, & \max_d(\mathbf{y}^*) == k \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

$$\text{Min:} \quad \tau_k = \begin{cases} 1, & \min_d(\mathbf{y}^*) == k \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

$$\text{Mode:} \quad \tau_k = \begin{cases} 1, & \text{mode}_d(\mathbf{y}^*) == k \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

$$\text{Median:} \quad \tau_k = \begin{cases} 1, & \text{median}_d(\mathbf{y}^*) == k \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$