Constraint-Based Calibration for Reliable Deep Learning Models

by

Balamurali MURUGESAN

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, NOVEMBER 7, 2025

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Jose Dolz, Thesis supervisor
Department of Software and IT Engineering, École de technologie supérieure

Mr. Ismail Ben Ayed, Thesis Co-Supervisor
Department of System Engineering, École de technologie supérieure

Mr. Matthew Toews, Chair, Board of Examiners
Department of System Engineering, École de technologie supérieure

Mr. Christian Desrosiers, Member of the Jury
Department of Software and IT Engineering, École de technologie supérieure

Ms. Ipek Oguz, External Independent Examiner
Department of Electrical and Computer Engineering, Vanderbilt University

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON OCTOBER 29, 2025

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

This thesis would not have been possible without the invaluable guidance of my supervisors, Dr. Jose Dolz and Dr. Ismail Ben Ayed. I am deeply grateful for their unwavering support and the opportunity they gave me to pursue my graduate studies.

I want to extend my sincere thanks to my committee members, Dr. Christian Desrosiers and Dr. Matthew Toews, for their constructive feedback and timely support. I also want to acknowledge the faculty at ÉTS Montréal for creating an inspiring and supportive academic environment throughout my time here.

I also thank Adrian Galdran, Hervé Lombaert, Riadh Kobbi, and Hadi Chakor for sharing their expertise and providing feedback whenever I needed it. I deeply appreciate my research collaborators—Bingyuan Liu, Julio Silva-Rodríguez, Sukesh Adiga V, Rajarshi Bhattacharya, and Rukhshanda Hussain—whose generous assistance and insightful discussions greatly enhanced the quality of this work. I am particularly grateful to Gnana Praveen, Raghav Mehta, and Akhil Meethal for introducing me to various aspects of research and providing key insights on preparing research manuscripts.

I would like to thank my mentors from my internships at Nuance (Erik Larsson, Aman Gokrani, and Nate Bodenstab) and Amazon (Bingyuan Liu, Lei Chen, Edouard Belval, Elad Ben Avraham, Oren Nuriel, and Thomas Delteil).

To my friends, thank you for your knowledge-sharing, encouragement, and camaraderie. I am especially grateful to Saypraseuth Mounsaveng, Heitor Rapela Medeiros, Sina Hajimiri, Farzad Beizaee, Fereshteh Shakeri, Masih Aminbeidokhti, Mélanie Gaillochet, Atif Belal, Kunal Mahatha, Rithu Krishna Sindhu, Shambhavi Mishra, Gustavo Vargas Hákim, David Osowiechi, Shakeeb Murtaza, and Karthik Gopinath for making this journey a joyful one. I also thank my longtime friends Kaushik Sarveswaran, Koushik Srivatsan, Mohankumar Sivasankaran, and Shyam Ayyasamy for staying connected virtually.

Finally, I want to express my heartfelt gratitude to my family. To my wife, Maheswari, thank you for your unwavering belief, patience, understanding, and endless encouragement, especially during the most challenging times. And to our newborn, Geethamagizhan, thank you for the new perspective and joy you brought into my life during this demanding period. I am also thankful to my extended family for their timely help and support. A special thanks to my brother, Badrinath, for always being there for our parents, allowing me to focus on my work without worry. Finally, I owe everything to my parents, Murugesan and Valli, whose sacrifices and boundless love have provided me with countless opportunities. I am forever grateful.

Calibrage de modèles en apprentissage profond

Balamurali MURUGESAN

RÉSUMÉ

Malgré les progrès indéniables réalisés dans les tâches de reconnaissance visuelle grâce aux réseaux de neurones profonds, des données récentes montrent que ces modèles sont mal calibrés, ce qui entraîne des prédictions trop fiables. Les pratiques standard de minimisation de la perte d'entropie croisée pendant l'apprentissage favorisent la correspondance des probabilités softmax prédites avec les attributions d'étiquettes uniques. Néanmoins, cela produit une activation pré-softmax de la classe correcte nettement supérieure aux activations restantes, ce qui aggrave le problème de mauvais calibrage. Des observations récentes issues de la littérature sur la classification suggèrent que les fonctions de perte intégrant une maximisation implicite ou explicite de l'entropie des prédictions offrent des performances de calibrage de pointe. Malgré ces résultats, l'impact de ces pertes sur la tâche pertinente de calibrage des réseaux de segmentation d'images médicales, les nouvelles pertes spécifiques à la tâche de segmentation et le langage visuel reste inexploré.

Dans le premier objectif, nous nous référons à l'un des travaux antérieurs de notre groupe, qui propose une perspective unifiée d'optimisation sous contraintes des pertes de calibrage de pointe actuelles. Plus précisément, ces pertes sont considérées comme des approximations d'une pénalité linéaire (ou d'un terme lagrangien) imposant des contraintes d'égalité sur les distances logit. Cela met en évidence une limitation importante de ces contraintes d'égalité strictes sous-jacentes, dont les gradients qui en résultent poussent constamment vers une solution non informative, ce qui pourrait empêcher d'atteindre le meilleur compromis entre performance discriminante et calibration du modèle lors de l'optimisation par gradient. Suite à nos observations, nous étendons à la segmentation d'images médicales la pénalité de généralisation simple et flexible proposée, qui impose une marge contrôlable sur les distances logit. Nous fournissons des expériences et des études d'ablation complètes sur cinq benchmarks de segmentation publics différents, axés sur diverses cibles et modalités, soulignant les capacités de généralisation de l'approche proposée. Nos résultats empiriques démontrent la supériorité de notre méthode par rapport aux pertes de calibration de pointe, tant en termes de calibration que de performance discriminante.

Dans le deuxième objectif, nous proposons une perspective d'optimisation sous contrainte du lissage spatial des étiquettes variables (SVLS), démontrant qu'il peut être considéré comme une perte d'entropie croisée standard associée à une contrainte implicite imposant aux prédictions softmax de correspondre à une proportion de classe souple des pixels environnants. Notre formulation montre que le SVLS ne dispose pas d'un mécanisme permettant de contrôler explicitement l'importance de la contrainte, ce qui peut entraver le processus d'optimisation, car il devient difficile d'équilibrer efficacement la contrainte avec l'objectif principal. Suite à ces observations, nous proposons une solution simple et flexible basée sur des contraintes d'égalité sur les distributions logit. La contrainte proposée est appliquée par une pénalité

linéaire simple, qui intègre un mécanisme explicite pour contrôler son poids. Notre approche offre non seulement une stratégie plus efficace pour modéliser les distributions logit, mais diminue également implicitement les valeurs logit, ce qui se traduit par des prédictions moins surconfiantes. Nous menons des expériences approfondies et des études d'ablation sur plusieurs benchmarks de segmentation d'images médicales, incluant diverses cibles et modalités, et démontrons la supériorité de notre méthode par rapport aux pertes d'étalonnage les plus récentes. De plus, plusieurs études d'ablation valident empiriquement les choix de conception de notre approche et démontrent son caractère agnostique vis-à-vis du modèle.

Dans le troisième objectif, nous proposons une approche par contraintes par classe et par région pour résoudre le problème d'étalonnage erroné dans les modèles de segmentation sémantique. Plus précisément, nous formulons une solution qui prend en compte les spécificités de chaque catégorie et des différentes régions en introduisant des pondérations de pénalité indépendantes par classe et par région. Ceci contraste avec les travaux antérieurs, où une pondération de pénalité scalaire uniforme est utilisée, quelles que soient les catégories ou les régions. De plus, nous transposons le problème contraint à son homologue d'optimisation duale sans contrainte en utilisant une méthode lagrangienne augmentée (ALM). Cela évite d'ajuster manuellement chaque pondération de pénalité et permet, grâce à une série d'étapes itératives internes et externes, de trouver la valeur optimale de chaque pondération de pénalité, laquelle peut être apprise de manière adaptative. Des expériences approfondies sur deux benchmarks de segmentation populaires et deux structures de segmentation bien connues démontrent la supériorité de notre approche par rapport à un ensemble d'approches d'étalonnage récentes et pertinentes.

Dans le quatrième objectif, nous démontrons empiriquement que les stratégies d'adaptation CLIP courantes, telles que les adaptateurs, l'apprentissage rapide et le réglage rapide au moment du test, dégradent considérablement les capacités de calibrage de la ligne de base zéro-shot en présence de dérive distributionnelle. Pour ces stratégies d'adaptation, nous démontrons que la cause sous-jacente du mauvais calibrage est en fait l'augmentation des plages logit. Cela contraste avec les travaux récents sur le calibrage des modèles entièrement supervisés, qui suggèrent que la cause inhérente du mauvais calibrage est plutôt l'augmentation de sa norme, due à la perte d'entropie croisée standard utilisée pour l'apprentissage. Sur la base de ces observations, nous présentons une solution simple et indépendante du modèle, qui consiste à mettre à l'échelle la plage logit de chaque échantillon en fonction des logits zéro-shot. Nous présentons également plusieurs alternatives pour adapter notre solution, qui peuvent être mises en œuvre au moment de l'apprentissage ou de l'inférence. Des expériences approfondies sur des référentiels de classification OOD courants démontrent empiriquement l'efficacité de nos approches pour réduire l'erreur de calibrage, tout en conservant les performances discriminantes.

Mots-clés: Calibrage de réseau, segmentation d'image, incertitude, modèles vision-langage, adaptation à quelques prises de vue, généralisation de domaine, adaptation au temps de test

Constraint-Based Calibration for Reliable Deep Learning Models

Balamurali MURUGESAN

ABSTRACT

Despite the undeniable progress in visual recognition tasks fueled by deep neural networks, there exists recent evidence showing that these models are poorly calibrated, resulting in over-confident predictions. The standard practices of minimizing the cross-entropy loss during training promote the predicted softmax probabilities to match the one-hot label assignments. Nevertheless, this yields a pre-softmax activation of the correct class that is significantly larger than the remaining activations, which exacerbates the miscalibration problem. Recent observations from the classification literature suggest that loss functions that embed implicit or explicit maximization of the entropy of predictions yield state-of-the-art calibration performances. Despite these findings, the impact of these losses in the relevant task of calibrating medical image segmentation networks, novel losses specific to task of segmentation, and vision-language remains unexplored.

In the first objective, we refer to one of the earlier works from our group which provides a unifying constrained-optimization perspective of current state-of-the-art calibration losses. Specifically, these losses are viewed as approximations of a linear penalty (or a Lagrangian term) imposing equality constraints on logit distances. This points to an important limitation of such underlying hard equality constraints, whose ensuing gradients constantly push towards a non-informative solution, which might prevent from reaching the best compromise between the discriminative performance and calibration of the model during gradient-based optimization. Following these insights, we extend the proposed simple and flexible generalization penalty which imposes a controllable margin on logit distances to medical image segmentation. We provide comprehensive experiments and ablation studies on seven different public segmentation benchmarks that focus on diverse targets and modalities, highlighting the generalization capabilities of the proposed approach. Our empirical results demonstrate the superiority of the margin based label smoothing compared to state-of-the-art calibration losses in both calibration and discriminative performance.

In the second objective, we provide a constrained-optimization perspective of Spatially Varying Label Smoothing (SVLS), 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.

In the third objective, we propose a class and region-wise constraint approach to tackle the miscalibration issue in semantic segmentation models. In particular, we formulate a solution that considers the specificities of each category and different regions by introducing independent class and region-wise penalty weights. This contrasts with the prior work, where a uniform scalar penalty weight is employed, regardless of categories or regions. Furthermore, we transfer the constrained problem to its dual unconstrained optimization counterpart by using an Augmented Lagrangian method (ALM). This alleviates the need for manually adjusting each penalty weight and allows, through a series of iterative *inner* and *outer* steps, to find the optimal value of each penalty weight, which can be learned in an adaptive manner. Comprehensive experiments on two popular segmentation benchmarks, and with two well-known segmentation backbones, demonstrate the superiority of our approach over a set of relevant recent calibration approaches.

In the fourth objective, we empirically demonstrate that popular CLIP adaptation strategies, such as Adapters, Prompt Learning, and Test-Time Prompt Tuning, substantially degrade the calibration capabilities of the zero-shot baseline in the presence of distributional drift. For these adaptation strategies, we expose that the underlying cause of miscalibration is, in fact, the increase of the logit ranges. This contrasts with recent work in calibrating fully-supervised models, which suggests that the inherent cause of miscalibration is the increase of its norm instead, due to the standard cross-entropy loss used for training. Based on these observations, we present a simple, and model-agnostic solution, which consists in scaling the logit range of each sample based on the zero-shot logits. We further present several alternatives to accommodate our solution, which can be implemented either at training or inference time. Comprehensive experiments on popular OOD classification benchmarks empirically demonstrate the effectiveness of our approaches to reduce the miscalibration error, while keeping the discriminative performance.

Keywords: Network calibration, Image Segmentation, Uncertainty, Vision-language models, Few-shot adaptation, Domain generalization, Test-time adaptation

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYNMS

ETS	École de Technologie Supérieure
ASC	Agence Spatiale Canadienne
DSC	Dice Metric for Segmentation
HD	Hausdorff Distance for Segmentation
ECE	Expected Calibration Error
CECE	Class-Wise Expected Calibration Error
ACE	Adaptive Calibration Error
TACE	Threshold Adaptive Calibration Error
FL	Focal Loss
LS	Label Smoothing
BWCR	Boundary-weighted logit consistency Regularizer
SVLS	Spatial Varying Label Smoothing
ECP	Explicit confidence penalty
MbLS	Margin based label smoothing
NACL	Neighbor Aware CaLibration
CRaC	Class and Region-Adaptive Constraints
ZS-NORM	Zero-Shot Logit Normalization
ZS-PEN	Zero-Shot Penalty
SaLS	Sample-adaptive Logit Scaling

ACDC	Automated Cardiac Diagnosis Challenge
FLARE	Fast and Low GPU memory Abdominal Organ Segmentation
BraTS	Brain Tumor Segmentation (BRATS) 2019 Challenge
PROMISE	Prostate Cancer Dataset
KiTS	Kidney Tumor Segmentation Challenge
MRBrainS	Medical Image Computing and Computer Assisted Intervention Brain Segmentation Challenge
CoOp	Context Optimization
CoCoOp	Conditional Context Optimization
ProGrad	Prompt-aligned Gradient
MaPLe	Multi-modal Prompt Learning
TPT	Test-time Prompt Tuning
C-TPT	Calibrated Test-time Prompt Tuning
KgCoOp	Knowledge-guided Conditional Context Optimization
CAM	Class Activation Map
MedIA	Medical Image Analysis
MICCAI	International Conference on Medical Image Computing and Computer-Assisted Intervention
ECCV	European Conference on Computer Vision
WACV	Winter Conference on Applications of Computer Vision

LIST OF SYMBOLS AND UNITS OF MEASUREMENTS

$\mathcal{D}$	Train, Valid, and Test Dataset
N	Number of Training Samples
$\mathcal{X}$	Input Data
$\mathcal{Y}$	Target Label
$\mathbf{l}$	Prediction Logits
$\hat{y}$	Predicted Label
s	Prediction softmax
$\hat{p}$	Prediction probability
$\mathbf{d}(\mathbf{l})$	Logit distance
Ω	Spatial Domain
K	Number of Classes
$\mathcal{L}$	Loss Function
$\mathbf{u}$	Uniform Distribution
$\mathbf{m}$	Margin
α	Hyper-Parameter in Label Smoothing
γ	Hyper-Parameter in Focal Loss
τ	Class distribution of surrounding pixels
$\mathcal{H}$	Entropy
λ	Balancing Factor

$\mathcal{I}$	Inner Region
$\mathcal{O}$	Outer Region
ρ	Penalty parameter
$\mathcal{D}_{\mathrm{KL}}$	KL Divergence
$\mathbf{w}$	Spatial Kernel
$d_1 \times d_2$	Patch Size
σ	Gaussian filter standard deviation
B	Batch Size
$\mathcal{T}$	Text Prompts
$\boldsymbol{t}$	CLIP text embedding
$\boldsymbol{z}$	CLIP vision embedding
$l_i^{\text{ZS-min}}$	min logit magnitudes of zero-shot
$l_i^{\text{ZS-max}}$	max logit magnitudes of zero-shot
l_i^{min}	min logit magnitudes of prediction
l_i^{max}	max logit magnitudes of prediction

INTRODUCTION

0.1 Motivation

Deep Neural Networks (DNNs) have not only achieved remarkable results but have consistently surpassed previous state-of-the-art methodologies across numerous demanding benchmarks (Krizhevsky, Sutskever & Hinton, 2012; Everingham *et al.*, 2015; Wang *et al.*, 2018; Rajpurkar, Zhang, Lopyrev & Liang, 2016), fundamentally reshaping the landscape of artificial intelligence and demonstrating an unparalleled capacity for learning intricate patterns from complex data. This paradigm shift is strikingly evident in fields ranging from computer vision (He, Zhang, Ren & Sun, 2016b; Ren, He, Girshick & Sun, 2015; Dosovitskiy *et al.*, 2021b), where DNNs have achieved human-level performance on tasks like image recognition and object detection, to natural language processing (Sutskever, Vinyals & Le, 2014; Vaswani *et al.*, 2017; Brown *et al.*, 2020), where they power sophisticated machine translation, and text generation, systems with unprecedented fluency and accuracy. Furthermore, their influence is felt in speech recognition (Panayotov, Chen, Povey & Khudanpur, 2015; Guoguo Chen,Shuzhou Chai, 2021), achieving remarkable reductions in error rates, and even in complex strategic domains like game playing (Silver *et al.*, 2016), previously thought to be the exclusive domain of human expertise. Critically, their impact extends profoundly into medical image analysis (Shin *et al.*, 2016; Ronneberger, Fischer & Brox, 2015b; Milletari, Navab & Ahmadi, 2016b; Gulshan *et al.*, 2016), enabling significant advancements in tasks such as automated detection of diseases (e.g., cancer, diabetic retinopathy), and precise segmentation of anatomical structures for surgical planning – often surpassing the accuracy and efficiency of traditional methods.

The deployment of these state-of-the-art deep learning networks is still in early stages in practical applications like medical diagnosis (Huang, Ruan, Xing & Feng, 2024), autonomous vehicle control (Loquercio, Segu & Scaramuzza, 2020), or financial risk assessment (Blasco, Sánchez & García, 2024), where besides accuracy, the uncertainty score associated with the

prediction plays a vital role in decision making. A high uncertainty score can flag potentially unreliable predictions, prompting human review, triggering fallback mechanisms, or informing a more cautious approach to the model's output (Kendall & Gal, 2017; Amini, Schwarting, Soleimany & Rus, 2020). For instance, the model might output a prediction of "malignant tumor" with an associated uncertainty score of 30%, which could arise because the image quality is poor, or the tumor has unusual characteristics not well-represented in the training dataset. This high uncertainty would flag the prediction as potentially unreliable, prompting a human radiologist to review the image and model's findings more carefully before making a diagnosis. Hence, it is necessary to understand the source of uncertainty, and define appropriate measures to quantify the reliability based on the intended application.

The predictive uncertainty (Hüllermeier & Waegeman, 2021) is in general separated into data uncertainty (also statistical or aleatoric uncertainty) and model uncertainty (also systemic or epistemic uncertainty), a crucial framework for understanding the limitations and potential pitfalls of any predictive model, particularly within the complex landscape of deep learning. Data uncertainty, often viewed as irreducible, stems from the inherent noise, randomness, and ambiguity present within the data itself. In contrast, model uncertainty, which is theoretically reducible, originates from the limitations of the model's structure, the training process, or the amount and representativeness of the training data. Disentangling these two sources of uncertainty, and understanding the type of uncertainty leads to targeted improvements (Kendall & Gal, 2017). Besides the above-mentioned types of uncertainty, deep learning models are very susceptible to distribution shifts (Liang, Li & Srikant, 2018), as it is impossible to cover all possible scenarios, hence modeling distributional uncertainty is mandatory in building trustworthy and interpretable systems.

In case of classification, the aleatoric uncertainty quantification is generally obtained from the softmax outputs (Van Amersfoort, Smith, Teh & Gal, 2020). The maximum confidence

associated with the final prediction can serve as the uncertainty metric, higher the softmax score, more confident is the model. A more holistic quantification would involve usage of entropy, where confident predictions have high entropy compared to uncertain – uniform softmax values. The epistemic uncertainty is captured through Bayesian approaches (Blundell, Cornebise, Kavukcuoglu & Wierstra, 2015), where the model intends to predict probability distributions. Besides, techniques like Monte Carlo Dropout (Gal & Ghahramani, 2016), and Deep Ensembles (Fort, Hu & Lakshminarayanan, 2019) are common approaches to obtain alternate distributions. The variance of these distributions quantify the uncertainty, where lower values mean certainty in the prediction. Though Bayesian approaches are better at modelling the uncertainty, it can be computationally expensive and challenging to scale to large deep learning models. In contrast, softmax analysis is simple and computationally efficient, as it comes directly from the standard output layer.

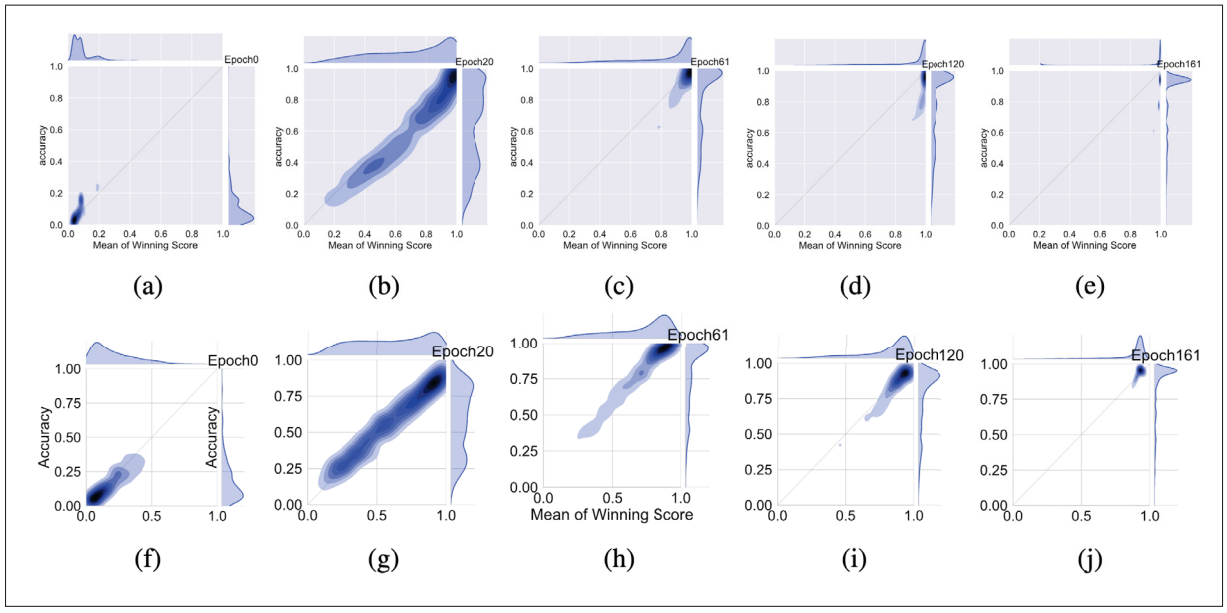

Figure 0.1 Joint density plots of accuracy vs confidence (captured by the mean of the winning softmax score) on the CIFAR-100 validation set at different training epochs for the VGG-16 deep neural network. **Top Row:** Cross Entropy, **Bottom Row:** MixUp

The predictive uncertainty obtained from maximum softmax score should be indicative of the actual likelihood of correctness. However, recent studies (Thulasidasan, Chennupati, Bilmes, Bhattacharya & Michalak, 2019b) have shown that many latest models are inherently overconfident due to their complex architectures and the standard loss functions that are designed to maximize accuracy. For instance, consider the case of VGG-16 network optimized for CIFAR-100 dataset, comparing the average winning score and accuracy for the validation data using joint density plots in the top row of the Figure 0.1 shows that confidence (captured by the winning score) as well as accuracy start out low and gradually increase as the network learns. However, it can be noted that confidence always surpasses accuracy in the later stages of training; accuracy saturates while confidence continues to increase resulting in a very sharply peaked distribution of winning scores and an overconfident model.

Most modern DNNs, when trained for classification in a supervised learning setting, are trained using one-hot encoded labels that have all the probability mass in one class (Müller, Kornblith & Hinton, 2019b); the training labels are thus zero-entropy signals that admit no uncertainty about the input. The DNN is thus, in some sense, trained to become overconfident. This inherent tendency toward overconfidence is handled through model calibration (Guo, Pleiss, Sun & Weinberger, 2017b; Wenger, Kjellström & Triebel, 2020), a procedure which ensures that the model's confidence score matches with the discriminative performance. The density plots comparing the average winning score and accuracy similar to the earlier discussed case are shown in the bottom row of the Figure 0.1, it can be seen that the calibrated model has predicted softmax scores which are better indicators of the actual likelihood of a correct prediction.

0.2 Reasons behind model miscalibration

It has been observed that some recent changes in modern neural networks are responsible for model miscalibration (Guo *et al.*, 2017b; Mukhoti *et al.*, 2020b; Minderer *et al.*, 2021). The underlying general cause is that modern neural networks' high capacity makes them vulnerable

to miscalibration, which is tightly correlated to the concepts of *over-parameter*, *overfitting* and *over-confidence*.

0.2.1 Model Complexity

The design and depth of deep models help in providing improved predictive performance, and it also impacts the model calibration. Studies have shown that over-parameterized models tend to push the prediction very close to the target label, an important tradeoff between generic and overfitted model (Guo *et al.*, 2017b). The recent pretrain-finetune paradigm which involves updating only few parameters offers the possibility of reducing overfitting (Desai & Durrett, 2020; Hu *et al.*, 2022b). The choice of parameters considered for optimizing can also help with calibration, like to avoid attention mechanisms updates, and focus only on the feedforward and normalization layers (Ye *et al.*, 2023). Moreover, Vision Transformers (ViT) and ConvNeXt have shown better calibration over Convolution Neural Networks (CNNs) (Liu *et al.*, 2022c; Bai, Mei, Yuille & Xie, 2021).

0.2.2 Data Issues

The data and annotations play a crucial role in developing deep learning models for the target task. Likewise, to obtain desirable calibration, the quantity and quality of data is key. Training over-parametrized network with less or incomplete data can result in overconfident model. The usage of noisy and biased data for training can also severely miscalibrate the model. For example, when dataset is imbalanced, meaning that few classes are significantly more prevalent than the other, the model tends to become overconfident for those few classes and uncertain about the remaining classes, leading to poor calibration. Ensemble methods (Lakshminarayanan, Pritzel & Blundell, 2017b; Malinin, Mlodozeniec & Gales, 2019) and data augmentation (Thulasidasan *et al.*, 2019b; Müller *et al.*, 2019b) have shown encouraging results in mitigating the effects of noisy data on model performance.

0.2.3 Learning Objective and Regularization

The loss functions are primarily chosen based on the target objective, and in most cases, it is to increase the performance of the model, which can impact the calibration quality. For instance, in case of classification tasks, the weights are optimized for the likelihood of correct class predictions. However, optimizing only for accuracy can lead to over-confident predictions, particularly in situations where learning the features of a particular class is relatively easier than the other. Regularization techniques have always been incorporated along with the target objective to avoid model overfitting aiding in model calibration. Lately, techniques likes batch normalization (Ioffe & Szegedy, 2015) and dropout (Srivastava, Hinton, Krizhevsky, Sutskever & Salakhutdinov, 2014) have been designed specifically for deep learning models to alleviate the over-confident issue. Focal loss (Lin, Goyal, Girshick, He & Dollár, 2017) has shown its effectiveness in handling the training progress for samples with different difficulty levels.

0.3 Calibration in image segmentation

The earlier discussed problems also applies to tasks beyond classification, like semantic segmentation - classifying every single pixel in an image with a corresponding class label. The direct solutions from classification literature have shown promising results, and there have been some preliminary attempts to utilize the spatial nature of segmentation tasks (Wang, Gong & Wang, 2023). Importantly, most of the existing methods have primarily focused on medical imaging domain which requires capturing the ambiguity between classes especially in the boundary regions (Kohl *et al.*, 2018). Stochastic segmentation networks (Monteiro *et al.*, 2020) captures correlations between pixels by modelling the logit map as a low-rank multivariate normal distribution. Deep Deterministic Uncertainty (Mukhoti, van Amersfoort, Torr & Gal, 2021) showed that feature space densities could also be used to estimate the

uncertainty. Local temperature scaling (Ding *et al.*, 2021) predicts a temperature for individual pixels based on the spatial variations. Model ensembles besides achieving better segmentation also helps in improving calibration (Mehrtash, Wells, Tempany, Abolmaesumi & Kapur, 2020). Inference from Variational U-Net (Fuchs, Gonzalez & Mukhopadhyay, 2021) showed that reliable predictive estimates could be obtained.

0.4 Domain shift calibration

Domain generalization is a key requirement for model deployment as predictions for unseen data could be unreliable (Torralba & Efros, 2011). Hence, it is necessary to have a model calibrated not only for in-distribution data, but also be reliable for out-of-distribution (unseen) data. Earlier, techniques based on unsupervised domain adaptation (Park, Bastani, Weimer & Lee, 2020) were used to predict the covariate shifts and correct for the shift through importance weighting. Further, to reduce the effective distribution disparity between the target and calibration domains, (Gong *et al.*, 2021) leverages developing models exposed to multiple domains. Post processing methods (Tomani, Gruber, Erdem, Cremers & Buettner, 2021b) after performing perturbations on the validation set have also been effective in calibration. Recently, pre-trained vision-language models like CLIP have shown to be inherently better in domain generalization. To further improve the performance for the target tasks, adaptation techniques are being considered, which disturbs the calibration for out-of-distribution samples.

0.5 Research Objectives and Contributions

In this thesis, we mainly focus on the miscalibration caused through learning objectives. The standard practices of minimizing the cross-entropy loss during training promote the predicted softmax probabilities to match the one-hot label assignments. Nevertheless, this yields a pre-softmax activation of the correct class that is significantly larger than the remaining activations, which exacerbates the miscalibration problem. Recent observations from the classification

literature suggest that loss functions that embed implicit or explicit maximization of the entropy of predictions yield state-of-the-art calibration performances. Despite these findings, the impact of these losses in the relevant task of calibrating medical image segmentation networks, novel losses specific to task of segmentation, and vision-language remains unexplored.

In the first objective, we refer to one of the earlier works from our group which provides a unifying constrained-optimization perspective of current state-of-the-art calibration losses. Specifically, these losses are viewed as approximations of a linear penalty (or a Lagrangian term) imposing equality constraints on logit distances. This points to an important limitation of such underlying hard equality constraints, whose ensuing gradients constantly push towards a non-informative solution, which might prevent from reaching the best compromise between the discriminative performance and calibration of the model during gradient-based optimization. Following our observations, we extend the proposed simple and flexible generalization penalty which imposes a controllable margin on logit distances to medical image segmentation. We provide comprehensive experiments and ablation studies on five different public segmentation benchmarks that focus on diverse targets and modalities, highlighting the generalization capabilities of the proposed approach. Our empirical results demonstrate the superiority of our method compared to state-of-the-art calibration losses in both calibration and discriminative performance. Importantly, through this work, we established a medical image segmentation benchmark by comparing our method with the standard classification specific loss functions.

In the second objective, we provide a constrained-optimization perspective of Spatially Varying Label Smoothing (SVLS), 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.

In the third objective, we propose a class and region-wise constraint approach to tackle the miscalibration issue in semantic segmentation models. In particular, we formulate a solution that considers the specificities of each category and different regions by introducing independent class and region-wise penalty weights. This contrasts with the prior work, where a uniform scalar penalty weight is employed, regardless of categories or regions. Furthermore, we transfer the constrained problem to its dual unconstrained optimization counterpart by using an Augmented Lagrangian method (ALM). This alleviates the need for manually adjusting each penalty weight and allows, through a series of iterative *inner* and *outer* steps, to find the optimal value of each penalty weight, which can be learned in an adaptive manner. Comprehensive experiments on two popular segmentation benchmarks, and with two well-known segmentation backbones, demonstrate the superiority of our approach over a set of relevant recent calibration approaches.

In the fourth objective, we empirically demonstrate that popular CLIP adaptation strategies, such as Adapters, Prompt Learning, and Test-Time Prompt Tuning, substantially degrade the calibration capabilities of the zero-shot baseline in the presence of distributional drift. For these adaptation strategies, we expose that the underlying cause of miscalibration is, in fact, the increase of the logit ranges. This contrasts with recent work in calibrating fully-supervised models, which

suggests that the inherent cause of miscalibration is the increase of its norm instead, due to the standard cross-entropy loss used for training. Based on these observations, we present a simple, and model-agnostic solution, which consists in scaling the logit range of each sample based on the zero-shot logits. We further present several alternatives to accommodate our solution, which can be implemented either at training or inference time. Comprehensive experiments on popular OOD classification benchmarks empirically demonstrate the effectiveness of our approaches to reduce the miscalibration error, while keeping the discriminative performance.

0.6 Thesis Outline

The organization of the work reported in the thesis is described in this section. This introductory chapter first establishes the motivation for model calibration in classification, explaining the common reasons for miscalibration. It then highlights the importance of calibration for image segmentation, where it leads to more reliable boundaries, and for domain generalization, ensuring trustworthy classification of out-of-distribution data. Lastly, it provides a summary of observed problems and its respective solutions contributing to this thesis. **Chapter 1** introduces the mathematical concept of model calibration in image classification, defining a perfectly calibrated model as one whose confidence scores are both equal to and maximal. The chapter then highlights the three common calibration strategies and provides background on medical image segmentation and vision-language pre-training and adapters. The rest of the chapter shows the existing calibration methods in segmentation, and vision-language foundation models. **Chapter 2** provides a unified view of state-of-the-art calibration losses, framing them as approximations of a linear penalty that imposes equality constraints on logit distances. Building on this observation, the chapter adapts a method that uses a controllable margin on logit distances for medical image segmentation. It then presents a thorough validation of this approach, demonstrating its superiority against current methods across diverse segmentation benchmarks. The content of this chapter corresponds to the journal article *"Calibrating Segmentation networks with*

Margin-Based Label Smoothing" published in Journal of Medical Image Analysis (MedIA), one of the leading journals in the field of medical imaging. **Chapter 3** argues that Spatially Varying Label Smoothing (SVLS) can be viewed as a cross-entropy loss that implicitly constrains predictions to match the label distribution of surrounding pixels. The chapter identifies a key limitation of SVLS: its lack of a strategy to balance this implicit constraint with the primary segmentation objective. To address this, the chapter proposes a solution that uses a simple linear penalty to better model the logit distributions. This proposed solution is then thoroughly analyzed and comprehensively compared against other state-of-the-art segmentation methods, showcasing its effectiveness. This chapter corresponds to the journal article entitled *"Neighbor-Aware Calibration of Segmentation Networks with Penalty-Based Constraints"* published in the Journal of Medical Image Analysis (MedIA), recognized as one of the premier journals within the community. A preliminary version of this work was initially published in Medical Image Computing and Computer-Assisted Intervention (MICCAI), a leading conference in the field. **Chapter 4** explores the possibility of gaining finer control over the spatial-aware calibration loss introduced in the previous chapter, specifically for different categories and regions. It demonstrates how the Augmented Lagrangian Method (ALM) can be used to automatically determine the weights for class and region-wise penalties. The chapter validates this method's superiority by testing it on two popular segmentation benchmarks and two different segmentation backbones. The content presented in this chapter corresponds to the conference proceeding titled *"Class and Region-Adaptive Constraints for Network Calibration"* published in Medical Image Computing and Computer-Assisted Intervention (MICCAI), one of the premier medical imaging conferences. **Chapter 5** empirically demonstrates that popular CLIP adaptation strategies substantially degrade the zero-shot's calibration performance in the presence of distributional drift, and expose that the underlying cause is increase of the logit ranges. Based on these analysis, the chapter provides a simple, and model-agnostic solution, which consists in scaling the logit range of each sample based on the zero-shot logits. The chapter further demonstrates

that the proposed solution reduces the miscalibration error, while keeping the discriminative performance on standard OOD classification benchmarks. The content presented in this chapter corresponds to the conference proceeding titled *"Robust calibration of large vision-language adapters"* published in European Conference on Computer Vision (ECCV), considered one of the top conferences in computer vision. Finally, the **Conclusion** chapter summarizes the calibration solutions developed for least explored domains of medical image segmentation, and vision-language adapters. It then discusses the limitations of these methods and proposes future work, such as using intensity information to better model spatial awareness and analyzing the dataset-specific behavior of zero-shot predictions. Moreover, the chapter also introduces the possibility of applying calibration to other stages of medical imaging pipeline, and to use conformal prediction for more formal uncertainty guarantees.

0.7 Published Work

The main findings in this thesis have led to the following publications.

– Journals:

1. **Murugesan Balamurali**, Adiga Vasudeva Sukesh, Liu Bingyuan, Ben Ayed Ismail, Dolz Jose. "Neighbor-aware calibration of segmentation networks with penalty-based constraints". *Medical Image Analysis (MedIA) - 2025*.
2. **Murugesan Balamurali**, Liu Bingyuan, Lombaert Herve, Ben Ayed Ismail, Dolz Jose. "Calibrating segmentation networks with margin-based label Smoothing". *Medical Image Analysis (MedIA) - 2023*.

– Conferences:

1. **Murugesan Balamurali**, Silva-Rodriguez Julio, Ben Ayed Ismail, Dolz Jose. "Robust Calibration of Large Vision-Language Adapters". *European Conference on Computer Vision (ECCV) - 2024*.

2. **Murugesan Balamurali**, Silva-Rodriguez Julio, Ben Ayed Ismail, Dolz Jose. "Class and Region-Adaptive Constraints for Network Calibration". *Medical Image Computing and Computer Assisted Intervention (MICCAI) - 2024*.
3. **Murugesan Balamurali**, Adiga Vasudeva Sukesh, Liu Bingyuan, Ben Ayed Ismail, Ben Ayed Ismail, Dolz Jose. "Trust your neighbours: Penalty-based constraints for model calibration". *Medical Image Computing and Computer Assisted Intervention (MICCAI) - 2023*.

The other findings in this thesis have led to the following publications.

– Conferences:

1. **Murugesan Balamurali**, Hussain Rukhshanda, Bhattacharya Rajarshi, Ben Ayed Ismail, Dolz Jose. "Prompting classes: Exploring the Power of Prompt Class Learning in Weakly Supervised Semantic Segmentation". *Winter Conference on Applications of Computer Vision (WACV) - 2024*.

0.8 Code Availability

The code implementations of the proposed objective functions in this thesis have been made publicly available through the following links. Besides, we also incorporated our state-of-the-art medical image segmentation calibration loss function NACL in the Medical Open Network for Artificial Intelligence (MONAI) framework, a key artificial intelligence framework for processing medical imaging in healthcare.

1. Robust Calibration of Large Vision-Language Adapters
    a. https://github.com/Bala93/CLIPCalib
2. Class and Region-Adaptive Constraints for Network Calibration
    a. https://github.com/Bala93/CRac
3. Neighbor-Aware Calibration of Segmentation Networks with Penalty-Based Constraints

   a. https://github.com/Bala93/NACL
   b. https://docs.monai.io/en/stable/losses.html#naclloss
4. Calibrating segmentation networks with margin-based label
   a. https://github.com/Bala93/MarginLoss

CHAPTER 1

BACKGROUND

1.1 Preliminaries

Let $\mathcal{D}(\mathcal{X}, \mathcal{Y}) = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{N}$ be the training dataset, with $\mathbf{x}^{(i)} \in \mathcal{X} \subset \mathbb{R}^{\Omega_i}$ representing the i^{th} image, Ω_i the spatial image domain, and $\mathbf{y} \in \mathcal{Y} \subset \{0, 1\}$ its corresponding ground-truth label with K classes, provided as one-hot encoding. Given an input image $\mathbf{x}^{(i)}$, a neural network parameterized by θ generates a logit vector, defined as $f_\theta(\mathbf{x}^{(i)}) = \mathbf{l}^{(i)} \in \mathbb{R}^K$. To simplify the notations, we omit sample indices, as this does not lead to ambiguity, and just use $\mathbf{l} = (l_k)_{1 \leq k \leq K} \in \mathbb{R}^K$ to denote logit vectors. Note that the logits are the inputs of the softmax probability predictions of the network, which are computed as:

$$\mathbf{s} = (s_k)_{1 \leq k \leq K} \in \mathbb{R}^K; \quad s_k = \frac{\exp^{l_k}}{\sum_j^K \exp^{l_j}} \tag{1.1}$$

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

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

1.2 Calibration Methods

In this section, we categorize the state-of-the-art calibration methods into post-hoc methods, regularization methods and uncertainty estimation methods. Besides, we discuss ensemble methods that combine different calibration methods.

1.2.1 Post-hoc Methods

Post-hoc methods involve modifying the predictions of pre-trained model based on the separately held datasets intended for calibration. The two common approaches include non-parametric methods like histogram binning (Zadrozny & Elkan, 2001), isotonic regression (Zadrozny & Elkan, 2002c) and parametric methods like Platt scaling (Platt *et al.*, 1999). Isotonic regression involving fitting a monotonic function to the predictions, while Platt scaling involve learning a parameter model with least calibration error. The parametric methods are more commonly used due to their low complexity, interpretablity, and efficiency. *Temperature Scaling (TS)* and its extensions such as attended TS (Mozafari, Gomes, Leão, Janny & Gagné, 2018), *Dirichlet calibration* (Kull *et al.*, 2019), *Bin-wise TS (BTS)* (Ji, Jung, Yoon, Kim *et al.*, 2019) are some of the widely adapted techniques for Platt scaling. As post-hoc methods decouple calibration from training, it is convenient for the deployment. However, these methods effectiveness largely on the calibration data, and simply finding a temperature value is not enough to capture the possible variations in real data.

1.2.2 Regularization Method

Regularization is key component in preventing neural network models from making overconfident predictions. In this section, we discuss works that either explicitly or implicitly regularizes deep networks for better calibration.

1.2.2.1 Explicit Regularization

The addition of explicitly adding a regularization term (L2 or L1) to provide better generalization, have been effective in model calibration (Guo *et al.*, 2017b). The overconfident predictions generally have peaked softmax distributions, hence enforcing entropy as regularization (Pereyra, Tucker, Chorowski, Kaiser & Hinton, 2017) ensures that the models are penalized for those predictions aiding in calibration. Recently, in an interesting study (Liu, Ben Ayed, Galdran & Dolz, 2022b) showed that popular calibration methods like label smoothing (Müller *et al.*, 2019b),

and (Lin *et al.*, 2017) are also inherently regularizing the entropy. Besides, the work also showed that, constraining the logits is better for calibration compared to operating in label or the softmax space. In one of the works (Liang, Zhang, Wang & Jacobs, 2020), authors have attempted to directly regularize with the non-differential term by obtaining the difference between confidence and accuracy. There is a rising trend (Kumar, Sarawagi & Jain, 2018a; Bohdal, Yang & Hospedales, 2021) to find the differential approximations of calibration error and use it along with the standard losses.

1.2.2.2 Implicit Regularization

Focal loss (Lin *et al.*, 2017) was initially proposed to handle class-imbalance issue in object detection. Recently, it has been shown that, the design of focal loss (Mukhoti *et al.*, 2020b) helps in calibration by implicitly reducing cross entropy with an entropy maximizer as regularizer. Moreover, in the experiment, it was observed that the hyper-parameter γ could be better controlled to improve calibration, and later proposed (Ghosh, Schaaf & Gormley, 2022) sample dependent γ. Instead of adding penalty terms to objective functions, regularization to mitigate model miscalibration could also be provided through data augmentation. Label smoothing (Müller *et al.*, 2019b) soften hard labels with a smoothing parameter in cross entropy, while Mixup (Thulasidasan *et al.*, 2019b) combines inputs and labels to obtain a synthetic sample for objective optimization.

In comparison to post-hoc methods, regularization methods can provided calibrated model without additional processing. Moreover, most of these methods doesn't necessarily increase the model complexity or the training procedures.

1.2.3 Ensemble Methods

Instead of applying each method independently, we can always combine two or more methods to have better calibration. One straightforward way is to combine non-post-hoc methods with post-hoc methods. For instance, performing Temperature Scaling (TS) after employing the

regularization method and implicit calibration (Kumar *et al.*, 2018a; Bohdal *et al.*, 2021). The combination of regularization techniques like mixup training and label smoothing have also shown improved calibration (Thulasidasan *et al.*, 2019b). Methods like MC-dropout (Gal & Ghahramani, 2016), Gumbel-softmax sampling (Jang, Gu & Poole, 2017b; Wang, Lawrence & Niepert, 2021a), and ensembles (Lakshminarayanan *et al.*, 2017b) introduce randomness in the training or inference mechanism to alleviate model miscalibration. The preliminary attempts include starting with different pre-trained weights, or dropping connections randomly during training.

Remarks: The latter set of ensemble works requires multiple inference runs to perform approximations. This increases computational overhead significantly as compared to previous methods. On the other hand, these methods have shown promise in uncertainty quantification and estimation.

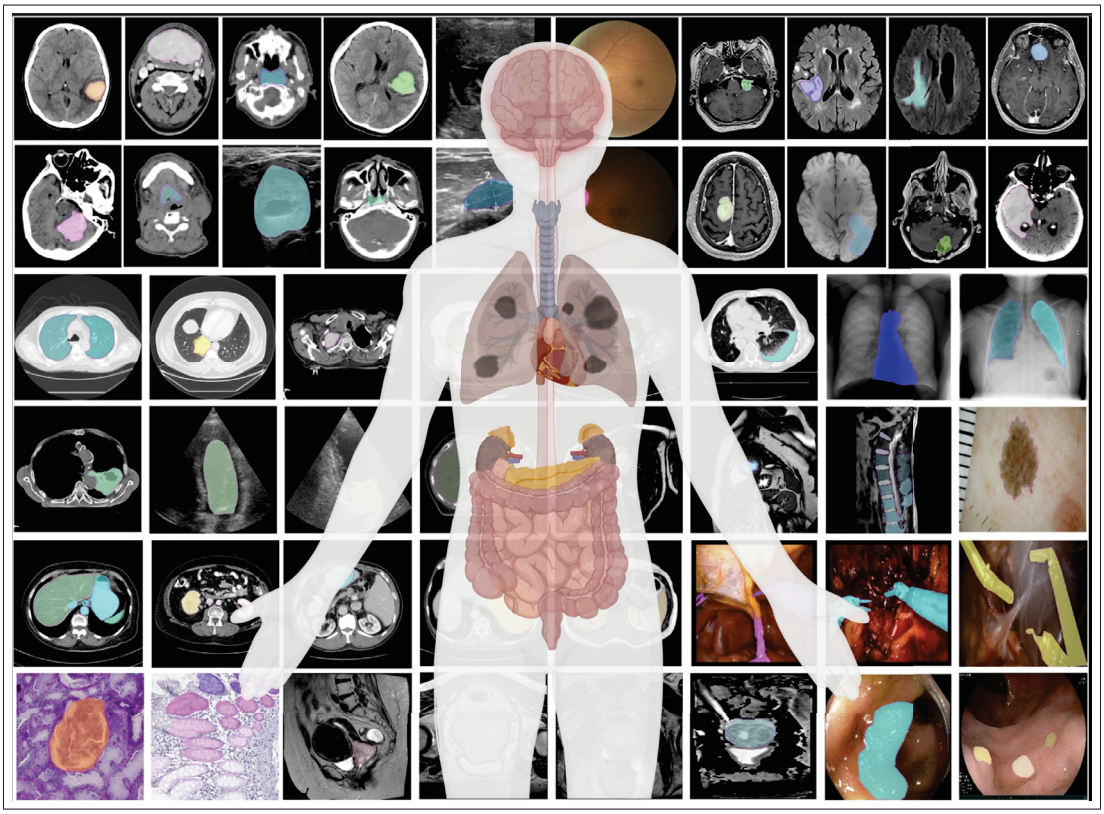

Figure 1.1 Illustration of various medical segmentation benchmarks including different imaging modalities and region of interests

1.3 Medical Image Segmentation

Medical image segmentation refers to the process of delineating organs or abnormal regions from the imaging modalities like CT, MRI, and is one of the key components in the medical image analysis pipeline (Litjens *et al.*, 2017). The results obtained from this stage guide the doctors to make critical decisions including diagnosis, and treatment planning. For instance, in case of tumor identification and removal, segmentation of tumor cells can provide an idea about the malignancy of the cancer cells and can later be used to surgically remove them through guidance systems (Ranjbarzadeh *et al.*, 2021). A few examples of the imaging modalities, and its respective target region annotations can be found in Figure 1.1. Before the advent of deep learning, this process was either done manually or through conventional techniques involving manual feature design (Bensch & Ronneberger, 2015). Despite their efficiency and interpretability, the techniques were unable to deliver the anticipated accuracy on the benchmark datasets. Similar to the AlexNet (Krizhevsky, Sutskever & Hinton, 2017) moment in ImageNet classification benchmark, and SegNet (Badrinarayanan, Kendall & Cipolla, 2017) for CamVid dataset, ISBI 2015 Cell Segmentation model (Maška *et al.*, 2014) had U-Net (Ronneberger *et al.*, 2015b) which overhauled the medical imaging segmentation domain. The promising results of U-Net in this challenge laid the groundwork for the subsequent works (Zhou, Siddiquee, Tajbakhsh & Liang, 2020; Isensee *et al.*, 2021; Chen *et al.*, 2021; Guo *et al.*, 2022; Cao *et al.*, 2022; Maška *et al.*, 2023; Ma *et al.*, 2024; Wang, Zheng, Zhang, Cui & Li, 2024c).

1.4 Vision-Language Models

Text-driven pre-training of image representation, so-called vision-language models (VLMs) is revolutionizing the paradigm of transfer learning. These models can integrate massive web-scrabbled data sources thus learning robust feature representations. In particular, models such as CLIP (Radford *et al.*, 2021) or ALIGN (Jia *et al.*, 2021) train joint multi-modal embedding spaces via contrastive learning of paired images and text, using dual encoder architectures. Such strong vision-language alignment has demonstrated robust open-vocabulary zero-shot generalization capabilities (Radford *et al.*, 2021; Zhai *et al.*, 2022). The overall outline of a typical Vision-

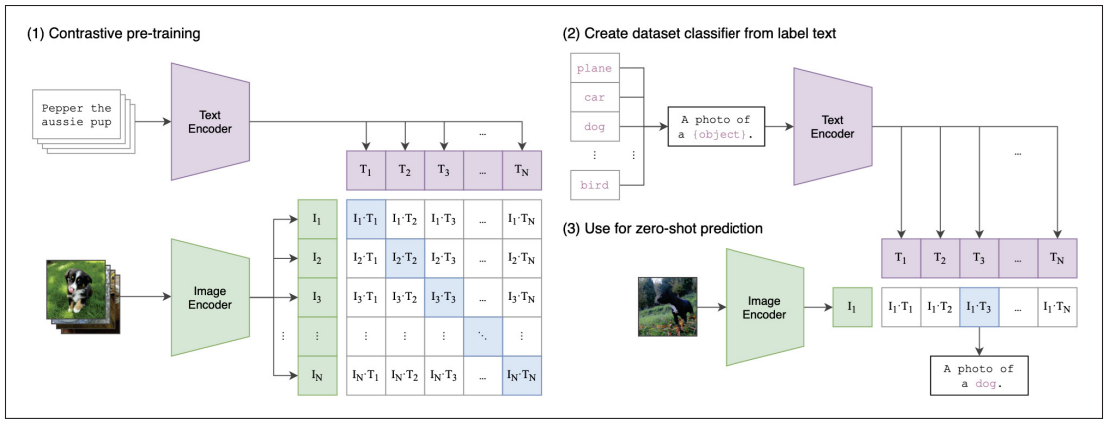

Figure 1.2 Summary of CLIP – Image and Text encoder are trained to predict the correct pairings of batch of (image, text) samples. In test time, zero-shot linear classifier is constructed with the text encoder embeddings of the possible target classes along with the input image

language model, in this case, CLIP is shown in Figure 1.2. Given such potential, transferring pre-trained VLMs to a wide variety of tasks is gaining increasing popularity. Nevertheless, this process faces particular challenges. First, large-scale pre-training usually involves also scaling network sizes, which is a computational bottleneck for low-resource adaptation scenarios. Second, recent attempts to fine-tune VLMs have demonstrated a deterioration of their robustness against domain drifts (Kumar, Raghunathan, Jones, Ma & Liang, 2022; Wortsman *et al.*, 2022), especially when available data is limited. Thus, an emerging core of recent literature is focusing on novel alternatives to overcome these limitations. More concretely, freezing the pre-trained backbone, and reusing such features by training a small set of parameters, via Prompt Learning (Zhou *et al.*, 2022c; Zhou, Yang, Loy & Liu, 2022a; Hantao Yao,Rui Zhang, 2023; Zhu, Niu, Han, Wu & Zhang, 2023; Khattak, Rasheed, Maaz, Khan & Khan, 2023), or black-box Adapters (Gao *et al.*, 2024; Zhang *et al.*, 2022b; Yu, Lu, Jin, Chen & Wang, 2023b; Silva-Rodriguez, Hajimiri, Ayed & Dolz, 2024; Ouali, Bulat, Martinez & Tzimiropoulos, 2023; Li *et al.*, 2024; Zhang *et al.*, 2023), is getting increasing attention.

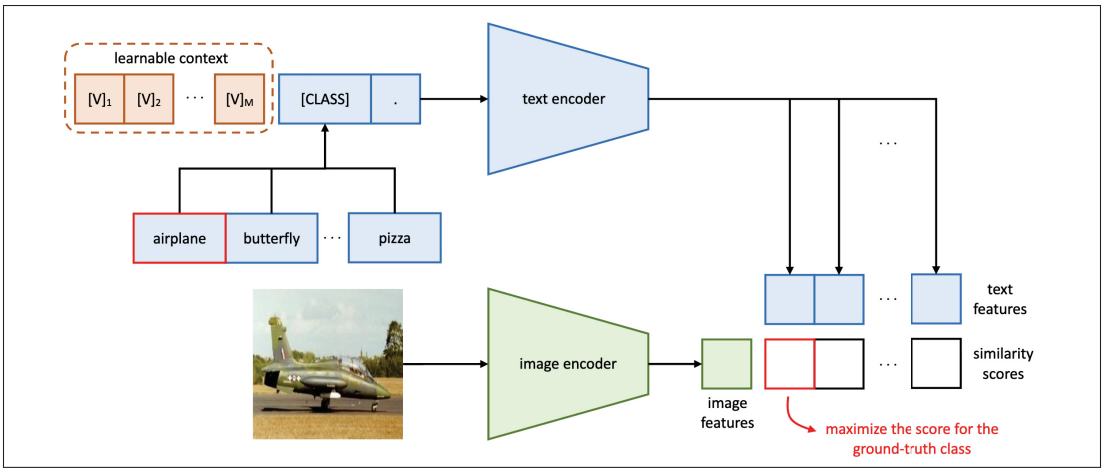

Figure 1.3 An example of one of the earlier versions of prompt learning – CoOp. The prompt's context uses a set of learnable vectors which are optimized by minimizing the classification loss

1.4.1 Prompt based learning

CLIP models have shown encouraging results by hand-crafting personalized text descriptions of the target visual representation (Menon & Vondrick, 2023). The automatizing of this cumbersome process raises the concept of Prompt Learning (PL) (Zhou *et al.*, 2022c), a family of methods to adapt CLIP that inserts a set of continuous learnable tokens in the original text prompt at the input of the VLM language encoder. While the CLIP model remains frozen, PL optimizes the most discriminative text input, given a few-shot support set (Zhou *et al.*, 2022c,a; Khattak *et al.*, 2023; Zhu *et al.*, 2023). CoOp (Zhou *et al.*, 2022c) represents one of the initial attempts to study the effect of prompt tuning on different tasks, and proposed to learn the prompt's context words. CoCoOp (Zhou *et al.*, 2022a), on the other hand, designed a simple network to predict the input text prompt through image features, as CoOp failed to match the zero-shot performance on generic tasks. Figure 1.3 showcases the pipeline of CoOp, where during inference, the concatenated version of target classes embedding along with the learned contexts is used. TPT (Shu *et al.*, 2022) extends PL to address time-test adaptation scenarios by updating the prompt for a batch with original and augmented samples through entropy minimization.

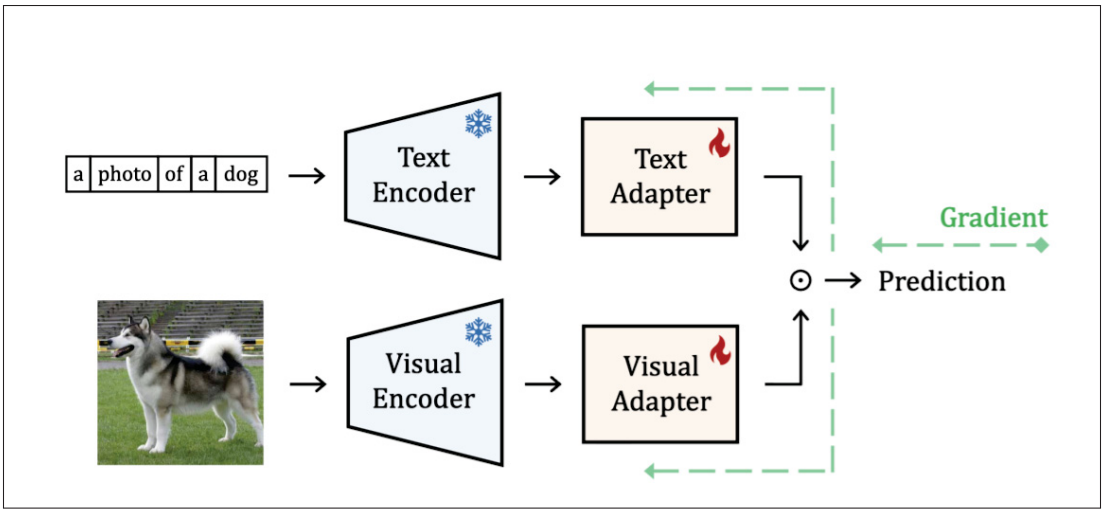

Figure 1.4 CLIP-Adapter – one of the preliminary versions of adapters which does not require to calculate and propagate the gradients through CLIP's encoder

1.4.2 Black-box Adapters

Prompt Learning involves using the CLIP's encoder throughout the entire training process as the backpropagation of the gradient has to pass through it to update the prompts, which results in large computational constraints (Gao *et al.*, 2024). Adapter-based techniques provide an alternative to Prompt Learning for aligning to downstream tasks, leveraging uniquely pre-computed features with minimal additional parameters. A base version of such methods involves training a linear classifier via logistic regression, typically referred to as Linear Probing (Radford *et al.*, 2021). Nevertheless, leveraging only the vision features does not fully exploit the potential of VLMs. To this end, several methods have proposed enhanced Adapters, which further rely on zero-shot text-driven class-wise prototypes. In particular, Clip-Adapter (Gao *et al.*, 2024) introduced additional fully connected layers and operated on the vision or language branch through residual style feature combination. The layout of CLIP-Adpater can be found in 1.4 showing that encoder weights are frozen while updating the post encoders weights. Training-free methods such as Tip-Adapter (Zhang *et al.*, 2022b) resorted to a key-value cache model based on the available few-shot supports. Likewise, TaskRes (Yu *et al.*, 2023b) introduced additional learning parameters and applied a residual modification of the text representation, which led to a better

initialization when learning from few-shot supervision. More recently, (Silva-Rodriguez *et al.*, 2024) provided a wider look at the coupling of vision and text features in such Adapters, by pointing out that these methods are sensitive to hyper-parameters and largely build up their improved performance on initializing the logistic classifier weights with the zero-shot prototypes, which in itself is strong baseline. Moreover, they proposed a simple solution, coined CLAP, for a better distillation of such prototypes.

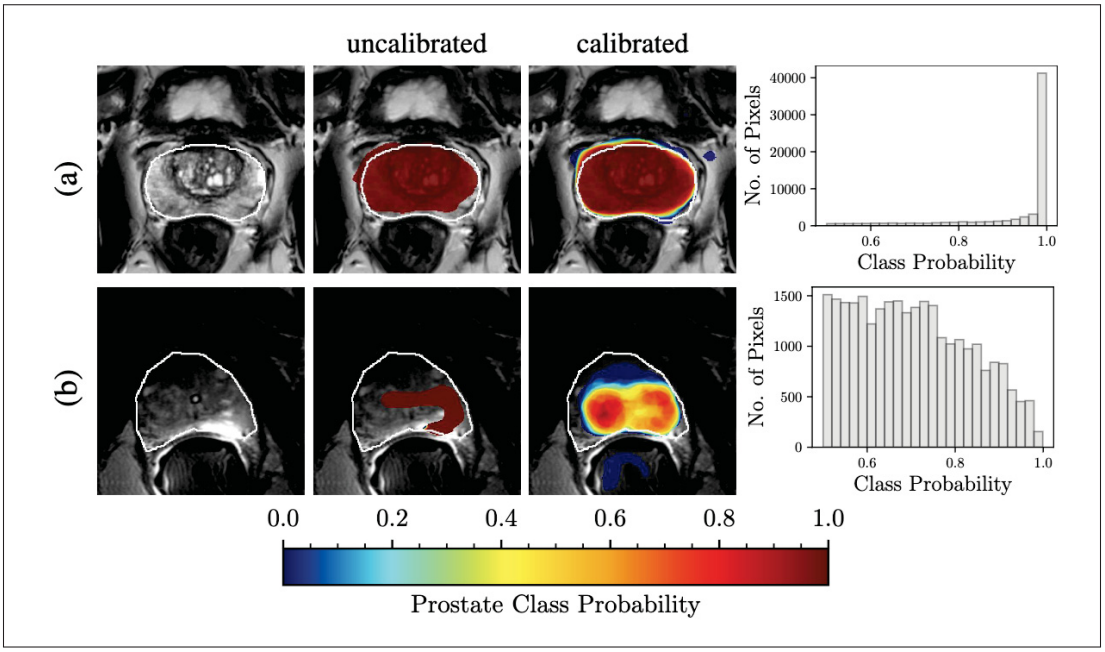

Figure 1.5 Illustration of calibrated model providing better segmentation and uncertainty estimate compared to the uncalibrated model. In the top row, both the models overall provide confident right predictions, whereas in the bottom row for a challenging domain shifted version, calibrated model provide uncertain predictions (desirable), compared to confident wrong predictions by uncalibrated model. The distribution of class probabilities demonstrates the prediction uncertainty

1.5 Calibration in Medical Image Segmentation

Recent literature has focused on either estimating the predictive uncertainty or on leveraging this uncertainty to improve the discriminative performance of segmentation models (Wang *et al.*, 2019b; Zou *et al.*, 2023; Linmans, Elfwing, van der Laak & Litjens, 2023; Adiga, Dolz & Lombaert, 2024). Nevertheless, research to improve both the calibration and segmentation

performance of CNN-based segmentation models is scarce and has been largely disregarded. In Figure 1.5, we showcase a contrastive example to demonstrate the importance of having a segmentation model which is also better calibrated. (Jena & Awate, 2019) proposed a novel deep segmentation framework rooted in generative modeling and Bayesian decision theory, which allowed to define a principled measure of uncertainty associated with label probabilities. Recent findings (Fort *et al.*, 2019), however, suggest that current state-of-the-art Bayesian neural networks have tendency to find solutions around a single minimum of the loss landscape and, consequently, lack diversity. In contrast, ensembling deep neural networks typically results in more diverse predictions, and therefore obtain better uncertainty estimates. This observation aligns with the recent work in (Jungo, Balsiger & Reyes, 2020; Mehrtash *et al.*, 2020), which evaluates several uncertainty estimation approaches and concludes that ensembling outperforms other methods. To promote model diversity within the ensemble, (Larrazabal, Martínez, Dolz & Ferrante, 2021) integrate an orthogonality constraint in the learning objective, showing significant gains over the non-constrained set. More recently, (Karimi & Gholipour, 2022) argue that training a single model in a multi-task manner on several different datasets yields better calibration on the different tasks compared to its single-task counterpart. Nevertheless, these methods incur in high computationally expensive steps as they involve training either multiple models or a single model on multiple datasets. In an orthogonal direction, several recent methods have overcome this limitation and proposed lighter alternatives. For example, (Ding *et al.*, 2021) extends the naive temperature scaling by integrating a simple CNN to predict the pixel-wise temperature values in a post-processing step. Despite the improvement observed over the naive TS, this method inherits the limitations of temperature scaling and related post-processing approaches. In addition, (Islam & Glocker, 2021) apply a weight matrix with a Gaussian kernel across the one-hot encoded expert labels to obtain soft class probabilities, adding into the standard Label smoothing a spatial-awareness. A potential limitation of this strategy is that the modification of the hard labels is done without considering the behaviour of the model, systematically disregarding those samples which are more, or less, confident. This contrasts with the proposed approach, which does not modify the hard assignments, but directly controls the confidence of the model in the logit space. However, despite these initial efforts,

and to the best of our knowledge, a comprehensive evaluation of calibration methods in multiple medical image segmentation benchmarks has not been conducted yet.

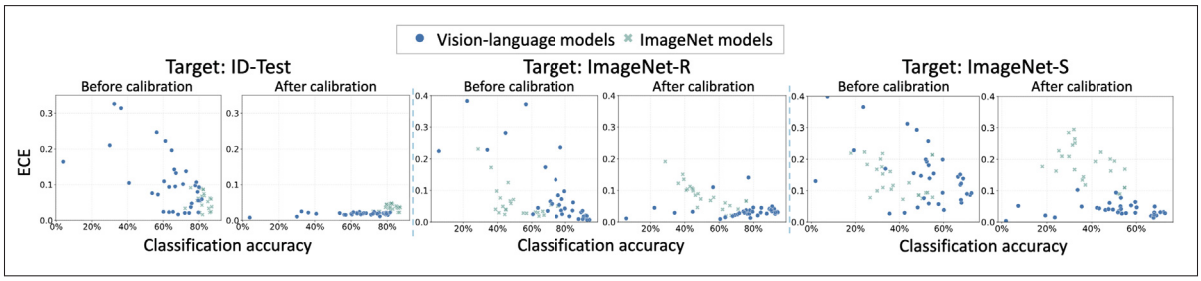

Figure 1.6 VLMs are well-calibrated by temperature scaling on both ID and OOD test sets compared to ImageNet-trained models

1.6 Calibration in Vision-Language Models

(Tu, Deng, Campbell, Gould & Gedeon, 2024) have investigated the factors that affect the uncertainty estimation performance of Vision-Language Models (VLMs). They have shown that, VLMs coupled with temperature scaling surpass other models in their ability to estimate uncertainty accurately (LeVine, Pikus, Raja & Gil, 2023), Figure 1.6 is provided as reference. Moreover, it has also been noted that VLMs can be calibrated with datasets that have different label sets or label hierarchy levels. Through thorough analysis (Wang *et al.*, 2024b), it has been shown that fine-tuned VLMs often suffer from miscalibration, especially in the open-vocabulary setting. The authors demonstrate the correlation between the calibration and the textual distribution gap, and also show that after prompt learning, VLMs tend to be overconfident on classes far away from base classes. To overcome this problem, Distance-Aware Calibration, a simple and effective temperature scale is performed with distance between predicted text labels and base classes as a guidance. Likewise, current prompt-tuning methods typically lead to a trade-off between base and new classes, compromising one of them. The proposed Dynamic Outlier Regularization (DOR) (Wang, Li & Wei, 2024a), a simple yet effective regularization that ensures calibration performance on both base and new classes. Instead of adaptation, if we fine-tune the models, (Oh *et al.*, 2024) argues that the existing models do not adequately achieve satisfactory OOD generalization and calibration simultaneously. To support, the work shows

that both classification and calibration errors are bounded from above by the ID calibration error and the smallest singular value of the covariance matrix over the ID input representation. The proposed CaRot reduces the upper bound by conducting constrained multimodal contrastive learning with EMA self-distillation. In test-time adaptation setting, (Yoon *et al.*, 2024) showed that the calibration of CLIP models is significantly influenced by the prompt choice, with certain prompts offering superior calibration with the same prediction accuracy level. It has been identified that the critical difference between these prompts can be characterized by the distribution of the class-embedded text features, with a noticeable negative correlation between the dispersion of the text features and the calibration error. This paper introduces Calibrated Test-time Prompt Tuning (C-TPT), which is used to jointly optimize the prompt during test time to achieve better calibration by maximizing Average Text Feature Dispersion (ATFD). Recently, O-TPT (Sharifdeen, Munir, Baliah, Khan & Khan, 2025) showed that the calibration of test-time prompt tuning of VLMs can be accomplished by introducing orthogonalization constraints on the textual features by enforcing the angular distance between them.

CHAPTER 2

CALIBRATING SEGMENTATION NETWORKS WITH MARGIN-BASED LABEL SMOOTHING

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

[a] Department of Software Engineering, École de Technologie Supérieure,
[b] Department of System Engineering, École de Technologie Supérieure,
1100 Notre-Dame West, Montreal, Quebec, Canada H3C 1K3
[c] Universitat Pompeu Fabra, Barcelona, Spain

Paper published in Medical Image Analysis, July 2023

Presentation

This chapter presents the article *"Calibrating segmentation networks with margin-based label smoothing"* (Murugesan, Liu, Galdran, Ayed & Dolz, 2023b) published in Medical Image Analysis (**MedIA**) Volume 87, Page 102826, 2023.

Abstract

Despite the undeniable progress in visual recognition tasks fueled by deep neural networks, there exists recent evidence showing that these models are poorly calibrated, resulting in over-confident predictions. The standard practices of minimizing the cross-entropy loss during training promote the predicted softmax probabilities to match the one-hot label assignments. Nevertheless, this yields a pre-softmax activation of the correct class that is significantly larger than the remaining activations, which exacerbates the miscalibration problem. Recent observations from the classification literature suggest that loss functions that embed implicit or explicit maximization of the entropy of predictions yield state-of-the-art calibration performances. Despite these findings, the impact of these losses in the relevant task of calibrating medical image segmentation networks remains unexplored. In this work, we provide a unifying constrained-optimization perspective of current state-of-the-art calibration losses. Specifically, these losses could be viewed as approximations of a linear penalty (or a Lagrangian term) imposing equality

constraints on logit distances. This points to an important limitation of such underlying equality constraints, whose ensuing gradients constantly push towards a non-informative solution, which might prevent from reaching the best compromise between the discriminative performance and calibration of the model during gradient-based optimization. Following our observations, we propose a simple and flexible generalization based on inequality constraints, which imposes a controllable margin on logit distances. Comprehensive experiments on a variety of public medical image segmentation benchmarks demonstrate that our method sets novel state-of-the-art results on these tasks in terms of network calibration, whereas the discriminative performance is also improved. The code is available at https://github.com/Bala93/MarginLoss

2.1 Introduction

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

Quantifying the predictive uncertainty of modern DNNs has gained popularity recently, with several alternatives to train better calibrated models. A simple yet effective approach consists in integrating a post-processing step that modifies the predicted probabilities of a trained neural network (Guo *et al.*, 2017b; Zhang, Kailkhura & Han, 2020c; Tomani, Gruber, Erdem, Cremers & Buettner, 2021a; Ding *et al.*, 2021). This strategy, however, presents several limitations. First, the choice of the transformation parameters, such as temperature scaling, is highly dependent on the dataset and network. And second, under domain drift, post-hoc calibration performance largely degrades (Ovadia *et al.*, 2019), resulting in unreliable predictions. A more principled alternative is to explicitly maximize the Shannon entropy of the model predictions during training, which can be achieved by augmenting the learning objective with a

term that penalizes confident output distributions (Pereyra *et al.*, 2017). Furthermore, recent efforts to quantify the quality of predictive uncertainties have focused on investigating the effect of the entropy on the training labels (Xie, Wang, Wei, Wang & Tian, 2016; Müller *et al.*, 2019b; Mukhoti *et al.*, 2020b). Findings from these works evidence that, popular losses, which modify the hard-label assignments, such as label smoothing (Szegedy, Vanhoucke, Ioffe, Shlens & Wojna, 2016) and focal loss (Lin *et al.*, 2017), implicitly integrate an entropy maximization objective and have a favourable effect on model calibration. As shown comprehensively in the recent study in (Mukhoti *et al.*, 2020b), these losses, with implicit or explicit maximization of the entropy, represent the state-of-the-art in model calibration in visual and non-visual recognition tasks.

Despite this progress, the benefit of these calibration losses remains unclear in medical image segmentation. Indeed, only a handful of works have addressed this important problem, mostly focusing on the calibration assessment of standard segmentation losses (Mehrtash *et al.*, 2020), i.e., cross-entropy and Dice. From a clinical perspective, semantic segmentation is of pivotal importance in several downstream tasks, such as diagnostic, surgical planning, treatment assessment, or following-up disease progress. In these important steps, clinicians equipped with segmentation uncertainty can make better decisions, and build trust on the system. For example, a clinician faced with a large segmentation error localized in a particular area of an image and a small error at any other region, without knowledge of the segmentation uncertainty, may decide to dismiss the segmentation entirely. On the other hand, by providing precise estimates of the segmentation uncertainties, the clinician could evaluate whether these regions lie in low uncertainty or high uncertainty areas, which will facilitate the assessment of the quality of the segmentation per region. We stress that if clinicians place unwarranted confidence on regions with inaccurate uncertainty estimates, the resulting decision might have catastrophic consequences. Thus, we believe that it is of great significance and interest to study methods for confidence calibration of segmentation models in the context of medical imaging.

The contributions of this work are summarized as follows:

- We provide a unifying constrained-optimization perspective of current state-of-the-art calibration losses. Specifically, these losses could be viewed as approximations of a linear

penalty (or a Lagrangian term) imposing equality constraints on logit distances. This points to an important limitation of such underlying hard equality constraints, whose ensuing gradients constantly push towards a non-informative solution, which might prevent from reaching the best compromise between the discriminative performance and calibration of the model during gradient-based optimization.

- Following our observations, we propose a simple and flexible generalization based on inequality constraints, which imposes a controllable margin on logit distances.
- We provide comprehensive experiments and ablation studies on five different public segmentation benchmarks that focus on diverse targets and modalities, highlighting the generalization capabilities of the proposed approach. Our empirical results demonstrate the superiority of our method compared to state-of-the-art calibration losses in both calibration and discriminative performance.

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

2.2 Related work

Post-processing approaches: Including a post-processing step that transforms the probability predictions of a deep network (Guo *et al.*, 2017b; Zhang *et al.*, 2020c; Tomani *et al.*, 2021a; Ding *et al.*, 2021) is a straightforward yet efficient strategy to mitigate miscalibrated predictions. Among these methods, *temperature scaling* (Guo *et al.*, 2017b), a variant of Platt scaling (Platt

et al., 1999), employs a single scalar parameter over all the pre-softmax activations, which results in softened class predictions. Despite its good performance on in-domain samples, (Ovadia *et al.*, 2019) demonstrated that temperature scaling does not work well under data distributional shift. (Tomani *et al.*, 2021a) mitigated this limitation by transforming the validation set before performing the post-hoc calibration step, whereas (Ma & Blaschko, 2021) introduced a ranking model to improve the post-processing model calibration.

Probabilistic and non-probabilistic approaches: have been also investigated to measure the uncertainty of the predictions in modern deep neural networks. For example, prior literature has employed Bayesian neural networks to approximate inference by learning a posterior distribution over the network parameters, as obtaining the exact Bayesian inference is computationally intractable in deep networks. These Bayesian-based models include variational inference (Blundell *et al.*, 2015; Louizos & Welling, 2016), stochastic expectation propagation (Hernández-Lobato & Adams, 2015) or dropout variational inference (Gal & Ghahramani, 2016). A popular non-parametric alternative is ensemble learning, where the empirical variance of the network predictions is used as an approximate measure of uncertainty. This yields improved discriminative performance, as well as meaningful predictive uncertainty with reduced miscalibration. Common strategies to generate ensembles include differences in model hyperparameters (Wenzel, Snoek, Tran & Jenatton, 2020), random initialization of the network parameters and random shuffling of the data points (Lakshminarayanan, Pritzel & Blundell, 2017a), Monte-Carlo Dropout (Gal & Ghahramani, 2016; Zhang, Dalca & Sabuncu, 2019), dataset shift (Ovadia *et al.*, 2019) or model orthogonality constraints (Larrazabal *et al.*, 2021). However, a main drawback of this strategy stems from its high computational cost, particularly for complex models and large datasets.

Explicit and implicit penalties: Modern classification networks trained under the fully supervised learning paradigm resort to training labels provided as binary one-hot encoded vectors. Therefore, all the probability mass is assigned to a single class, resulting in minimum-entropy supervisory signals (i.e., entropy equal to zero). As the network is trained to follow this distribution, we are implicitly forcing it to be overconfident (i.e., to achieve a minimum

entropy), thereby penalizing uncertainty in the predictions. While temperature scaling artificially increases the entropy of the predictions, (Pereyra *et al.*, 2017) included into the learning objective a term to penalize confident output distributions by explicitly maximizing the entropy. In contrast to tackling overconfidence directly on the predicted probability distributions, recent works have investigated the effect of the entropy on the training labels. The authors of (Xie *et al.*, 2016) explored adding label noise as a regularization, where the disturbed label vector was generated by following a generalized Bernoulli distribution. Label smoothing (Szegedy *et al.*, 2016), which successfully improves the accuracy of deep learning models, has been shown to implicitly calibrate the learned models, as it prevents the network from assigning the full probability mass to a single class, while maintaining a reasonable distance between the logits of the ground-truth class and the other classes (Müller *et al.*, 2019b). More recently, (Mukhoti *et al.*, 2020b) demonstrated that focal loss (Lin *et al.*, 2017) implicitly minimizes a Kullback-Leibler (KL) divergence between the uniform distribution and the softmax predictions, thereby increasing the entropy of the predictions. Indeed, as shown in (Müller *et al.*, 2019b; Mukhoti *et al.*, 2020b), both label smoothing and focal loss implicitly regularize the network output probabilities, encouraging their distribution to be close to the uniform distribution. To our knowledge, and as demonstrated experimentally in the recent studies in (Müller *et al.*, 2019b; Mukhoti *et al.*, 2020b), loss functions that embed implicit or explicit maximization of the entropy of the predictions yield state-of-the-art calibration performances.

2.2.0.0.1 Calibration in medical image segmentation

Recent literature has focused on either estimating the predictive uncertainty or on leveraging this uncertainty to improve the discriminative performance of segmentation models (Wang *et al.*, 2019b). Nevertheless, research to improve both the calibration and segmentation performance of CNN-based segmentation models is scarce. (Jena & Awate, 2019) proposed a novel deep segmentation framework rooted in generative modeling and Bayesian decision theory, which allowed to define a principled measure of uncertainty associated with label probabilities. Recent findings (Fort *et al.*, 2019), however, suggest that current state-of-the-art Bayesian neural

networks have tendency to find solutions around a single minimum of the loss landscape and, consequently, lack diversity. In contrast, ensembling deep neural networks typically results in more diverse predictions, and therefore obtain better uncertainty estimates. This observation aligns with the recent work in (Jungo *et al.*, 2020; Mehrtash *et al.*, 2020), which evaluates several uncertainty estimation approaches and concludes that ensembling outperforms other methods. To promote model diversity within the ensemble, (Larrazabal *et al.*, 2021) integrate an orthogonality constraint in the learning objective, showing significant gains over the non-constrained set. More recently, (Karimi & Gholipour, 2022) argue that training a single model in a multi-task manner on several different datasets yields better calibration on the different tasks compared to its single-task counterpart. Nevertheless, these methods incur in high computationally expensive steps as they involve training either multiple models or a single model on multiple datasets. In an orthogonal direction, several recent methods have overcome this limitation and proposed lighter alternatives. For example, (Ding *et al.*, 2021) extends the naive temperature scaling by integrating a simple CNN to predict the pixel-wise temperature values in a post-processing step. Despite the improvement observed over the naive TS, this method inherits the limitations of temperature scaling and related post-processing approaches. In addition, (Islam & Glocker, 2021) apply a weight matrix with a Gaussian kernel across the one-hot encoded expert labels to obtain soft class probabilities, adding into the standard Label smoothing a spatial-awareness. A potential limitation of this strategy is that the modification of the hard labels is done without considering the behaviour of the model, systematically disregarding those samples which are more, or less, confident. This contrasts with the proposed approach, which does not modify the hard assignments, but directly controls the confidence of the model in the logit space. However, despite these initial efforts, and to the best of our knowledge, a comprehensive evaluation of calibration methods in multiple medical image segmentation benchmarks has not been conducted yet.

2.3 Preliminaries

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

$$\mathbf{s} = (s_k)_{1 \leq k \leq K} \in \mathbb{R}^K; \quad s_k = \frac{\exp^{l_k}}{\sum_j^K \exp^{l_j}} \tag{2.1}$$

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

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

Miscalibration of DNNs. To train fully supervised discriminative deep models, the standard cross-entropy (CE) loss is commonly used as the training objective. We argue that, from a calibration performance, the supervision of CE is suboptimal. Indeed, CE reaches its minimum when the predictions for all the training samples match the hard (binary) ground-truth labels, i.e., $s_k = 1$ when k is the ground-truth class of the sample and $s_k = 0$ otherwise. Minimizing the CE implicitly pushes softmax vectors $\mathbf{s}$ towards the vertices of the simplex, thereby magnifying the distances between the largest logit $\max_k(l_k)$ and the rest of the logits, yielding over-confident predictions and miscalibrated models.

2.4 A constrained-optimization perspective of calibration

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

2.4.1 Definition of logit distances

Let us first define the vector of logit distances between the winner class (i.e., the class with the highest logit: $\arg\max_j(l_j)$) and the remaining classes as:

$$\mathbf{d}(\mathbf{l}) = (\max_j(l_j) - l_k)_{1\leq k\leq K} \in \mathbb{R}^K \tag{2.2}$$

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

2.4.2 Penalty functions in constrained optimization

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

$$\mathbf{d}(\mathbf{l}) = \mathbf{0} \tag{2.3}$$

The general principle of a soft-penalty optimizer is to replace a hard constraint of the form in Eq. 2.3 by adding an additional term $\mathcal{P}(\mathbf{d}(\mathbf{l}))$ into the main objective function to be minimized. Soft penalty $\mathcal{P}$ should be a continuous and differentiable function, which reaches its global minimum when the constraint is satisfied, i.e., it verifies: $\mathcal{P}(\mathbf{d}(\mathbf{l})) \geq \mathcal{P}(\mathbf{0}) \, \forall \, \mathbf{l} \in \mathbb{R}^K$. Thus, when the constraint is violated, i.e., when $\mathbf{d}(\mathbf{l})$ deviates from $\mathbf{0}$, the penalty term $\mathcal{P}$ increases.

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

$$\mathcal{L}_{\text{LS}} = -\sum_k y_k^{\text{LS}} \log s_k = -\sum_k ((1 - \alpha) y_k + \frac{\alpha}{K}) \log s_k \tag{2.4}$$

where $\alpha \in [0, 1]$ is the smoothing hyper-parameter. It is straightforward to verify that cross-entropy with label smoothing in Eq. 2.4 can be decomposed into a standard cross-entropy term augmented with a Kullback-Leibler (KL) divergence between uniform distribution $\mathbf{u} = \frac{1}{K}$ and the softmax prediction:

$$\mathcal{L}_{\text{LS}} \overset{\text{c}}{=} \mathcal{L}_{\text{CE}} + \frac{\alpha}{1 - \alpha} \mathcal{D}_{\text{KL}} \left(\mathbf{u} || \mathbf{s}\right) \tag{2.5}$$

where $\overset{\text{c}}{=}$ stands for equality up to additive and/or non-negative multiplicative constants. Now, consider the following bounding relationships between a linear penalty (or a Lagrangian) for equality constraint $\mathbf{d}(\mathbf{l}) = \mathbf{0}$ and the KL divergence in Eq. 2.5.

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

$$\mathcal{D}_{KL}\left(\mathbf{u} || \mathbf{s}\right) - \log(K) \overset{\text{c}}{\leq} \frac{1}{K} \sum_k (\max_j(l_j) - l_k) \overset{\text{c}}{\leq} \mathcal{D}_{KL}\left(\mathbf{u} || \mathbf{s}\right) \tag{2.6}$$

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

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

Prop. 1 means that LS is (approximately) optimizing a linear penalty (or a Lagrangian) for logit-distance constraint $\mathbf{d}(\mathbf{l}) = \mathbf{0}$, which encourages equality of all logits; see the illustration in Figure 2.1, top-left.

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

$$\mathcal{L}_{\text{FL}} = -\sum_k (1 - s_k)^\gamma y_k \log s_k \geq \mathcal{L}_{\text{CE}} - \gamma \mathcal{H}(\mathbf{s}) \tag{2.7}$$

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

$$\mathcal{H}(\mathbf{s}) = -\sum_k s_k \log(s_k) \tag{2.8}$$

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

$$-\mathcal{H}(\mathbf{s}) \overset{\text{c}}{=} \mathcal{D}_{\text{KL}}(\mathbf{s}||\mathbf{u}) \tag{2.9}$$

which is a reversed form of the KL term in Eq. 2.5.

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

2.4.3 Margin-based Label Smoothing (MbLS)

Our previous analysis shows that LS, FL and ECP are closely related from a constrained-optimization perspective, and they could be seen as approximations of a linear penalty for imposing constraint $\mathbf{d}(\mathbf{l}) = \mathbf{0}$, pushing all logit distances to zero; see Figure 2.1, top-left. Clearly, enforcing this constraint in a hard way yields a non-informative solution where all the classes have exactly the same logit and, hence, the same class prediction: $s_k = \frac{1}{K} \forall K$. While this trivial solution is not reached in practice when using soft penalties (as in LS, FL and ECP) jointly with CE, we argue that the underlying equality constraint $\mathbf{d}(\mathbf{l}) = \mathbf{0}$ has an important limitation, which might prevent from reaching the best compromise between the discriminative performance and calibration of the model during gradient-based optimization. Figure 2.1, left, illustrates this: With the linear penalty for constraint $\mathbf{d}(\mathbf{l}) = \mathbf{0}$ in the top-left of the Figure, the derivative with respect to logit distances is a strictly positive constant (left-bottom), yielding during training *a gradient term that constantly pushes towards the trivial, non-informative solution* $\mathbf{d}(\mathbf{l}) = \mathbf{0}$ (or equivalently $s_k = \frac{1}{K} \forall K$). To alleviate this issue, we propose to replace the equality constraint $\mathbf{d}(\mathbf{l}) = \mathbf{0}$ with the more general inequality constraint $\mathbf{d}(\mathbf{l}) \leq \mathbf{m}$, where $\mathbf{m}$ denotes the K-dimensional vector with all elements equal to $m > 0$. Therefore, we include a margin m into the penalty, so that the logit distances in $\mathbf{d}(\mathbf{l})$ are allowed to be below m when

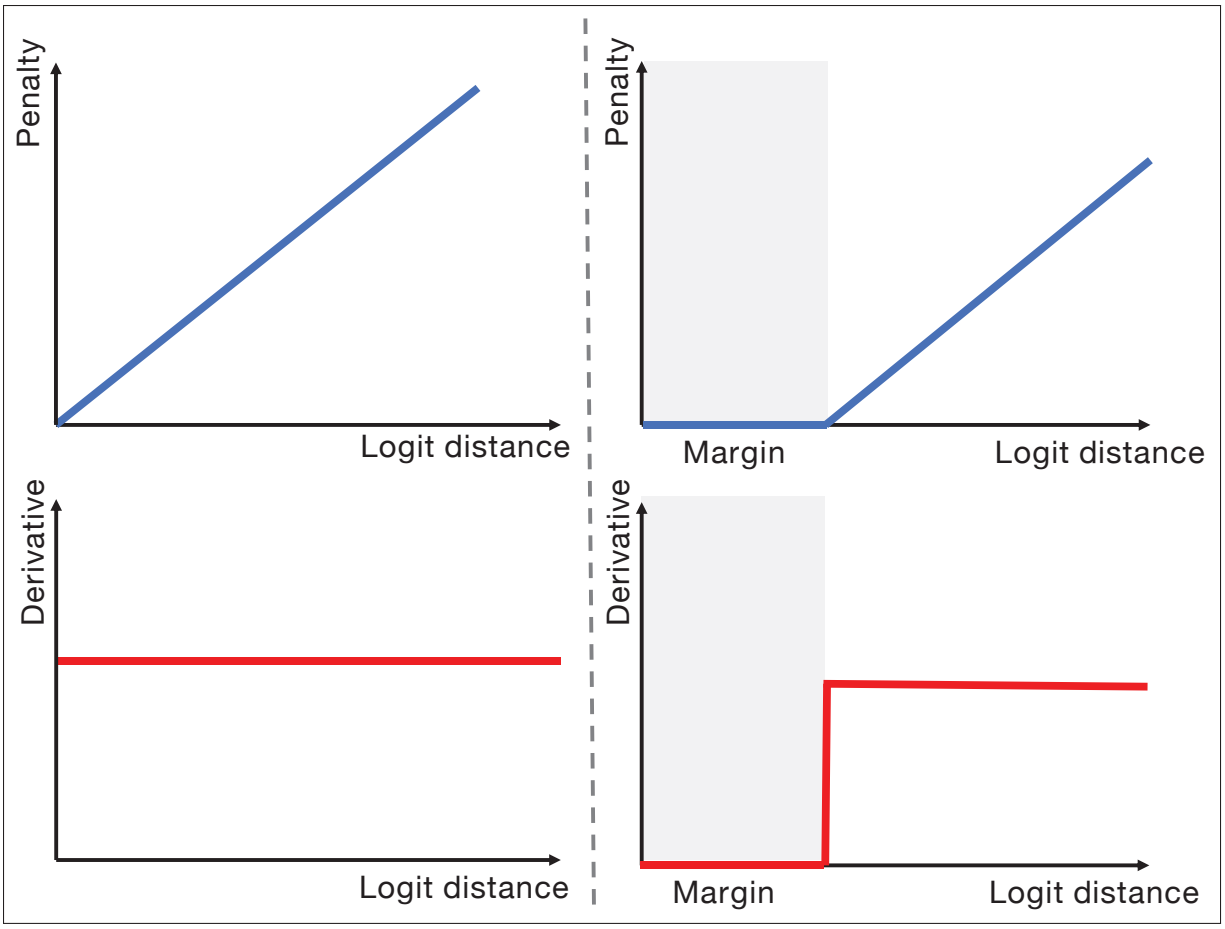

Figure 2.1 Illustration of the linear (left) and margin-based (right) penalties for imposing logit-distance constraints, along with the corresponding derivatives. Note that while the derivative of the linear penalty for constraint $\mathbf{d}(\mathbf{l}) = \mathbf{0}$ constantly pushes towards the trivial solution $s_k = \frac{1}{K} \forall K$ (i.e., LS, FL and EPC), the derivative of the proposed model only pushes towards zero those logits above the given margin

optimizing the main learning objective:

$$\min \quad \mathcal{L}_{\text{CE}} \quad \text{s.t.} \quad \mathbf{d}(\mathbf{l}) \leq \mathbf{m}, \quad \mathbf{m} > \mathbf{0} \tag{2.10}$$

The intuition behind adding a strictly positive margin m is that, unlike the linear penalty for constraint $\mathbf{d}(\mathbf{l}) = \mathbf{0}$ (Figure 2.1, left), the gradient is back-propagated only on those logits where the distance is above the margin (Figure 2.1, right). This contrasts with the linear penalty, for

which there exists always a gradient, and its value is the same across all the logits, regardless of their distance.

Even though the constrained problem in Eq. 2.10 could be solved by a Lagrangian-multiplier algorithm, we resort to a simpler unconstrained approximation by ReLU function:

$$\min \quad \mathcal{L}_{\text{CE}} + \lambda \sum_{k} \max(0, \max_{j}(l_j) - l_k - m) \tag{2.11}$$

Here, the non-linear ReLU penalty for inequality constraint $\mathbf{d}(\mathbf{l}) \leq \mathbf{m}$ discourages logit distances from surpassing a given margin m, and λ is a trade-off weight balancing the two terms. It is clear that, as discussed in Sec. 2.4, several competitive calibration methods could be viewed as approximations for imposing constraint $\mathbf{d}(\mathbf{l}) = \mathbf{0}$ and, therefore, correspond to the special case of our method when setting the margin to $m = 0$. Our comprehensive experiments in the next section demonstrate clearly the benefits of introducing a strictly positive margin m.

Note that our model in Eq. 2.11 has two hyper-parameters, m and λ. We fixed λ to 0.1 in our experiments for all the benchmarks, and tuned only the margin m over validation sets. In this way, when comparing with the existing calibration solutions, we use the same budget of hyper-parameter optimization (m in our method vs. α in LS or γ in FL).

2.5 Experiments

2.5.1 Experimental Setting

2.5.1.1 Datasets

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

2.5.1.1.1 Automated Cardiac Diagnosis Challenge (ACDC)(Bernard *et al.*, 2018)

This dataset consists of 100 patient exams containing cardiac MR volumes and its respective multi-class segmentation masks for both diastolic and systolic phases. The segmentation mask contains four classes, including the left ventricle (LV), right ventricle (RV), myocardium (Myo) and background. Following the standard practices on this dataset, 2D slices are extracted from the available volumes and resized to 224×224. Last, the dataset is randomly split into independent training (70), validation (10) and testing (20) sets.

2.5.1.1.2 Brain Tumor Segmentation (BRATS) 2019 Challenge

(Menze *et al.*, 2015; Bakas *et al.*, 2017, 2018) The dataset contains 335 multi-modal MR scans (FLAIR, T1, T1-contrast, and T2) with their corresponding Glioma segmentation masks. The classes representing the mask include tumor core (TC), enhancing tumor (ET) and whole tumor (WT). Each volume of dimension 155×240×240 is resampled and slices containing only background are removed from the training. The patient volumes are randomly split to 235, 35, 65 for training, validation, and testing, respectively.

2.5.1.1.3 MRBrainS18 (Mendrik *et al.*, 2015b)

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

2.5.1.1.4 Fast and Low GPU memory Abdominal oRgan sEgmentation (FLARE) Challenge (Ma *et al.*, 2021b)

The dataset contains 360 volumes of multi-organ abdomen CT including liver, kidneys, spleen and pancreas and their corresponding pixel-wise masks. The different resolutions are resampled

to a common space and cropped to 192×192×30. The volumes are then randomly split to 240 for training, 40 for validation, 80 for testing.

2.5.1.1.5 PROMISE

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

2.5.1.1.6 HIPPOCAMPUS (HPC)

(Antonelli *et al.*, 2022) : The data set consists of 260 MRI images acquired at the Vanderbilt University Medical Center, Nashville, US. This data set was selected due to the precision needed to segment such a small object in the presence of a complex surrounding environment. T1-weighted MPRAGE was used as the imaging sequence. The corresponding target ROIs were the anterior and posterior of the hippocampus, defined as the hippocampus proper and parts of the subiculum. The data is split to 185, 25, 50 for training, validation, and testing, respectively.

2.5.1.1.7 Breast UltraSound Images (BUSI)

(Al-Dhabyani, Gomaa, Khaled & Fahmy, 2020) The datasets consists of ultrasound images of normal, benign and malignant cases of breast cancer along with the corresponding segmentation maps. We use only benign and mailgnant images, which results in a total of 647 images resized to a resolution of 256 × 256 to benchmark our results. We considered 445 images for training, 65 images for validation, and the remaining 137 images for testing.

Note that in all datasets, images are normalized to be within the range [0-1]. Furthermore, for the datasets containing multiple image modalities (i.e., MRBrainS and BRATS), all available

modalities are concatenated in a single tensor, which is fed to the input of the neural network. In addition, there exists one dataset for which the low amount of available images impeded us to generate a proper training, validation and testing split (MRBrainS). In this case, we performed leave-one-out-cross-validation in our experiments, whereas the other datasets followed standard training, validation and testing procedures, using a single split in the experiments.

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

$$ECE = \sum_{i=1}^{M} \frac{|B_i|}{N} |A_i - C_i| \tag{2.12}$$

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

$$CECE = \sum_{i=1}^{M} \sum_{j=1}^{K} \frac{|B_{i,j}|}{N} |A_{i,j} - C_{i,j}| \tag{2.13}$$

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

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

Table 2.1 The discriminative performance (DSC and ASD) obtained by the different models across seven popular medical image segmentation benchmarks. Best method is highlighted in bold, whereas second best approach is underlined

Dataset	Region	CE + DICE		FL		ECP		LS		SVLS		Ours	
		DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓
ACDC	RV	79.8	0.75	71.4	1.27	75.4	0.87	81.5	0.68	64.2	1.86	**86.6**	**0.42**
	MYO	79.5	0.46	73.4	0.61	75.1	0.53	80.5	0.42	66.4	1.79	**84.5**	**0.37**
	LV	88.8	0.35	84.6	0.47	83.9	0.41	88.6	0.32	79.5	1.21	**91.3**	**0.29**
	Mean	82.7	0.52	76.4	0.78	78.2	0.60	83.5	0.48	70.1	1.62	**87.5**	**0.36**
MRBrainS	GM	75.7	0.48	77.3	0.53	79.3	0.47	74.5	0.51	75.3	0.49	**80.0**	**0.39**
	WM	76.1	0.66	80.4	0.60	81.0	0.55	72.7	0.97	67.0	1.06	**83.1**	**0.46**
	CSF	78.0	0.46	79.3	0.40	80.3	0.39	77.2	0.46	**81.0**	0.39	80.7	**0.38**
	Mean	76.6	0.54	79.0	0.51	80.2	0.47	74.8	0.64	74.4	0.65	**81.3**	**0.41**
FLARE	Liver	94.2	0.43	95.1	**0.37**	95.2	0.56	95.2	1.44	94.9	1.47	**95.3**	1.52
	Kidney	94.1	0.37	94.6	0.32	**95.0**	**0.31**	94.7	0.38	94.6	0.40	94.5	0.35
	Spleen	90.4	0.61	92.4	0.55	92.4	0.68	**94.2**	0.38	93.2	0.56	94.0	**0.38**
	Pancreas	63.4	**1.41**	62.5	1.65	**64.9**	1.47	63.6	1.56	63.6	1.53	64.5	1.42
	Mean	85.5	**0.71**	86.2	0.72	86.9	0.75	86.9	0.94	86.6	0.99	**87.1**	0.92
BRATS	TC	74.6	4.98	85.4	3.13	83.4	2.63	80.7	2.96	76.3	3.48	**85.6**	**2.24**
	ET	72.9	3.23	79.9	2.58	78.3	1.81	77.3	1.59	74.9	2.31	**81.1**	1.62
	WT	85.4	2.93	88.9	2.72	88.9	2.41	87.9	2.48	88.8	2.27	**89.5**	**2.11**
	Mean	77.6	3.71	84.8	2.81	83.6	2.28	81.9	2.34	79.8	2.69	**85.4**	**1.99**
PROMISE	Prostate	0.751	1.17	72.9	1.42	73.6	1.27	71.3	1.72	76.6	1.27	**77.0**	**0.95**
	Tumor	32.8	4.10	36.1	3.35	34.4	2.48	35.0	3.29	39.6	**2.16**	**39.7**	2.34
	Mean	54.0	2.63	54.5	2.39	54.0	1.88	53.2	2.50	58.1	1.71	**58.3**	**1.64**
HPC	Anterior	87.4	0.47	87.9	0.46	87.4	0.49	87.9	0.49	**88.3**	**0.46**	87.6	0.49
	Posterior	85.7	0.43	85.2	0.48	85.3	0.45	85.7	**0.42**	84.9	0.48	85.7	0.43
	Mean	86.7	**0.45**	86.5	0.47	86.4	0.47	**86.8**	0.45	86.6	0.47	86.7	0.46
BUSI	Tumor	**70.9**	**13.1**	68.8	13.9	67.7	15.1	67.9	15.6	67.9	14.6	68.5	13.7

Table 2.2 The calibration performance (ECE and CECE) obtained by the different models across five popular medical image segmentation benchmarks. Best method is highlighted in bold, whereas second best approach is underlined. ∇ indicates the difference between the best model and our approach

Dataset	CE + DICE		FL		ECP		LS		SVLS		Ours	
	ECE ↓	CECE ↓	ECE ↓	CECE ↓	ECE ↓	CECE ↓	ECE ↓	CECE ↓	ECE ↓	CECE ↓	ECE ↓	CECE ↓
ACDC	0.137	0.084	0.113	0.116	0.109	0.095	0.081	0.107	0.176	0.135	**0.061**	**0.069**
MRBrainS	0.172	0.102	**0.020**	0.064	0.048	0.068	0.036	0.085	0.060	0.080	0.050	**0.058**
FLARE	0.058	0.034	**0.033**	0.035	0.037	**0.027**	0.055	0.050	0.039	0.036	0.038	0.028
BRATS	0.178	0.122	**0.097**	0.119	0.132	0.091	0.112	0.108	0.151	0.122	0.101	0.093
PROMISE	0.430	0.304	0.247	0.298	0.306	0.252	0.280	0.299	0.344	0.271	**0.232**	**0.237**
HPC	0.069	**0.079**	0.042	0.108	0.066	0.093	0.061	0.109	0.044	0.104	**0.033**	0.088
BUSI	0.250	0.278	0.220	0.305	0.237	0.365	0.229	0.333	0.226	0.305	**0.193**	**0.274**

2.5.1.2 Implementation Details

To empirically evaluate the proposed model, we conduct experiments comparing a state-of-the-art segmentation network on a multi-class scenario trained with different learning objectives. In particular, we first include standard loss functions employed in medical image segmentation, which include the common Cross-entropy (CE) and the duple composed by CE and DSC losses. Furthermore, we also include training objectives which have been proposed to calibrate neural networks, which represent nowadays the state-of-the-art for this task. This includes Focal loss (FL) (Lin *et al.*, 2017), Label Smoothing (LS) (Szegedy *et al.*, 2016) and ECP (Pereyra *et al.*, 2017). Last, we also compare to the recent Spatially-Varying LS (SVLS) (Islam & Glocker, 2021), which demonstrated to outperform the simpler LS version in the task of medical image segmentation. Following the literature, we have chosen the commonly used hyper-parameters and considered the values which provided the best DSC on the validation dataset. For FL, γ values of 1, 2, and 3 are considered. In case of ECP and LS, α and λ values of 0.1, 0.2, 0.3 are used. For our method, we considered the margins to be 5, 8, and 10. In the case of SVLS, the one-hot label smoothing is performed with a kernel size of 3. For the experiments, we fixed the batch size to 4, epochs to 100, and optimizer to ADAM. The learning rate of 1e-3 and 1e-4 are used for the first 50 epochs, and the next 50 epochs respectively. The models are trained on 2D slices, while the evaluation is done over 3D volumes.

Backbones. The main experiments are conducted on the popular UNet (Ronneberger *et al.*, 2015b). Nevertheless, to show the versatility of the proposed margin based label smoothing,

we have evaluated our model on other popular architectures in medical image segmentation including AttUNet (Oktay *et al.*, 2018), TransUNet (Chen *et al.*, 2021), and UNet++ (Zhou *et al.*, 2020).

2.5.2 Results

2.5.2.1 Main results

The discriminative quantitative results obtained by the proposed model, as well as prior literature, are reported in Table 3.1. We observe that across the different datasets, our model consistently achieves competitive discriminative performance, typically ranking as the best or second-best model in both region-based (i.e., DSC) and distance-based (i.e., ASD) metrics. This demonstrates that our method yields not only a better identification of target regions, but also an improvement in the boundary regions, highlighted by lower ASD values. An interesting observation is that, while other learning objectives typically result in performance gains compared to the standard CE loss, their superiority over the others depends on the selected dataset.

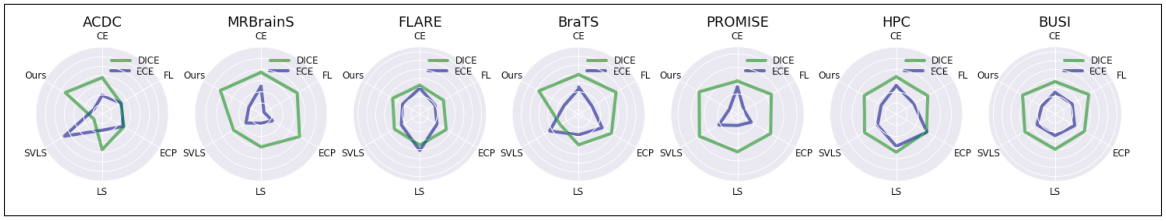

Figure 2.2 **Compromise between calibration and discriminative performance.** In order to get the best performance, we expect a model to achieve large DSC (*in green*) and small ECE (*in blue*) values

Table 3.2 summarizes the calibration performance, in terms of ECE and CECE of all the analyzed models. We can observe that, similar to the discriminative performance reported earlier, the proposed model typically ranks as best or second best method. An interesting observation is that, according to the results, focal loss provides well-calibrated models (i.e., low ECE and CECE values), whereas their discriminative performance is typically far from best performing models. As exposed in our motivation, one of the reasons behind this behaviour might be the undesirable

Table 2.3 Calibration performance of post-hoc calibration methods: temperature scaling (TS) and Local Temperature Scaling (LTS) (Ding *et al.*, 2021). Best method is highlighted in bold, whereas second best approach is underlined

Dataset	Method	CE		CE + DICE		FL		ECP		LS		SVLS		Ours	
		ECE	CECE	ECE	CECE	ECE	CECE	ECE	CECE	ECE	CECE	ECE	CECE	ECE	CECE
ACDC	Pre	0.079	0.073	0.137	0.084	0.113	0.116	0.109	0.095	0.081	0.107	0.176	0.135	**0.061**	**0.069**
	TS	0.077	0.073	0.135	0.084	0.112	0.117	0.105	0.095	0.084	0.109	0.174	0.135	**0.055**	**0.065**
	LTS	0.067	0.065	0.070	0.076	0.127	0.125	0.080	0.094	0.065	0.072	0.118	0.116	**0.041**	**0.046**
FLARE	Pre	0.045	0.029	0.058	0.034	**0.033**	0.035	0.037	**0.027**	0.055	0.050	0.039	0.036	0.038	0.028
	TS	0.040	0.030	0.051	0.036	**0.030**	0.038	0.032	**0.028**	0.042	0.039	0.039	0.038	0.033	0.029
	LTS	0.033	0.030	0.044	0.038	0.065	0.048	**0.026**	0.028	0.031	0.026	0.040	0.036	0.031	**0.026**
BRATS	Pre	0.131	**0.091**	0.178	0.122	**0.097**	0.119	0.132	0.091	0.112	0.108	0.151	0.122	0.101	0.093
	TS	0.13	**0.09**	0.177	0.122	**0.097**	0.119	0.131	0.091	0.111	0.108	0.149	0.121	0.098	0.093
	LTS	0.114	**0.089**	0.156	0.121	0.097	0.119	0.117	0.09	0.105	0.119	0.131	0.121	**0.089**	0.096
PROMISE	Pre	0.411	0.334	0.430	0.304	0.247	0.298	0.306	0.252	0.280	0.299	0.344	0.271	**0.232**	**0.237**
	TS	0.408	0.334	0.429	0.304	0.245	0.299	0.303	0.251	0.279	0.298	0.342	0.271	**0.229**	**0.237**
	LTS	0.294	0.283	0.312	0.263	0.209	0.291	0.230	0.235	0.255	0.257	0.234	0.238	**0.189**	**0.217**
HPC	Pre	0.052	0.091	0.069	**0.079**	0.042	0.108	0.066	0.093	0.061	0.109	0.044	0.104	**0.034**	0.088
	TS	0.051	0.091	0.068	**0.079**	0.042	0.108	0.065	0.093	0.059	0.108	0.044	0.104	**0.033**	0.089
	LTS	0.048	0.092	0.065	**0.08**	0.041	0.108	0.061	0.094	0.059	0.108	0.043	0.103	**0.032**	0.09
BUSI	Pre	0.230	0.334	0.250	0.278	0.220	0.305	0.237	0.365	0.229	0.333	0.226	0.305	**0.193**	**0.274**
	TS	0.229	0.333	0.250	0.278	0.236	0.365	0.219	0.305	0.229	0.333	0.225	0.305	**0.193**	**0.274**
	LTS	0.207	0.328	0.210	**0.257**	0.243	0.377	0.202	0.298	0.268	0.358	0.198	0.295	**0.182**	0.275

effect of pushing all logit distances to zero. Enforcing this constraint may alleviate the problem of overconfidence in deep networks, at the cost of providing non-informative solutions.

An interesting summary of these results is depicted in Figure 2.2, where we resort to radar plots to highlight the better compromise between discriminative and calibration performance shown by our model. In particular, a *well-calibrated* model should have a balanced compromise between a high discriminative power (*green line*) and low calibration metrics (*blue line*). This means that, following these plots, the larger the gap between green and blue lines, the better the compromise between discriminative and calibration performance.

Furthermore, to have a better overview of the general performance across different models, we follow the strategy followed in several MICCAI Challenges, e.g., MRBrainS (Mendrik *et al.*, 2015a), where the final ranking is given as the sum of individual ranking metrics: $R_T = \sum_{m=0}^{|M|} r_m$, where r_m is the rank of the segmentation model for the metric m (mean)[1]. Thus, if a model ranks first in terms of DSC in the FLARE dataset, it will receive one point, whereas five points will be added in case the model ranks fifth. The final ranking is obtained after the overall scores R_T for each model are sorted in ascending order, and ranked from 1 to n. Furthermore, to account for the different complexities of each sample, we follow the mean-case-rank strategy,

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

which has been employed in other MICCAI Challenges, e.g., (Maier *et al.*, 2017). In particular, we first compute the DSC, ASD, ECE, and CECE values for each sample, and establish each method's rank based on these metrics, separately for each case. Then, we compute the mean rank over all four evaluation metrics, per case, to obtain the method's rank for that given sample. Finally, we compute the mean over all case-specific ranks to obtain the method's final rank. Figure 2.3 provides the rank comparison through heatmap visualization. It can be inferred that, for both discriminative and calibration metrics, our methods achieves the highest rank. Interestingly, the proposed loss term yields very competitive discriminative results, outperforming the popular compounded CE+DSC loss. It is noteworthy to highlight that the optimization goal of these two terms are different. Networks trained with CE tend to achieve a lower average negative log-likelihood over all the pixels, whereas using Dice as loss function should increase the discriminative performance, in terms of Dice. Thus, it is expected that the compounded loss brings the better of both worlds. Nevertheless, we can observe that this is not what happens in practice. On the one hand, the networks trained with CE+DSC loss rank among the best discriminative models (third in DSC and second in ASD). On the other hand, their calibration performance is substantially degraded, ranking last and second-last in ECE and CECE, respectively. These results align with recent findings (Mehrtash *et al.*, 2020), which highlight the deficiencies of models trained with the DSC loss to deliver well-calibrated models. While adjusting the balancing hyperparameter could improve the performance on one task, the results on the other task would likely degrade due to the different nature of both learning objectives. Thus, based on these observations, we argue that obtaining a good compromise between calibration and segmentation quality is hardly attainable with the popular CE+DSC loss, and promote our model as a better alternative. Fig. 2.3 provides a heatmap visualization to compare the methods for different metrics using mean-case-specific strategy. As observed in sum rank approach, our methods consistently achieves the best rank in both discriminative and calibration metrics. Importantly, calibration methods like FL, and LS achieve promising results with ECE, while severely compromising the DSC.

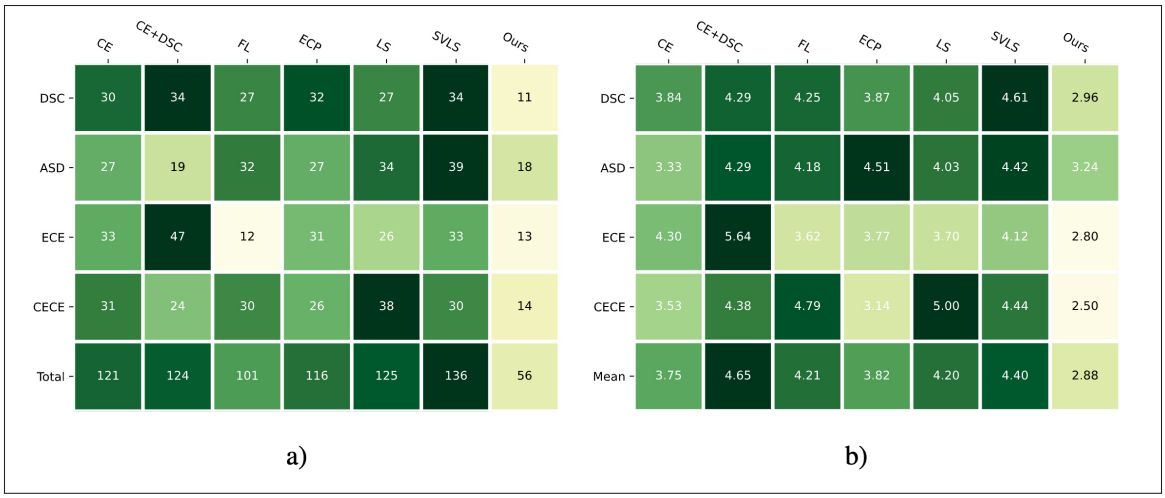

Figure 2.3 Ranking (*global* and *per-metric*) of the different methods based on the sum-rank and mean of case-specific approach

2.5.2.2 Comparison to post-hoc calibration

The proposed approach is orthogonal to post-hoc calibration strategies, which can still be used after training, as long as there exists an independent validation set to find the optimal hyperparameters (for example T in temperature scaling). To demonstrate this, we now report the performance of pre-scaling and post-scaling for ACDC, FLARE, and PROMISE datasets across the different approaches. In particular, we have included two post-hoc calibration strategies. First, we use the standard Temperature scaling approach, referred to as TS, where a single value for the entire image is employed. Furthermore, we also include the Local Temperature Scaling (LTS) method in (Ding *et al.*, 2021), which was recently proposed in the context of medical image segmentation and provides a temperature value at each image pixel. For both TS and LTS, the optimal temperature values are found by optimizing the network parameters to decrease the negative log likelihood on an independent validation set. From the quantitative comparison, which can be found in Table 2.3, it can be inferred that our method further benefits from scaling the raw softmax probability predictions. Interestingly, the calibration performance obtained by our method prior to temperature scaling still outperforms the results obtained by several other approaches even after applying LTS on their predictions. Another unexpected observation is that, under some settings, the use of temperature scaling (either TS or LTS) deteriorates the calibration

performance. We argue that this phenomenon could be due to noticeable differences between the validation and testing datasets. As empirically demonstrated in (Ovadia *et al.*, 2019), applying temperature scaling when differences between datasets exist might result in a negative impact. In addition, similar observations were reported in (Kock, Thielke, Chlebus & Meine, 2021), where the calibration performance of segmentation models on several datasets was degraded after applying temperature scaling.

2.5.2.3 Effects of logit margin constraints

In our motivation, we hypothesized that the suboptimal supervision delivered by CE in multi-class scenarios might likely result in poorly calibrated models, as the posterior probability assigned to each of the non-true classes cannot be directly controlled. Indeed, it is expected that by minimizing the CE the softmax vectors are pushed towards the vertex of the probability simplex. This implies that the distances between the largest logit and the rest are magnified, resulting in overconfident models. To validate this hypothesis, and to empirically demonstrate that our proposed term can alleviate this issue, we plot the average logit distributions across classes on two datasets. In particular, we first separate all the voxels based on their ground truth labels. Then, for each category group, we average the per-voxel vector of logit predictions for both CE and the proposed model, whose results are depicted in Figure 2.4. First, we can observe that a model trained with CE indeed tends to provide large logit differences, which intensifies overconfidence predictions. Furthermore, while the mean logit value of the target class is considerably large and greatly differs from the largest value across other categories, the differences with the remaining logits –from non-target classes– remain uncontrolled. In contrast, we can clearly observe the impact on the logit distribution when we include the proposed term into the learning objective. In particular, our margin-based term *i)* promotes similar values of the true class logit across classes and *ii)* encourages more equidistant logits between this and the remaining classes, which implicitly constraints the logit values of untargeted classes to be very close (mimicking a uniform distribution). These results empirically validate our hypothesis in regards of the weaknesses of CE and the benefits brought by our approach.

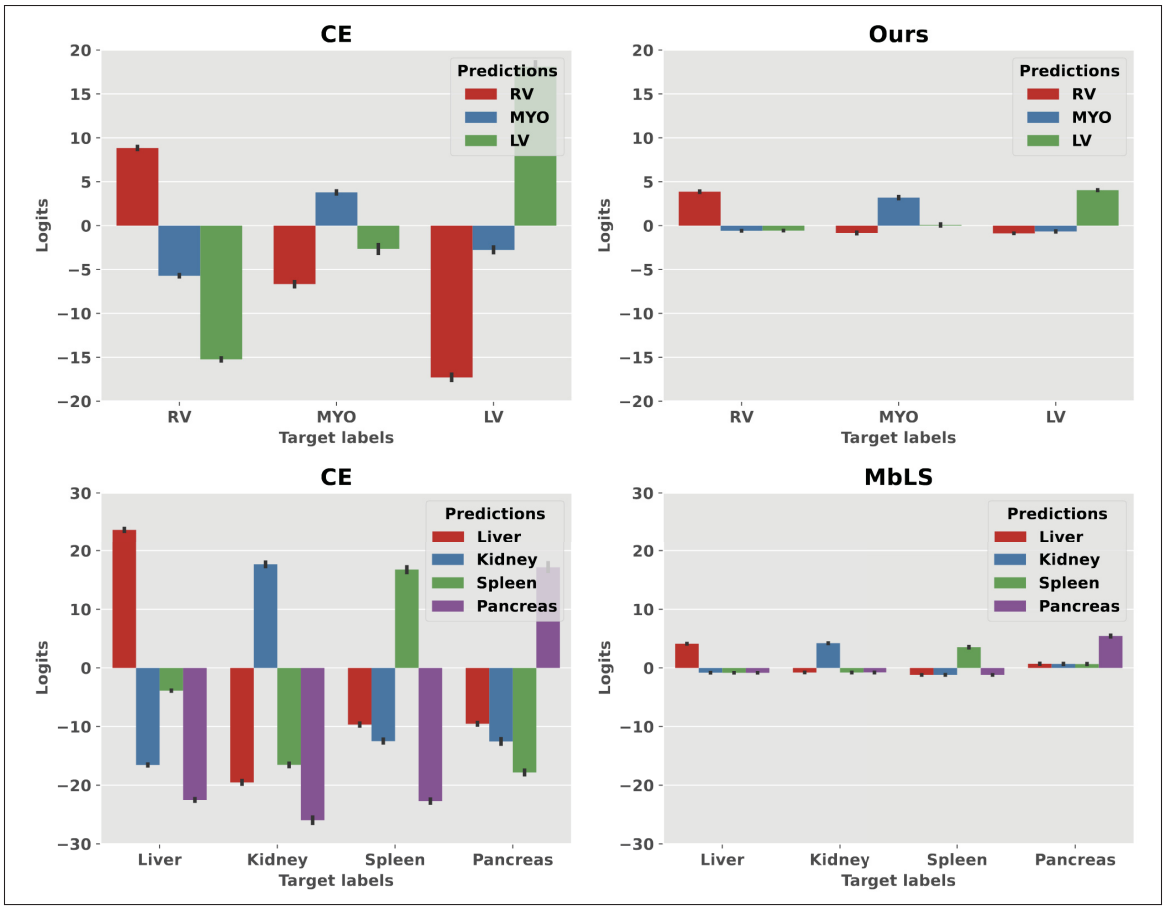

Figure 2.4 **Adopting the proposed term during training *substantially* reduces the logit distances, producing less overconfident predictions.** These plots depict the average predicted logit distributions for each target class –based on the ground truth– on ACDC (*top*) and FLARE (*bottom*) datasets when the model is trained with CE (*left*) and the proposed loss (*right*)

2.5.2.4 Calibration and discriminative performance under distribution shift

There have been recent empirical studies (Ovadia *et al.*, 2019; Minderer *et al.*, 2021) on the robustness of calibration models under distribution shift. In particular, (Minderer *et al.*, 2021) explores out-of-distribution calibration by resorting to ImageNet-C (Hendrycks & Dietterich, 2018), a computer vision dataset that contains images that have been synthetically corrupted, for example by including Gaussian noise. Inspired by these works, we now assess the robustness of our model in the presence of domain drift. To do this, we added Gaussian noise to the images

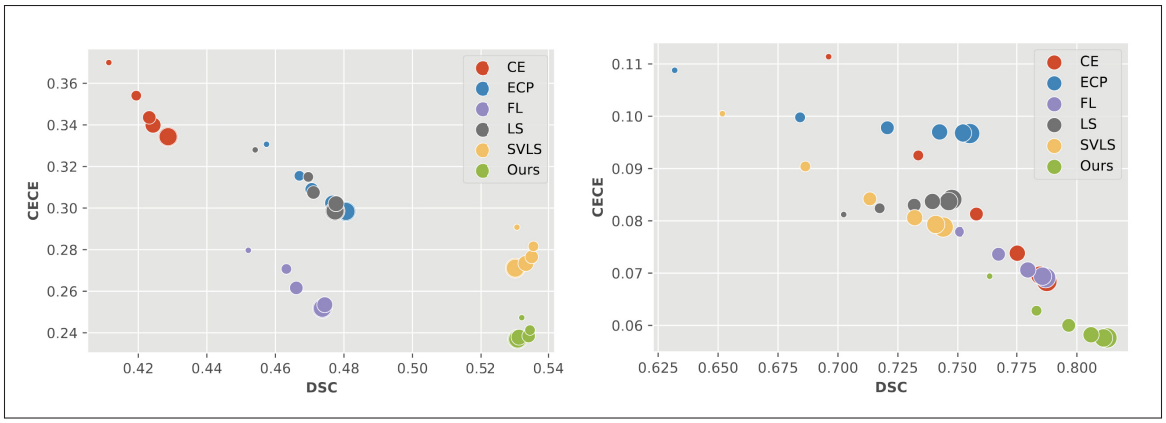

Figure 2.5 Robustness to distributional drift on PROMISE (*left*) and MRBrainS (*right*) datasets. Note that larger circles represent lower sigma values for the Gaussian noise corruptions

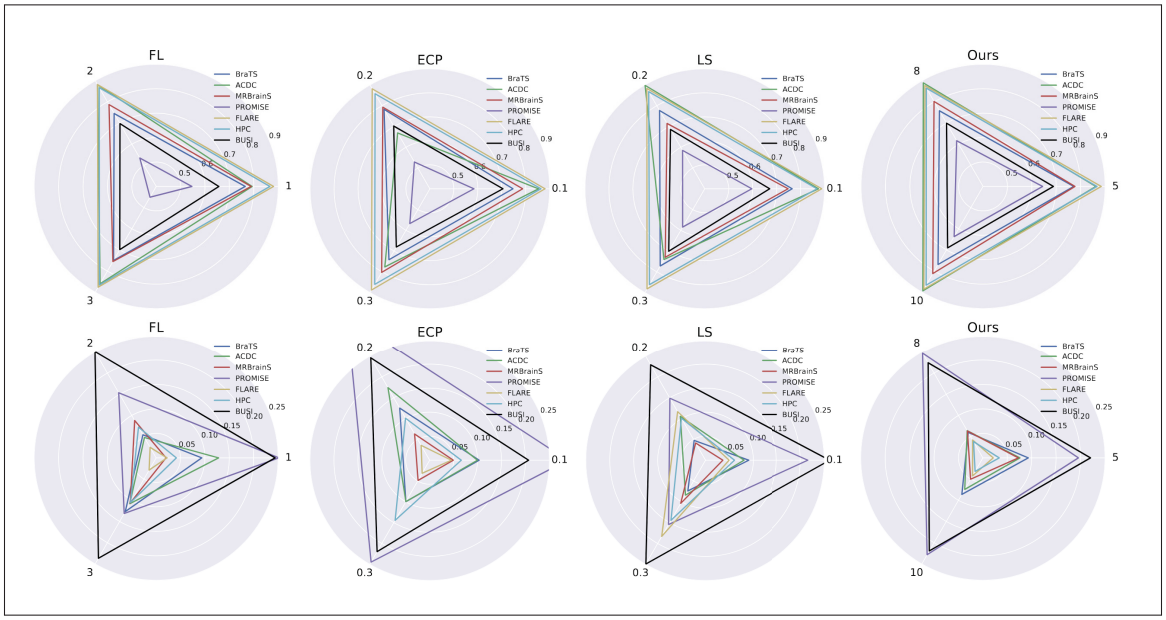

Figure 2.6 **Sensibility to hyperparameters across datasets.** For each method, we use the standard hyperparameters used in the literature and compare its variation across different datasets. The discriminative performance (DSC) is reported in the *top row*, whereas the calibration analysis (ECE) is depicted in the *bottom row*

on the testing set, with sigma values ranging from 0 to 0.05 with an increment of 0.01. From the plots in Figure 2.5 we can clearly observe that models trained with our objective function are less sensitive to noise, compared to prior state-of-the-art methods. More concretely, on the

PROMISE dataset, the discriminative and calibration performance of our approach remains almost invariant to image perturbations with different levels of Gaussian noise. Furthermore, while the results obtained by our method in the MRBrainS data are affected by noise, its performance degradation is significantly lower than nearly all previous approaches, being on par with the focal loss. Nonetheless, it is noteworthy to mention that despite the relative decrease in performance is similar between the proposed method and FL, their global performance differences are substantially large (e.g., 6-8% difference in DSC). Based on these results we can argue that the proposed method delivers higher performing models that are, in addition, more robust to distributional drifts produced by Gaussian noise.

2.5.2.5 On the impact of hyperparameters

We now assess the sensitivity of each model to the choice of the hyperparameters on each dataset. We stress that, for each method, we have selected a range of common values used in the literature. In particular, γ is set to 1.0, 2.0 and 3.0 in Focal loss, λ is fixed to 0.1, 0.2 and 0.3 in ECP and Label smoothing, whereas the margin values in our method are set to 5.0, 8.0 and 10. The discriminative (DSC) and calibration (ECE) performances obtained across the different hyperparameter values are depicted in Fig. 2.6. From this figure, it can be easily observed that, while prior approaches are very sensitive to the value of their balancing term, our method is significantly more robust to these changes. For example, the discriminative performance is drastically affected in both ECP and LS across several datasets when changing the value of the balancing term from 0.1 and 0.2 to 0.2 and 0.3, respectively. On the other hand, this phenomenon is more pronounced in the calibration metrics, where FL, ECP and LS show much higher variations than the proposed approach. A potential drawback that can be extracted from these findings is that, in order to get a well calibrated and high performance model, prior approaches might require multiple training iterations to find a satisfactory compromise. Furthermore, we believe that these large variations indicate that differences in the data –e.g., image contrast, target size and heterogeneity, or class distribution– might have a different impact on these losses, entangling the convergence of models trained with these terms.

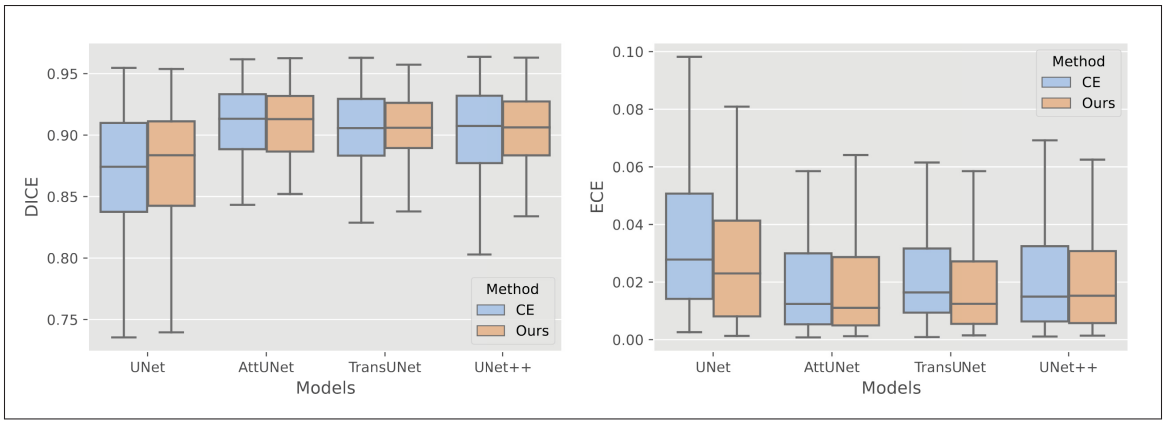

Figure 2.7 **Robustness to segmentation backbone**, which evaluates the standard cross-entropy and the proposed model on the FLARE segmentation benchmark

2.5.2.6 Robustness to backbone

In this experiment, we evaluate the proposed loss on several standard medical image segmentation networks, including: AttUNet (Oktay *et al.*, 2018), TransUNet (Chen *et al.*, 2021), and UNet++ (Zhou *et al.*, 2020). For this study, we consider the FLARE dataset due to its larger number of classes. The quantitative comparison of CE and our method for these backbones is presented in Fig. 2.7, from which it can be inferred that, irrespective of the backbone used, our method is capable of consistently achieving better model calibration compared to the popular cross-entropy loss, while yielding at par performance in the discrimination task. We can therefore say that the proposed term can be directly plugged into any segmentation network, and the improvement observed is consistent across different models.

2.5.2.7 Qualitative results and reliability diagrams

Figure 2.8 depicts the predicted segmentation masks (*top*), confidence maps (*middle*) and their corresponding reliability plots (*bottom*) on one subject across the different methods. While the segmentation masks reveal several differences in terms of discriminative performance, the confidence maps present more interesting observations. Note that, as highlighted in prior works (Liu *et al.*, 2022a), better calibrated models should show better edge sharpness, matching the

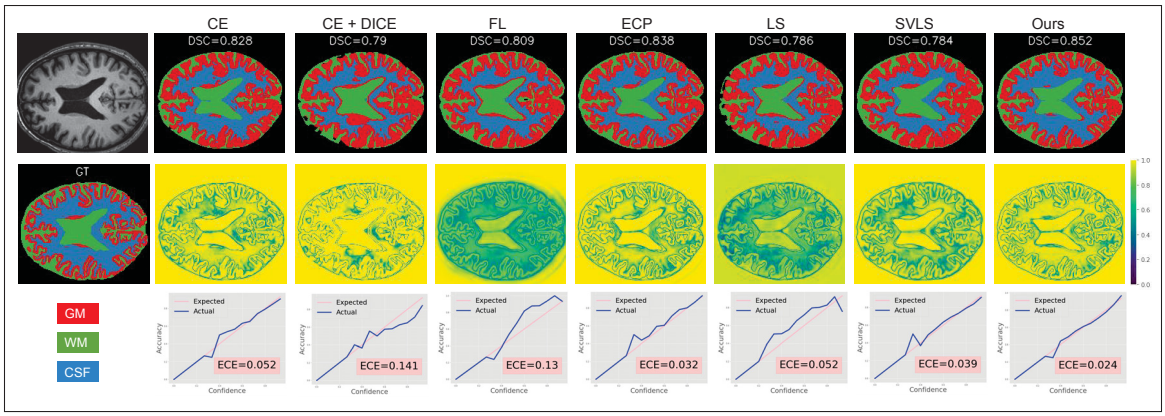

Figure 2.8 Qualitative results on MRBrainS dataset for different methods. In particular, we show the original image and the corresponding segmentation masks provided by each method (*top row*), the ground-truth (GT) mask followed by maximum confidence score of each method (*middle row*) and the respective reliability plots (*bottom row*). Methods from left to right: CE, CE+DICE, FL, ECP, LS, SVLS, Ours

expected property that the model should be less confident at the boundaries, whereas yielding more confident predictions in inner target regions. First, we can observe that adding the DSC loss term substantially degrades the confidence map compared to its single CE loss counterpart. In particular, the CE+DSC compounded loss tends to produce smoother edges, in terms of confidence, which is derived from worst calibrated models. Furthermore, while it can increase the confidence of predictions in several inner object regions, it decreases this score in others. In addition, we can clearly observe that the remaining analyzed methods provide a diverse span of confidence estimates, with several models providing highly unconfident inner regions (e.g., FL (Mukhoti *et al.*, 2020b) and LS (Szegedy *et al.*, 2016)). In contrast, our method yields confidence estimates that are sharp in the edges and low in within-region pixels, as expected in a well-calibrated model. These visual findings are supported by the reliability diagrams. Indeed, our model yields the best reliability diagram, as the ECE curves are closer to the diagonal, indicating that the predicted probabilities serve as a good estimate of the correctness of the prediction.

2.5.2.8 Choice of the penalty

In this work we have presented a unified constrained optimization perspective of existing calibration methods, showing that they can be seen as approximations of a linear penalty for imposing the constraint $\mathbf{d(l) = 0}$. In order to show that the improvement of the proposed formulation comes from the relaxation of this constraint, which has important limitations, we selected a linear penalty, similarly to the implicit underlying mechanism of LS, FL and ECP. Nevertheless, in this section we now address the question of whether we can further improve these results by employing other penalties to enforce the proposed constraint. In particular, we evaluate both the discriminative and calibration performance of our model when a quadratic penalty, i.e., L_2, is used to impose the constraint in Eq. 2.10. Results in Table 2.4 show that, even though both penalties behave similarly in terms of segmentation, the model trained with a quadratic penalty is typically worse calibrated. We argue that these differences might be due to the more aggressive behaviour of quadratic penalties when the constraint is not satisfied, which may eventually lead to near-to-uninformative solutions, similar to those obtained by FL, LS and ECP. Furthermore, we would like to highlight that in this experiment the margin m was fine-tuned for the L_2 penalty, whereas its controlling weight remained the same as in the L_1 term. Thus, further optimizing the penalty weight might alleviate its aggressive performance on large violations, potentially leading to superior performance of the proposed MbLS loss when the constraint is enforced via a quadratic penalty.

Table 2.4 Quantitative comparison across datasets of different penalty terms to impose the constraint $\mathbf{d(l) \leq m}$

	L1		L2	
	DSC	ECE ↓	DSC	ECE ↓
ACDC	0.875	0.061	0.874	0.064
FLARE	0.871	0.038	0.868	0.031
BRATS	0.854	0.101	0.845	0.101
PROMISE	0.583	0.232	0.549	0.279
HPC	0.867	0.033	0.864	0.042
BUSI	0.685	0.193	0.673	0.197

2.5.2.9 Impact of MbLS on the CE + Dice loss

The main goal of this work is to present existing calibration losses from a constrained optimization perspective, highlight their weaknesses and propose a potential solution to overcome the identified limitations. Furthermore, we stress that existing calibration losses do not include compounded losses that integrate segmentation terms, such as the Dice loss. Nevertheless, given the popularity of this joint learning objective in medical image segmentation, we assess in this section the impact of integrating the proposed constrained term into the duple CE + DSC. In particular, to better understand the impact of the proposed MbLS loss, as well as DSC loss, we depict the results for the standard CE, CE + DSC, MbLS and MbLS + DSC in Fig 2.9. The stacked normalized plots show that while these methods result in similar discriminative performance, the differences in calibration are more noticeable. More concretely, and as we demonstrated empirically in Table 2.2, models trained with the joint CE + DSC loss see their calibration performance degrade compared to the standard CE objective. By coupling the proposed approach with the DSC loss ($MbLS + DSC$) this degradation can be reduced thanks to the proposed penalty term. Nevertheless, the ECE results achieved by this joint term are significantly higher than those obtained by the proposed approach, which does not include the DSC loss. These findings demonstrate that *i*) the proposed MbLS can improve the calibration performance of the popular CE + DSC segmentation loss, and *ii)* despite the improvement in discrimination performance, losses integrating the DSC loss term present significant deficiencies to deliver well-calibrated models.

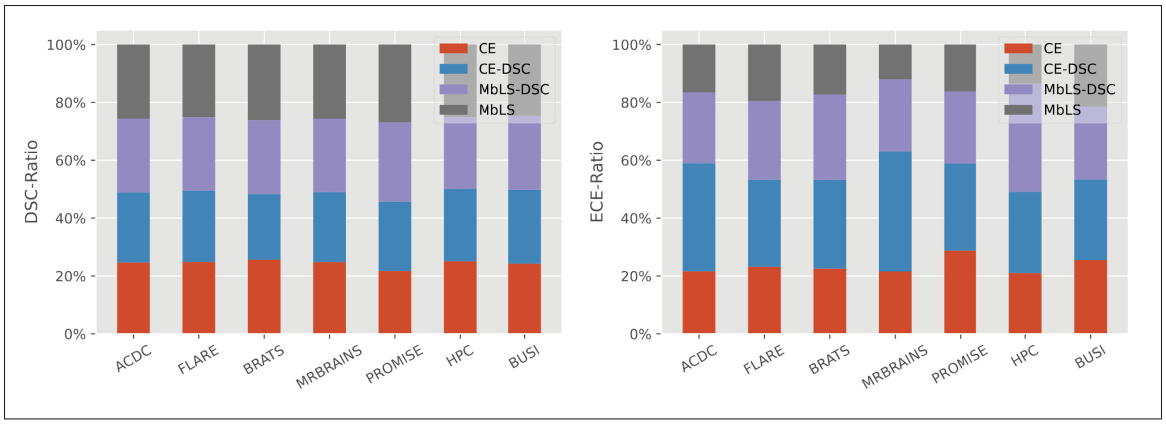

Figure 2.9 **Impact of MbLS on the DSC loss.** Normalized stacked bar plots to assess the impact of the proposed MbLS on the popular CE+DSC segmentation loss. Discriminative performance in terms of DSC is depicted in the *top* (the higher the ratio the better), whereas calibration is assessed in terms of ECE in the *bottom* (the lower the ratio the better)

2.6 Conclusion

Despite the popularity of network calibration in a broad span of applications, the connection between state-of-the-art calibration losses remains unexplored, and their impact on segmentation networks, particularly in the medical field, has largely been overlooked. In this work, we have demonstrated that these popular losses are closely related from a constrained optimization perspective, whose implicit or explicit constraints lead to non-informative solutions, preventing the model predictions to reach the best compromise between discriminative and calibration performance. To overcome this issue, we proposed a simple solution that integrates an inequality constraint into the main learning objective, which imposes a controlled margin on the logit distances. Through an extensive empirical evaluation, which contains multiple popular segmentation benchmarks, we have assessed the discriminative and calibration performance of state-of-the-art calibration losses in the important task of medical image segmentation. The results highlight several important benefits of the proposed loss. First, it achieves consistent improvements over state-of-the-art calibration and segmentation losses, both in terms of discriminative and calibration performance. Second, the proposed model is much less sensitive to hyperparameters changes compared to prior losses, which reduces the training time to find a

satisfactory compromise between discrimination and calibration tasks. In addition, the empirical observations support our hypothesis that the suboptimal supervision delivered by the standard cross-entropy loss likely results in poorly calibrated models, as model trained with this loss tend to produce largest logit differences. Thus, we advocate that the proposed loss term should be preferred to train models that provide higher discriminative performance, while yet delivering accurate uncertainty estimates.

CHAPTER 3

NEIGHBOR-AWARE CALIBRATION OF SEGMENTATION NETWORKS WITH PENALTY-BASED CONSTRAINTS

Balamurali Murugesan[a] , Sukesh Adiga Vasudeva[a] , Bingyuan Liu [a] , Herve Lombaert[a] , Ismail Ben Ayed[b] , Jose Dolz[a]

[a] Department of Software Engineering, École de Technologie Supérieure,
[b] Department of System Engineering, École de Technologie Supérieure,
1100 Notre-Dame West, Montreal, Quebec, Canada H3C 1K3

Paper published in Medical Image Analysis, April 2025

Presentation

This chapter presents the article *"Neighbor-aware calibration of segmentation networks with penalty-based constraints"* (Murugesan *et al.*, 2025) published in Medical Image Analysis (**MedIA**), Volume 101, Page 103501, 2025. The preliminary version (Murugesan *et al.*, 2023a) of the article was presented in International Conference on Medical Image Computing and Computer Assisted Intervention (**MICCAI**) Page. 572–581 2023.

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

3.1 Introduction

Despite the remarkable progress made by deep neural networks (DNNs) in a wide span or 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 & Ghahramani, 2016; Guo *et al.*, 2017b). 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.*, 2017b). Indeed, modern classification networks trained under the fully supervised learning paradigm resort to binary one-hot encoded vectors as supervisory 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.*, 2017b; Tomani *et al.*, 2021a; Zhang *et al.*, 2020c; 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.*, 2020b; Müller *et al.*, 2019b) or explicitly integrated as additional learning objectives (Pereyra *et al.*, 2017; Liu *et al.*, 2022b; Liu, Rony, Galdran, Dolz & Ben Ayed, 2023a).

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 & 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.*, 2020b; Müller *et al.*, 2019b; Liu *et al.*, 2022b). 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 & 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 & 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.

3.2 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.*, 2017b; Zhang *et al.*, 2020c; Tomani *et al.*, 2021a). 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.*, 2017b), 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.*, 2021a) proposes to transform the validation set before transforming the softmax distributions, whereas (Zhang *et al.*, 2020c) 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.*, 2019b; Mukhoti *et al.*, 2020b) 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.*, 2019b). In addition, (Mukhoti *et al.*, 2020b) 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.*, 2022b) 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.*, 2019b), little attention has been devoted to improving both the calibration and segmentation performance of deep models in the medical domain. (Jena & 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 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

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

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 & 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.*, 2022b), 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 & 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.

3.3 Methodology

3.3.1 Preliminaries

Notation. Let us denote the training dataset as $\mathcal{D} = \{(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)})\}_{n=1}^{N}$, 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, $\boldsymbol{x}^{(n)} \in \mathbb{R}^{\Omega_n}$ represents the n^{th} image, Ω_n the spatial image domain, and $\boldsymbol{y}^{(n)} \in \mathbb{R}^K$ 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 $\boldsymbol{x}$ will denote any image in the training set. Now, given an input image $\boldsymbol{x}$, a neural network parameterized by θ generates the set of logit predictions $f_\theta(\boldsymbol{x}) = \mathbf{l} \in \mathbb{R}^{\Omega_n \times K}$. Last, we use the softmax function, denoted as $\phi(\cdot)$ to obtain the predicted model probabilities $\hat{\boldsymbol{p}} = \phi(f_\theta(\boldsymbol{x})) \in \mathbb{R}^{\Omega_n \times K}$.

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:

$$\mathbb{P}(\hat{y} = y|\hat{p} = p) = p, \quad \forall p \in [0, 1], \tag{3.1}$$

where $\hat{y} = \arg\max(\hat{\boldsymbol{p}})$ is the predicted class of input image $\boldsymbol{x}$, and $\hat{p} = \max(\hat{\boldsymbol{p}})$ its associated confidence. Equation 3.1 tells us that, to be perfectly calibrated, when the model predicts the probability distribution $\phi(f_\theta(X))$ over the set of classes $[K] = \{1, 2, ..., K\}$, 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*.

3.3.2 A constrained optimization perspective of SVLS

Spatially Varying Label Smoothing (SVLS) (Islam & 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 probability in a given pixel is denoted as $\hat{\mathbf{p}} = [\hat{p}_0, \hat{p}_1, ..., \hat{p}_{k-1}]$. Let us now transform the surrounding patch of the segmentation mask around a given pixel into a unidimensional vector $\mathbf{y} \in \mathbb{R}^{\mathbf{d}}$, 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 $\mathbf{w} \in \mathbb{R}^{\mathbf{d}}$. Thus, for a given pixel j, and a class k, SVLS (Islam & Glocker, 2021) can be defined as:

$$\tilde{y}_j^k = \frac{1}{|\sum_i^d w_i|} \sum_{i=1}^{d} y_i^k w_i. \tag{3.2}$$

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

$$\mathcal{L} = -\sum_k \left(\frac{1}{|\sum_i^d w_i|} \sum_{i=1}^{d} y_i^k w_i \right) \log \hat{p}_j^k, \tag{3.3}$$

where $\hat{p}_j^k$ is the softmax probability for the class k at pixel j (the pixel in the center of the patch). Now, we can decompose this loss into:

$$\mathcal{L} = -\frac{1}{|\sum_i^d w_i|} \sum_k y_j^k \log \hat{p}_j^k \tag{3.4}$$

$$-\frac{1}{|\sum_i^d w_i|} \sum_k \left(\sum_{\substack{i=1 \\ i \neq j}}^{d} y_i^k w_i \right) \log \hat{p}_j^k, \tag{3.5}$$

[2] 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., $d_1 \times d_2$.

with j 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 j. Furthermore, let us denote $\tau_k = \sum_{\substack{i=1 \\ i \neq j}}^{d} y_i^k w_i$ 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. 3.4, and removing $|\sum_i^d w_i|$ as it multiplies both terms, the loss becomes:

$$\mathcal{L} = \underbrace{-\sum_k y_j^k \log \hat{p}_j^k}_{CE} \underbrace{- \sum_k \tau_k \log \hat{p}_j^k}_{\text{Constraint on } \boldsymbol{\tau}} . \tag{3.6}$$

As $\boldsymbol{\tau}$ is static prior, the second term in Eq. 3.6 can be replaced by a Kullback-Leibler (KL) divergence, leading to the following learning objective:

$$\mathcal{L} \overset{\text{c}}{=} \mathcal{L}_{CE} + \mathcal{D}_{KL}(\boldsymbol{\tau} || \hat{\boldsymbol{p}}), \tag{3.7}$$

where $\overset{\text{c}}{=}$ 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 $\boldsymbol{\tau} = \hat{\boldsymbol{p}}$, 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 pixel j.

3.3.3 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 $\boldsymbol{\tau}$ 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:

$$\min_{\theta} \quad \mathcal{L}_{CE} \quad \text{s.t.} \quad \boldsymbol{\tau} = \mathbf{l}, \tag{3.8}$$

where $\boldsymbol{\tau}$ now represents a desirable prior, and $\boldsymbol{\tau} = \mathbf{l}$ is a hard constraint. Note that the reasoning behind working directly on the logit space is two-fold. First, observations in (Liu *et al.*, 2022b) 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[3], their magnitudes are further decreased, which has a favorable effect on model calibration (Müller *et al.*, 2019b). We stress that despite both (Liu *et al.*, 2022b) and our method enforce constraints on the predicted logits, (Liu *et al.*, 2022b) is fundamentally different. In particular, (Liu *et al.*, 2022b) 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. 3.8 could be solved by a standard Lagrangian-multiplier algorithm, this method may be challenging to implement in practice. In particular, these approaches present important limitations in the context of deep networks, such as training instability due to the constraint prevailing the main objective term, i.e., CE, require convexity assumption to ensure convergence, and computational overhead derived from the iterative updates of the multipliers and constraints, becoming problematic in large-scale deep learning models (Bertsekas, 1995; Boyd & Vandenberghe, 2004). Therefore, we replace the hard constraint by a soft penalty of the form $\mathcal{P}(|\boldsymbol{\tau} - \mathbf{l}|)$, transforming our constrained

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

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 $\mathcal{P}(|\boldsymbol{\tau} - \mathbf{l}|) \geq \mathcal{P}(\mathbf{0}), \forall \boldsymbol{l} \in \mathbb{R}^K$, i.e., when the constraint is satisfied. Following this, when the constraint $|\boldsymbol{\tau} - \mathbf{l}|$ deviates from $\mathbf{0}$ the value of the penalty term increases. Thus, we can approximate the problem in Eq. 3.8 as the following simpler unconstrained problem:

$$\min_{\theta} \quad \mathcal{L}_{CE} + \lambda \sum_{k} |\tau_k - l_k|, \tag{3.9}$$

where the hyperparameter λ controls the importance of the penalty.

3.4 Experiments

3.4.1 Experimental Setting

3.4.1.1 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.*, 2021b). 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 192×192×30, 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 adjoined 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 & 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.*, 2015a). The purpose of this challenge is to segment the brain MRI scans into Gray Matter (GM), White Matter (WM), and Cerebralspinal fluid (CSF). The dataset contains paired T1, T2, and T1-IR sequences of 3D volumes (240×240×48) 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, using a single split in the experiments.

3.4.1.2 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.*, 2015a). 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.*, 2020b). Importantly, we obtain the calibration metrics only for the foreground regions following the recent literature (Islam & 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 & Caruana, 2005c), 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_{ij} 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 $C_{i,j} = \frac{1}{|B_{ij}|} \sum_{k \in B_{i,j}} \hat{p}_{kj}$ respectively. Hence, ECE and CECE is given by:

$$ECE = \sum_{i=1}^{M} \frac{|B_i|}{N} |A_i - C_i| \tag{3.10}$$

$$CECE = \sum_{i=1}^{M} \sum_{j=1}^{K} \frac{|B_{i,j}|}{N} |A_{i,j} - C_{i,j}| \tag{3.11}$$

3.4.1.3 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 & Glocker, 2021), and MbLS (Liu *et al.*, 2022b). Following the literature, we have chosen the following hyper-parameters for the different approaches: FL (γ=3.0), ECP (α=0.1), LS (λ=0.1), SVLS (σ=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.*, 2015b) 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)).

Table 3.1 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 ↑	HD ↓	DSC ↑	HD ↓	DSC ↑	HD ↓	DSC ↑	HD ↓	DSC ↑	HD ↓	DSC ↑	HD ↓	DSC ↑	HD ↓
ACDC	RV	0.799	3.10	0.580	9.37	0.751	4.93	0.796	3.34	0.791	2.89	0.812	**2.59**	**0.837**	3.02
	MYO	0.795	2.57	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	0.889	3.75	0.724	6.97	0.839	4.85	0.858	3.49	0.882	2.89	0.875	3.53	**0.905**	**2.59**
	Mean	0.828	3.14	0.620	7.30	0.782	4.44	0.809	3.30	0.824	2.81	0.827	2.99	**0.854**	**2.55**
FLARE	Liver	0.950	6.09	0.952	7.54	0.953	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	0.950	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	0.900	4.26
	Pancreas	0.636	7.95	0.626	7.80	0.649	7.77	0.637	6.45	0.650	6.91	0.596	8.61	**0.664**	7.37
	Mean	0.855	6.40	0.853	6.65	0.860	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	0.803	4.88	**0.804**	**3.98**
	ET	0.766	8.27	0.854	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	0.889	8.09	**0.893**	**6.78**
	Mean	0.789	6.96	0.848	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	12.73
	PZ	0.752	7.13	0.677	12.57	0.736	6.19	0.756	5.12	0.737	9.28	0.774	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	0.601	**8.29**	**0.607**	8.37
KiTS	Kidney	**0.786**	9.11	0.784	**8.74**	0.735	10.27	0.759	9.06	0.770	9.86	0.749	10.56	0.780	9.08
	Tumor	0.447	**13.09**	0.470	13.57	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**	0.627	11.15	0.550	12.88	0.602	12.83	0.619	12.91	0.588	13.71	**0.652**	12.42
MRBrainS	GM	0.754	1.73	0.672	2.81	0.747	2.23	0.707	2.12	0.725	1.71	0.741	2.09	**0.781**	**1.41**
	WM	0.759	2.91	0.598	5.60	0.783	2.73	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	0.800	1.41	0.769	1.71	**0.820**	**1.21**
	Mean	0.763	2.22	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 3.2 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 ↓	CECE ↓	ECE ↓	CECE ↓	ECE ↓	CECE ↓	ECE ↓	CECE ↓	ECE ↓	CECE ↓	ECE ↓	CECE ↓	ECE ↓	CECE ↓
ACDC	0.137	0.084	0.153	0.179	0.130	0.094	0.083	0.093	0.091	0.083	0.103	0.081	**0.048**	**0.061**
FLARE	0.058	0.034	0.053	0.059	0.037	**0.027**	0.055	0.049	0.039	0.036	0.046	0.041	**0.033**	0.031
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**	0.112	0.097
PROSTATE	0.430	0.304	0.271	0.381	0.306	0.252	0.304	0.301	0.335	0.272	0.322	**0.250**	**0.253**	0.254
KiTS	0.188	0.144	0.098	0.133	0.155	0.151	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	0.061	0.101	0.077	0.080	0.107	0.093	**0.027**	**0.056**

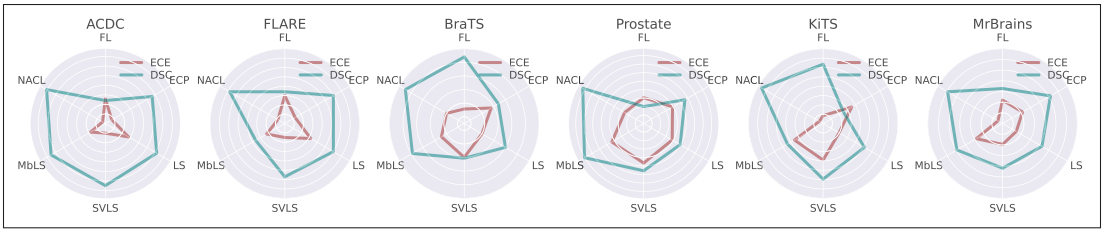

Figure 3.1 **Compromise between calibration and discriminative performance.** For each dataset, we show the discriminative (DSC) and calibration (ECE) results obtained by each method. We expect a *well-calibrated* model to achieve simultaneously large DSC (*in blue*) and small ECE (*in brown*) values

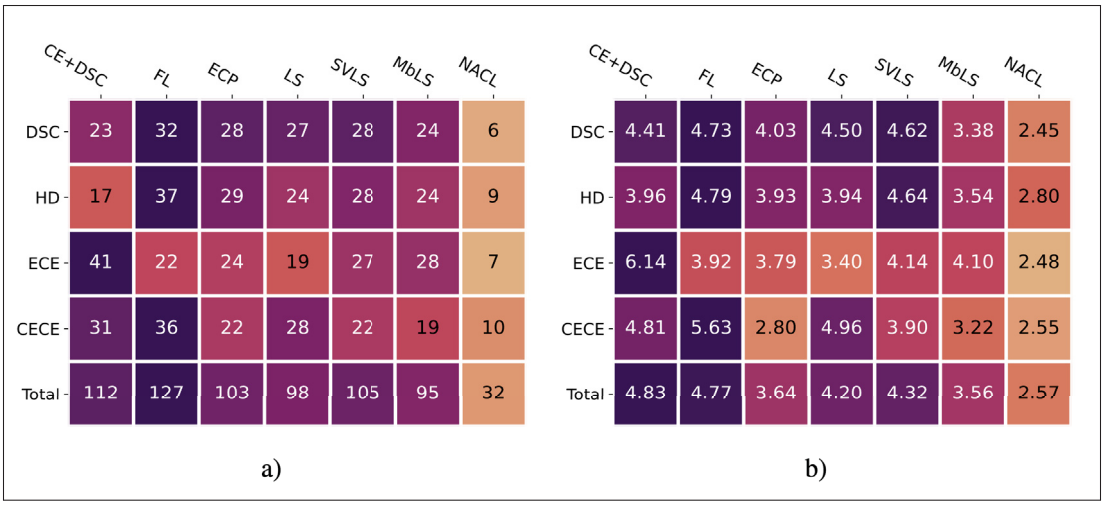

Figure 3.2 Ranking *global* and *per-metric* of the different methods based on the sum-rank and mean of case-specific approach

3.4.2 Results

3.4.2.1 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 3.1, 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 3.2) reveal that our approach consistently yields the best, and second best, uncertainty estimates across datasets and target objects. Furthermore, and as observed in Table 3.1, 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 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 Table 3.1 and 3.2 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 3.1 and 3.2. First, we resort to radar plots in Figure 3.1 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 implies 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.*, 2015a) and mean-case-rank (Maier *et al.*, 2017). As we can observe in the heatmaps provided in Fig. 3.2, 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.

In addition to the popular ECE and CECE metrics used in calibration, we further evaluate whether the observations above still hold when using adaptive binning schemes, such as Adaptive Calibration Error (ACE), and Thresholded Adaptive Calibration Error (TACE) (Nixon, Dusenberry, Zhang, Jerfel & Tran, 2019a). Following Eq. (3.11), the equispaced M bins are replaced with adaptive range R, where $R_{i,j}$ denotes the number of samples from j^{th} class in the i^{th} range and ACE is given by:

$$ACE = \sum_{i=1}^{R} \sum_{j=1}^{K} \frac{|R_{i,j}|}{N} |A_{i,j} - C_{i,j}| \tag{3.12}$$

1 TACE uses a threshold to prevent correct and confident predictions from dominating the calibration score. The results from these metrics, which are reported in Table 3.3, confirm the

trend observed with equally-spaced ECE metrics, where our model consistently yields very competitive performance.

Table 3.3 Calibration performance evaluated in terms of adaptative binning schemes, i.e., ACE (top) and TACE (bottom), across six popular medical image segmentation benchmarks

Dataset	CE+DSC	FL	ECP	LS	SVLS	MbLS	Ours
ACDC	0.137	0.155	0.129	0.089	0.091	0.109	**0.069**
FLARE	0.058	**0.033**	0.037	0.072	0.039	0.038	0.043
BraTS	0.223	**0.098**	0.139	0.121	0.125	0.132	0.114
PROSTATE	0.429	0.269	0.305	0.306	0.334	0.323	**0.253**
KiTS	0.188	0.099	0.155	0.126	0.162	0.156	**0.091**
MRBrainS	0.172	0.102	0.049	0.033	0.095	0.072	**0.031**
ACDC	0.151	0.224	0.151	0.093	0.138	0.081	**0.073**
FLARE	0.123	0.145	0.134	0.049	0.144	0.127	**0.031**
BraTS	0.201	0.146	0.145	0.108	0.141	**0.095**	0.097
PROSTATE	0.377	0.383	0.282	0.301	0.294	**0.249**	0.254
KiTS	0.187	0.151	0.161	0.141	0.174	0.147	**0.124**
MRBrainS	0.151	0.131	0.110	0.084	0.110	0.079	**0.057**

3.4.2.2 On the impact of constraining the logit space

Constraint over logits *vs* softmax. Recent evidence (Liu *et al.*, 2022b; 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 $\mathbf{l}$ by $\hat{\mathbf{p}}$ in Equation 3.9. From these results, reported in Figure 3.3, 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 3.4. 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

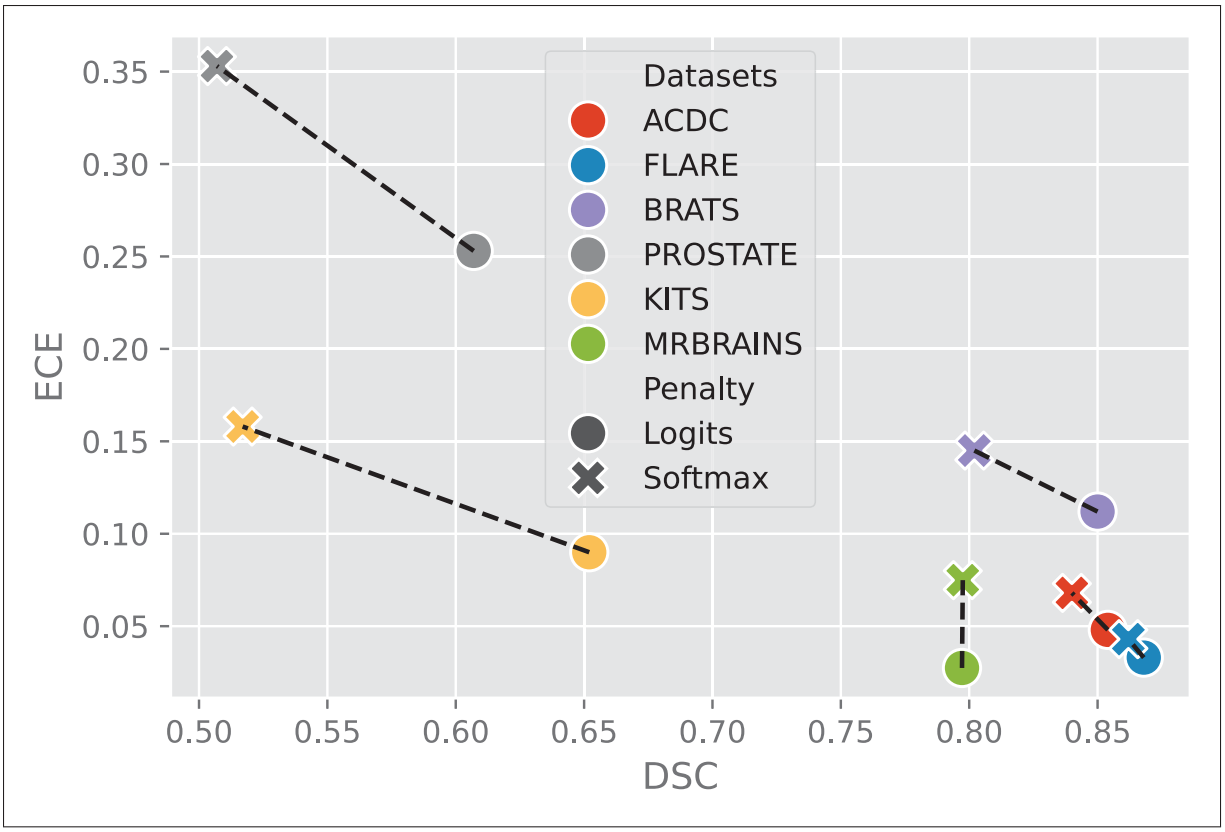

Figure 3.3 Impact of applying the penalty over softmax (*cross*) vs logits (*circle*) predictions across the different datasets

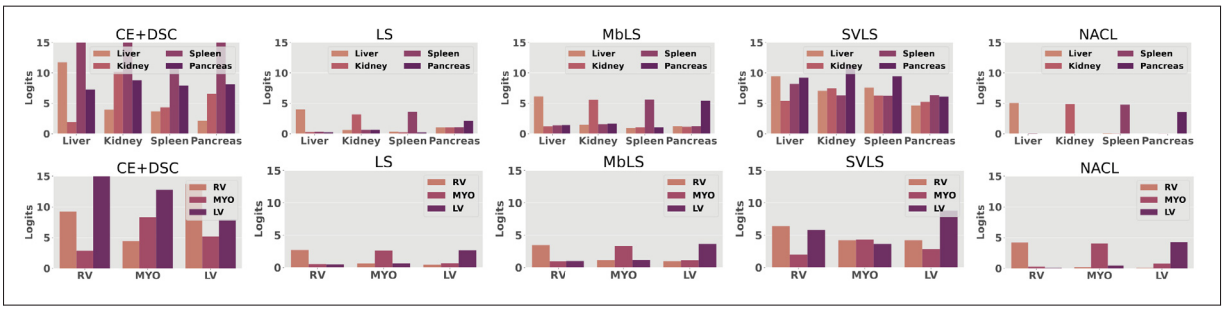

Figure 3.4 Distribution of logit predictions provided by a model trained with CE+DSC, LS, MbLS, SVLS and our approach (*from left to right*) on FLARE (*top*) and ACDC (*bottom*)

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 spatial information. In contrast, whereas LS and MbLS have a desired logit distribution for

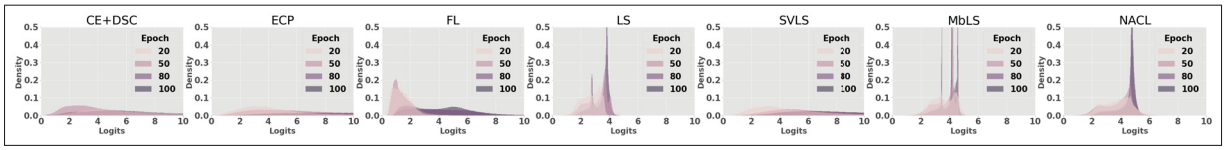

Figure 3.5 Histogram of global logit distribution over epochs obtained by the different approaches

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. 3.5. It is well known that, calibrated methods show a better regularization, restricting the range of logits to a particular range (Müller *et al.*, 2019b). 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.*, 2019b), 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.

3.4.2.3 On the impact of hyperparameters

In this experiment, we assess the sensitivity of the hyper-parameters in the different methods, and possibly find a setting which works best across datasets. For FL, γ values of 1, 2, and 3 are

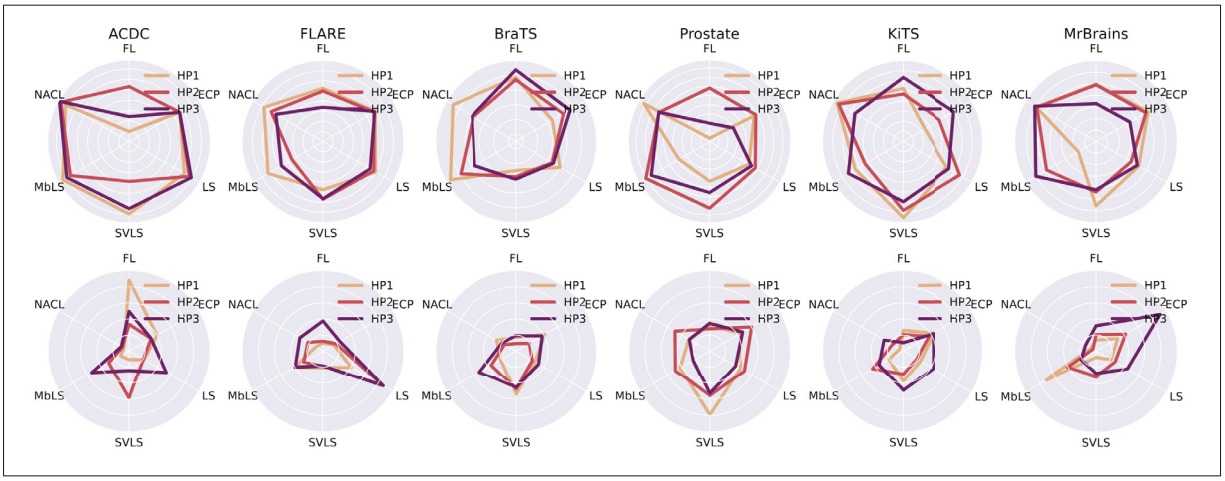

Figure 3.6 **Radar plots displaying hyperparameter-dependence performance (DSC on *top* and ECE in the *bottom*).** HP1, HP2 and HP3 denote the respective hyper-parameter set: FL (γ=[1,2,3]), ECP (λ =[0.1,0.2,0.3]), LS (α =[0.1,0.2,0.3]), MbLS (m=[3,5,10]) and SVLS (σ=[0.5,1,2], and ours (λ=[0.1,0.2,0.3]). Our method consistency provides best performance for 0.1 across datasets

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) performances using these hyper-parameters across the different datasets and depicted the results in Figure 3.6. From this figure, it can be observed that, our method is fairly consistent with a particular hyper-parameter (HP1). Moreover, while varying λ can lead to performance differences in our approach, these are smaller compared to existing approaches. Indeed, other methods presented larger performance variations, as discrimination and calibration metrics were highly sensitive to the hyper-parameter choice. For example, in ACDC, FL 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

What if we can fine-tune the hyperparameters? Providing a method that yields competitive results across datasets with its hyperparameters fixed brings several benefits in practice, e.g., it avoids extensive grid-search on a validation set. Nonetheless, one may argue that finding the

optimal value, in a per-dataset basis, should be considered, as the performance of the delivered model can be further improved. We perform in this section such analysis, and depict in Figure 3.7[4] the results when the hyperparameter value is selected based on the best DSC score on the validation set for each dataset. In these plots, the best method across each dataset can be identified based on the largest gap between DSC and ECE scores, as one would expect the highest discriminative (DSC) scores accompanied with the lowest calibration (ECE) values. Thus, while some methods perform satisfactorily in some datasets, such as SVLS in MRBrains, or FL in BraTS, our approach NACL brings the most consistent results across datasets, aligning with the observations in the previous section.

3.4.2.4 Effect of the prior

Ablation on different priors. A benefit of the proposed formulation, particularly compared to SVLS (Islam & 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 3.4 reveal that selecting a suitable prior can further improve the performance of our model.

Table 3.4 **Impact of using different priors**. We compare the discrimative and calibration performance of our approach across the six datasets when using different priors τ in Equation 3.9

	FLARE		ACDC		BraTS		PROSTATE		KiTS		MRBrainS	
Prior τ	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
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

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

[4] Note that Figure 3.7 is in fact a clean version of Figure 3.6, where only the performance of the best hyperparameter value is depicted.

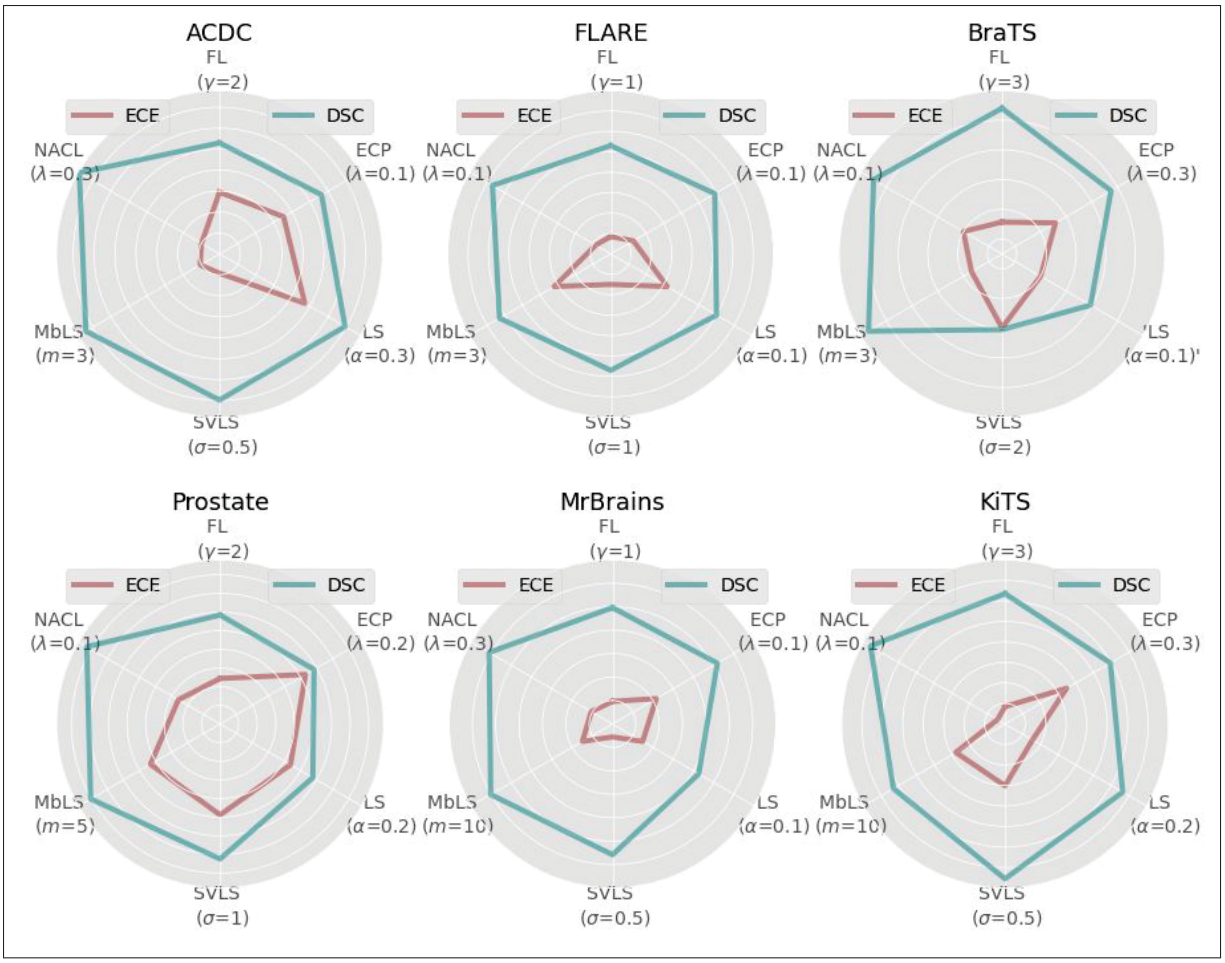

Figure 3.7 **What if we can fine-tune the best hyperparameter?** These plots depict the discriminative and calibration results when the optimal hyperparameter value (in brackets) is selected for each method

well to define a Gaussian prior in our formulation, whose results are depicted in Fig. 3.8. 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 3.3.2. These results show empirically that our method is able to better leverage the neighboring class information compared to SVLS.

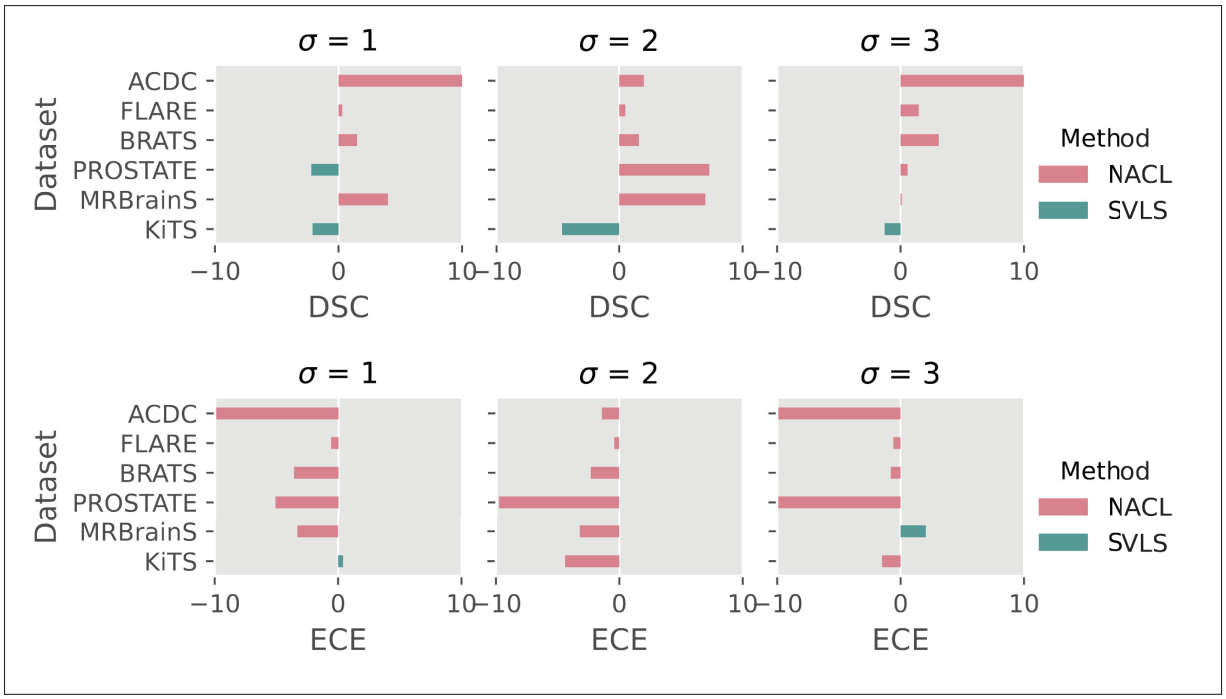

Figure 3.8 **Direct comparison of SVLS (Islam & Glocker, 2021) *vs.* NACL (Ours).** Relative error differences (%) between SVLS and our method when using the same Gaussian prior (with $\sigma = \{1, 2, 3\}$)

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: AttUNet (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. 3.9. 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.

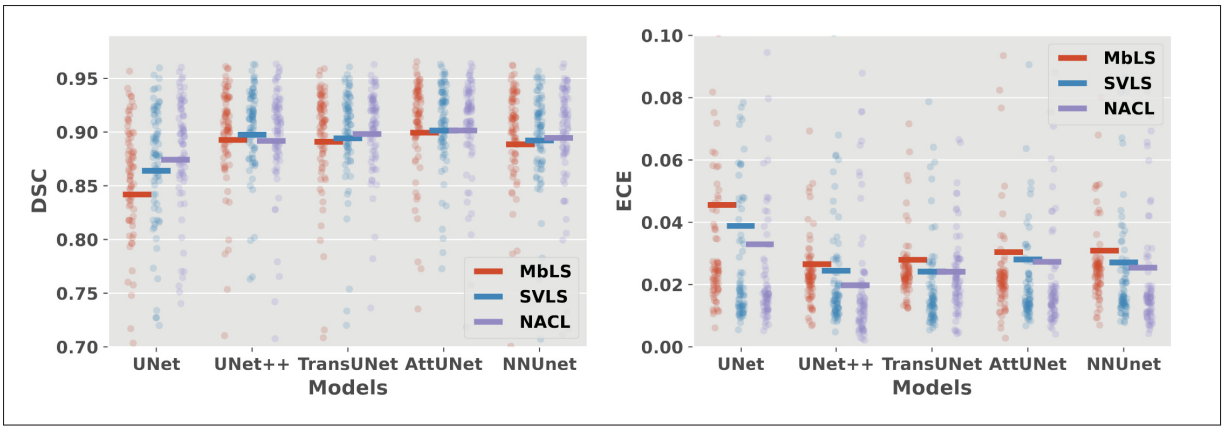

Figure 3.9 **Robustness to the segmentation backbone.** We evaluate the performance of competing approaches (i.e., MbLS and SVLS) on the FLARE dataset when using different architectures as segmentation backbones

3.4.2.6 Sensitivity to the number of training samples

In this experiment, we investigate whether varying the number of training samples impacts the performance of the different calibration methods, as well as the CE+DSC compounded loss. 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. 3.10 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.

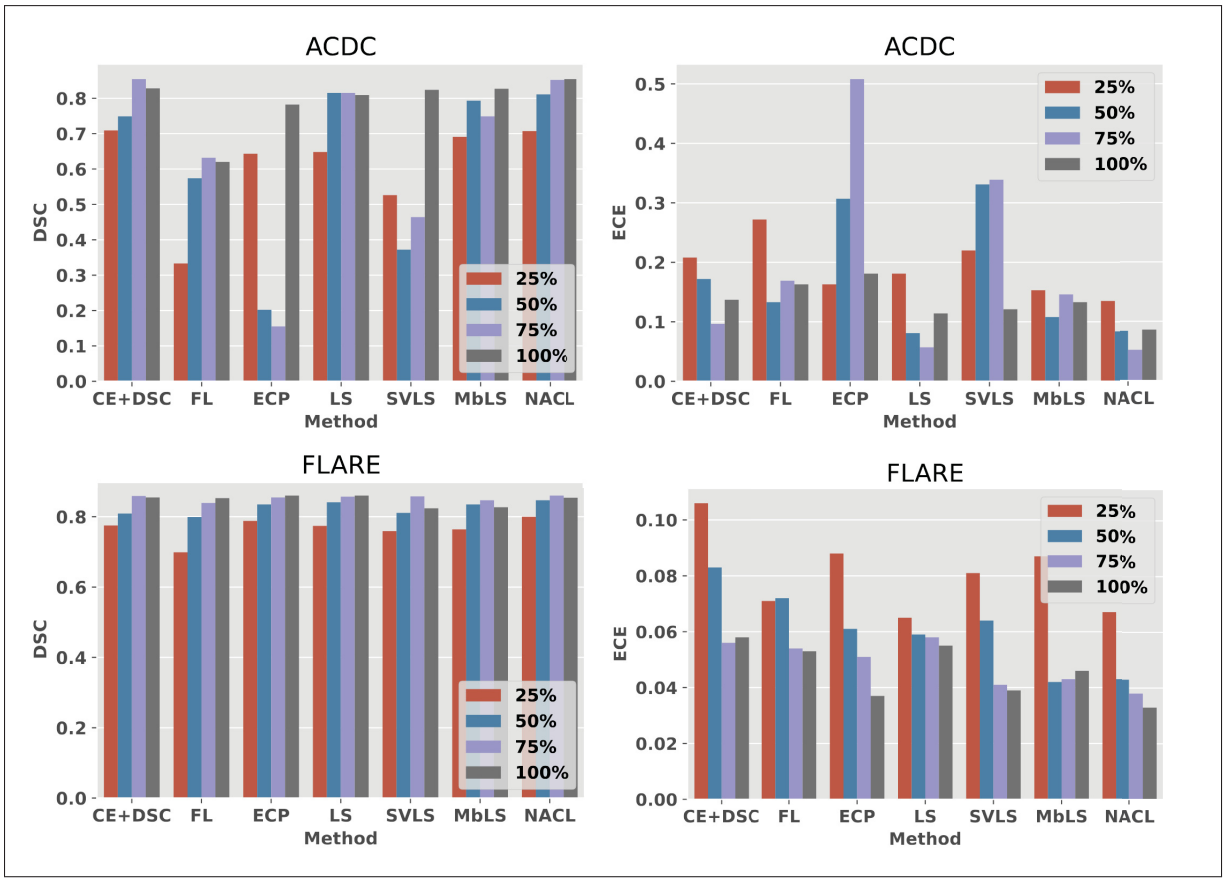

Figure 3.10 **Performance variation with number of labeled images.** These plots depict the performance of different approaches under several data labeled scenarios, going from 100% (i.e., original provided data) to 25% of images from the original dataset

3.4.2.7 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 3.5 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. 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.

Table 3.5 **Impact of different penalties.**Comparison of using a L_1 *vs* a L_2 penalty to impose the constraint in Equation 3.9

	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

3.4.2.8 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 target and predicted foregrounds, whose ECE and CECE values, against the DSC metric, are depicted in Figure 3.11. 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

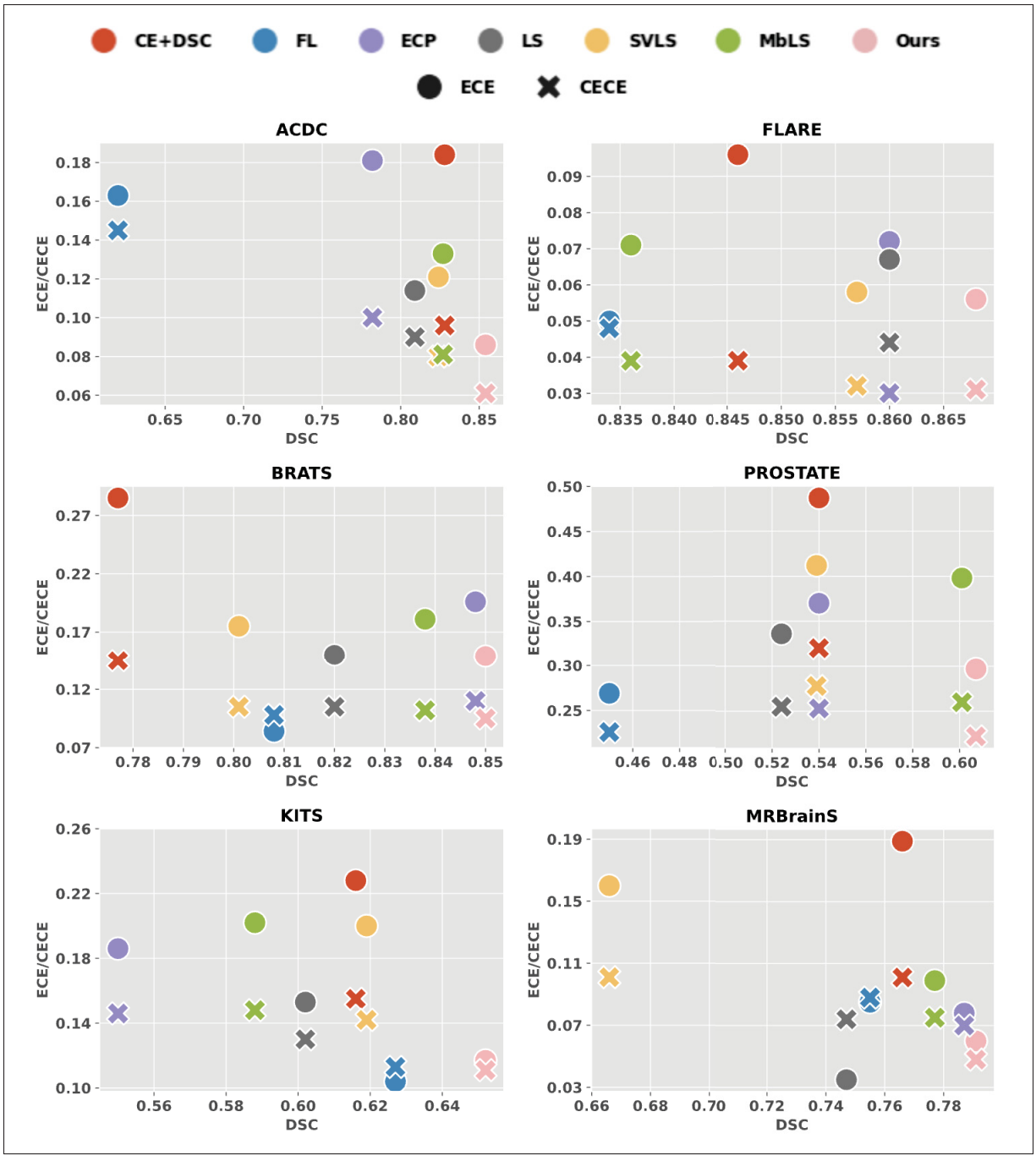

Figure 3.11 Scatter plots comparing DSC vs ECE/CECE when considering the foreground (prediction ∪ target) to compute the calibration metrics

does not change the message that the proposed approach offers a better alternative to existing calibration methods.

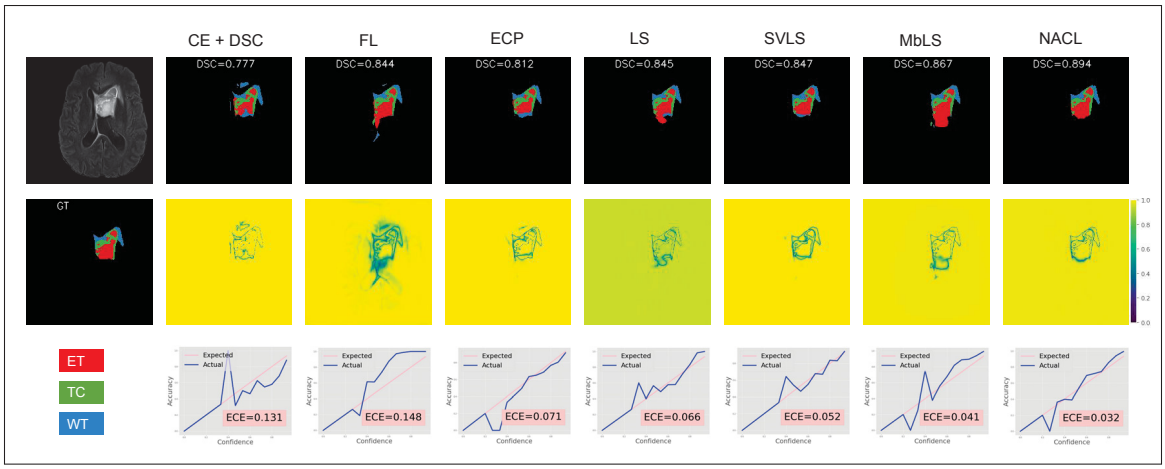

Figure 3.12 Qualitative results on BraTS dataset for different methods. In particular, we show the original image and the corresponding segmentation masks provided by each method (*top row*), the ground-truth (GT) mask followed by maximum confidence score of each method (*middle row*) and the respective reliability plots (*bottom row*). Methods from left to right: CE+DSC, FL, ECP, LS, SVLS, MbLS, and Ours

3.4.2.9 Qualitative results and reliability diagrams

We show now in Figure 3.12 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.*, 2022b), 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 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.

3.4.2.10 Robustness across multiple seeds

We now assess the robustness of the different methods across multiple seeds, whose average performance in depicted in Figure 3.13. More concretely, three different seeds are used to run the experiments three times, the same set of seeds are employed for all the methods and the mean over the three runs are reported. We can observe that, despite using different seeds, the findings from Figure 3.13 align with the main results reported in Table 3.1, with the proposed approach typically yielding the best segmentation and calibration performance. Indeed, looking closer to these results we can state that the proposed approach offers the best discriminative-calibration compromise, regardless of the dataset studied.

The detailed numerical values from the plots in Figure 3.13, altogether with their standard deviation, are presented in Table 3.6. These results showcase not only the superiority of the proposed approach, as it offers the best trade-off between discrimination and calibration performance, but also its robustness, since the standard deviation across seeds is typically lower compared to other methods. These observations underscore the potential of our approach from a clinical practice standpoint, as it typically yields the best results, yet being a robust strategy.

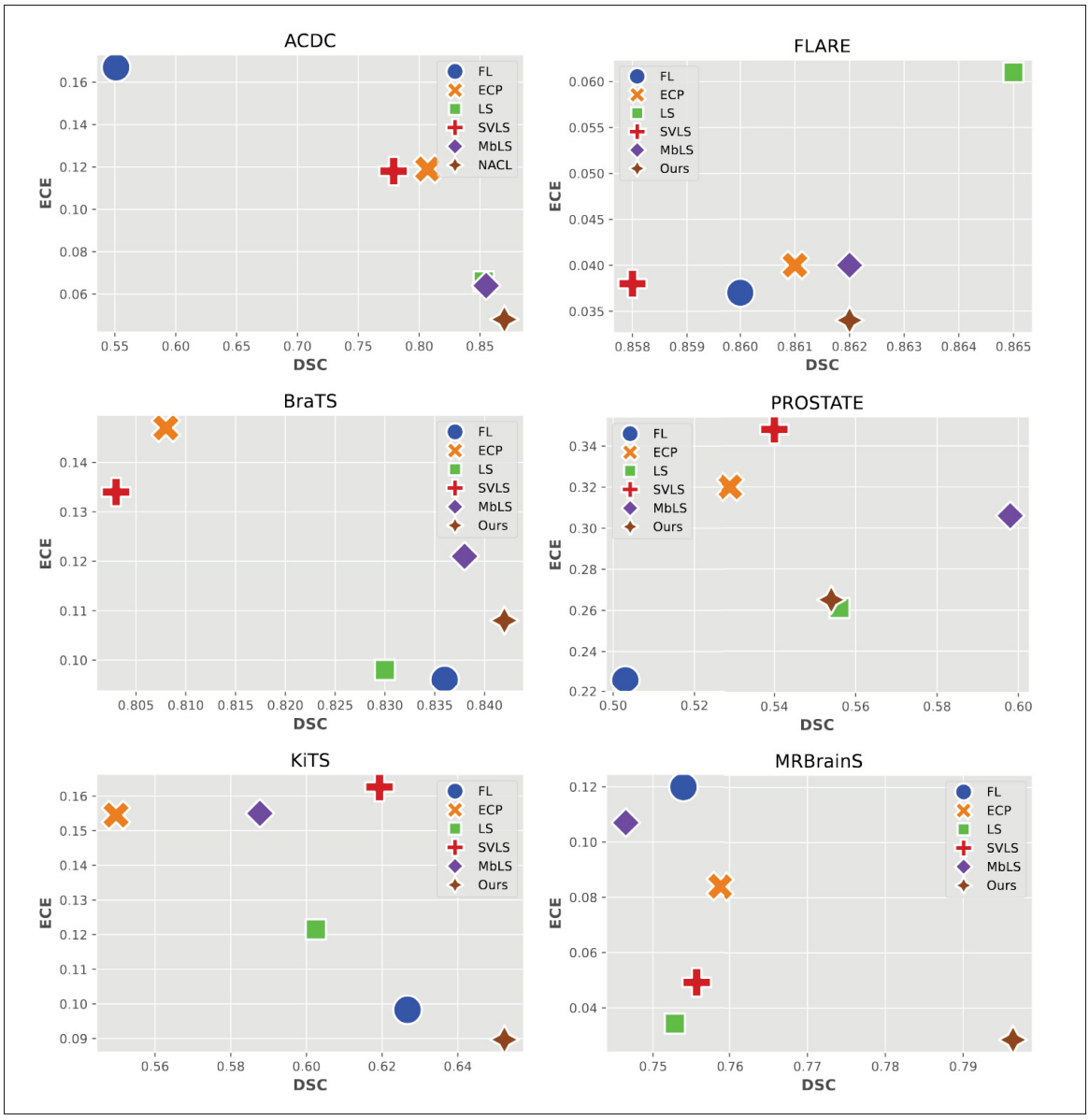

Figure 3.13 **Average results across three different seeds.** These scatter plots illustrate the average DSC vs ECE correlation for the different methods, and across multiple datasets, when three seeds are used

3.5 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

Table 3.6 **Segmentation and Calibration Results.** Average DSC and ECE scores across three seeds for six medical image segmentation benchmarks. Best method is highlighted in bold, and second best is underlined

Dataset	FL	ECP	LS	SVLS	MbLS	NACL
			DSC ↑			
ACDC	$55.05_{14.88}$	$74.81_{10.43}$	$85.30_{4.10}$	$77.94_{4.30}$	$\underline{85.52}_{2.47}$	$\mathbf{87.04}_{1.58}$
FLARE	$\underline{86.22}_{0.66}$	$86.05_{1.00}$	$\mathbf{86.51}_{0.71}$	$85.80_{0.27}$	$86.19_{2.29}$	$86.20_{0.58}$
BraTS	$83.62_{1.00}$	$80.81_{1.87}$	$82.96_{0.84}$	$80.29_{1.84}$	$\underline{83.81}_{0.10}$	$\mathbf{84.15}_{1.02}$
PROSTATE	$50.34_{4.75}$	$52.89_{1.24}$	$\underline{55.57}_{2.77}$	$53.96_{3.20}$	$\mathbf{59.79}_{0.51}$	$55.37_{5.45}$
KiTS	$61.11_{6.21}$	$61.69_{5.85}$	$\underline{64.18}_{4.39}$	$61.00_{2.49}$	$62.82_{3.78}$	$\mathbf{65.34}_{0.53}$
MRBrainS	$75.39_{7.85}$	$76.35_{2.56}$	$75.28_{3.68}$	$75.56_{4.28}$	$\underline{76.44}_{5.80}$	$\mathbf{79.64}_{1.22}$
			ECE ↓			
ACDC	$16.66_{6.87}$	$15.09_{5.56}$	$6.62_{1.61}$	$11.85_{2.78}$	$\underline{6.43}_{3.37}$	$\mathbf{4.81}_{0.60}$
FLARE	$4.37_{0.64}$	$3.96_{0.93}$	$6.13_{0.59}$	$\underline{3.80}_{0.25}$	$4.00_{0.49}$	$\mathbf{3.45}_{0.19}$
BraTS	$\mathbf{9.58}_{0.17}$	$14.71_{0.73}$	$\underline{9.82}_{1.34}$	$13.45_{1.10}$	$12.12_{0.90}$	$10.83_{1.17}$
PROSTATE	$\mathbf{22.63}_{4.04}$	$32.02_{3.85}$	$\underline{26.13}_{6.19}$	$34.76_{1.44}$	$30.56_{2.14}$	$26.52_{1.31}$
KiTS	$11.84_{4.03}$	$13.71_{2.60}$	$\mathbf{10.67}_{1.30}$	$14.75_{1.79}$	$14.69_{1.08}$	$\underline{11.33}_{2.26}$
MRBrainS	$11.99_{3.02}$	$8.21_{2.02}$	$\underline{3.43}_{2.27}$	$4.92_{2.45}$	$8.03_{5.21}$	$\mathbf{2.85}_{0.27}$

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 effectively. 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.

Limitations of the proposed approach. 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.

CHAPTER 4

CLASS AND REGION-ADAPTIVE CONSTRAINTS FOR NETWORK CALIBRATION

Balamurali Murugesan[a] , Julio Silva-Rodriguez[a] , Ismail Ben Ayed[b] , Jose Dolz[a]

[a] Department of Software Engineering, École de Technologie Supérieure,
[b] Department of System Engineering, École de Technologie Supérieure,
1100 Notre-Dame West, Montreal, Quebec, Canada H3C 1K3

Paper published in Medical Image Computing and Computer Assisted Intervention, October 2024

Presentation

This chapter presents the article *"Class and Region adaptive constraints for network calibration"* (Murugesan, Silva-Rodriguez, Ben Ayed & Dolz, 2024c) published in International Conference on Medical Image Computing and Computer Assisted Intervention (**MICCAI**) Page. 57–67 2024. The paper was granted an Early Accept by the conference committee.

Abstract

In this work, we present a novel approach to calibrate segmentation networks that considers the inherent challenges posed by different categories and object regions. In particular, we present a formulation that integrates class and region-wise constraints into the learning objective, with multiple penalty weights to account for class and region differences. Finding the optimal penalty weights manually, however, might be unfeasible, and potentially hinder the optimization process. To overcome this limitation, we propose an approach based on Class and Region-Adaptive constraints (CRaC), which allows to learn the class and region-wise penalty weights during training. CRaC is based on a general Augmented Lagrangian method, a well-established technique in constrained optimization. Experimental results on two popular segmentation benchmarks, and two well-known segmentation networks, demonstrate the superiority of CRaC compared to existing approaches. The code is available at: https://github.com/Bala93/CRac/

4.1 Introduction

Despite the remarkable progress achieved by deep neural networks (DNNs), they are susceptible to suffer from miscalibration, leading to overconfident predictions (Guo *et al.*, 2017b; Minderer *et al.*, 2021), even when they are incorrect. This issue becomes especially significant in safety-critical scenarios, such as medical diagnosis or treatment, where producing accurate uncertainty estimates is of paramount importance. An inherent cause of network miscalibration is known to be the implicit bias for low-entropy predictions caused by popular supervised losses, such as the cross-entropy, which encourages large differences between the logit of the ground truth category and the remaining classes (Mukhoti *et al.*, 2020b).

A myriad of approaches have emerged to mitigate network miscalibration, which mainly focus on either post-processing strategies or integrating additional learning objectives during training. The first family of approaches, i.e., *post-processing* methods, offers a simple alternative for modifying the softmax predictions in a post-hoc fashion by establishing a mapping from raw network outputs to well-calibrated confidences (Ding *et al.*, 2021; Guo *et al.*, 2017b; Gupta *et al.*, 2020; Tomani *et al.*, 2021a; Zhang *et al.*, 2020c). The second category involves incorporating additional regularization during training, typically penalizing low-entropy predictions. For example, (Pereyra *et al.*, 2017) introduced an explicit term that maximizes the Shannon entropy of the network predictions during training, which was later extended in (Larrazabal, Martínez, Dolz & Ferrante, 2023b) by penalizing low-entropy distributions only in incorrect predictions. Furthermore, popular losses for classification, such as Label smoothing (Szegedy *et al.*, 2016) or focal loss (Lin *et al.*, 2017), implicitly integrate an entropy maximization term, which has a favourable effect on calibration (Mukhoti *et al.*, 2020b; Müller *et al.*, 2019b). More recently, (Liu *et al.*, 2022b, 2023a; Murugesan *et al.*, 2023b) propose to enforce inequality constraints on the logit space, allowing to control the margin on logit distances, ultimately reducing overconfidence in the predictions. This provided more flexibility than systematically maximizing the entropy of the predictions, as in (Mukhoti *et al.*, 2020b; Müller *et al.*, 2019b), which results in gradients that continually push towards a non-informative solutions. Other works include the integration of pair-wise constraints between classes (Cheng & Vasconcelos, 2022) or augmenting the training

dataset by convex combinations of random pairs of images and their associated labels, e.g., MixUp (Thulasidasan *et al.*, 2019b). Nevertheless, even though these works have achieved remarkable progress in addressing miscalibration in both classification (Cheng & Vasconcelos, 2022; Guo *et al.*, 2017b; Gupta *et al.*, 2020; Mukhoti *et al.*, 2020b; Müller *et al.*, 2019b; Pereyra *et al.*, 2017; Tomani *et al.*, 2021a) and segmentation tasks (Ding *et al.*, 2021; Larrazabal *et al.*, 2023b; Liu *et al.*, 2022b; Murugesan *et al.*, 2023b), they disregard neighbour pixel relationships, in terms of classes, which is of significant relevance in semantic image segmentation.

Certainly, one of the factors contributing to the reduced performance of these losses in segmentation tasks arises from the uniform, or near-to-uniform, distribution enforced in the network predictions (whether logit or softmax predictions), which neglects the spatial context (Murugesan *et al.*, 2023a). To overcome this issue, and to integrate class-wise information of the surrounding pixels during training, Spatially Varying Label Smoothing (SVLS) (Islam & Glocker, 2021) introduced a label smoothing strategy that captures the structural uncertainty required in semantic segmentation. More specifically, SVLS uses a Gaussian kernel applied across the one-hot encoded ground truth, leading to class probabilities based on a soft combination of neighboring pixels. As exposed in (Murugesan *et al.*, 2023a, 2025), SVLS integrates an implicit penalty on softmax predictions, which enforces a prior based on soft class proportions of surrounding pixels. This strategy, however, lacks a mechanism to control the influence of the constraint over the main objective, potentially hindering the optimization process. To circumvent this limitation, authors presented a simple solution that combines the standard cross-entropy with an explicit penalty, where both the prior and its impact can be easily controlled.

Although the work proposed in (Murugesan *et al.*, 2023a, 2025) achieves greater calibration performance than existing alternatives, and integrates class-relationships across a pixel and its neighbours, it presents two major limitations: *1)* The scalar balancing weight that controls the importance of the penalty is equal for all classes, and for all the regions. This scenario is suboptimal, as it can hamper the network performance when some classes are more challenging to segment, or under-represented. Furthermore, this strategy considers than the weight of the penalty should be the same for a pixel inside the object (likely to have *low uncertainty*) than for a

pixel within the organ boundaries (likely to have *high uncertainty*). *2)* The value of the balancing weight is defined before network optimization, lacking an adaptive strategy during training. For example, as the training evolves, the cross-entropy loss pushes towards lower-entropy predictions, whereas the penalty weight is the same at the beginning and the end of the training.

Based on these findings, we can summarize our contributions as:

1. We propose a class and region-wise constraint approach to tackle the miscalibration issue in semantic segmentation models. In particular, we formulate a solution that considers the specificities of each category and different regions by introducing independent class and region-wise penalty weights. This contrasts with the prior work in (Murugesan *et al.*, 2023a), where a uniform scalar penalty weight is employed, regardless of categories or regions.
2. Furthermore, we transfer the constrained problem to its dual unconstrained optimization counterpart by using an Augmented Lagrangian method (ALM). This alleviates the need for manually adjusting each penalty weight and allows, through a series of iterative *inner* and *outer* steps, to find the optimal value of each penalty weight, which can be learned in an adaptive manner.
3. Comprehensive experiments on two popular segmentation benchmarks, and with two well-known segmentation backbones, demonstrate the superiority of our approach over a set of relevant recent calibration approaches.

4.2 Methodology

Notation. We denote the training dataset as $\mathcal{D}(\mathcal{X}, \mathcal{Y}) = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$, where $\mathbf{x}^{(n)} \in \mathcal{X} \subset \mathbb{R}^{\Omega_n}$ represents the n^{th} image, Ω_n its spatial image domain, and $\mathbf{y}^{(n)} \in \mathcal{Y} \subset \mathbb{R}^{K}$ the corresponding pixel-wise ground-truth annotation with K classes, which is provided as a one-hot encoding vector. Given an input image $\mathbf{x}^{(n)}$, a neural network parameterized by $\boldsymbol{\theta}$ generates a logit vector $f_{\boldsymbol{\theta}}(\mathbf{x}^{(n)}) = \mathbf{l}^{(n)} \in \mathbb{R}^{\Omega_n \times K}$, which can be converted into probability values with the softmax operator, $\text{softmax}(\mathbf{l}^{(n)}) = \mathbf{s}^{(n)} \in [0, 1]^{\Omega_n \times K}$. To simplify the notations, we omit sample indices, as this does not lead to any ambiguity.

4.2.1 Background

Despite its importance in dense prediction tasks, such as segmentation, very few approaches consider pixel spatial relationships across classes to address the miscalibration issue. Spatially Varying Label Smoothing (SVLS) (Islam & Glocker, 2021) integrates neighbour class information by softening the pixel label assignments with a discrete spatial Gaussian kernel. More recently, NACL (Murugesan *et al.*, 2023a, 2025) formally showed that SVLS actually enforces an implicit constraint on soft class proportions of surrounding pixels, and propose the following constrained optimization problem to alleviate the limitations of SVLS:

$$\min_{\boldsymbol{\theta}} \quad \mathcal{L}_{CE} \quad \text{s.t.} \quad \boldsymbol{\tau} = \mathbf{l}, \tag{4.1}$$

which can be approximated by incorporating an explicit penalty, whose overall learning objective is defined as:

$$\min_{\boldsymbol{\theta}} \quad \sum_{i \in \Omega} \sum_{k \in K} (-y_k^{(i)} \log(s_k^{(i)}) + \lambda |\tau_k^{(i)} - l_k^{(i)}|). \tag{4.2}$$

The first term in the above equation is the standard cross-entropy loss on a given pixel, the second term is a linear penalty over the pixel logit distributions, $\boldsymbol{\tau}$ is a prior, and λ the balancing hyperparameter that controls the importance of each term. With this objective, when the constraint $|\tau_k - l_k|$ deviates from 0 (i.e., τ_k and l_k are different) the value of the penalty term increases. Thus, as the prior $\boldsymbol{\tau} = \{\tau_0, ..., \tau_{K-1}\}$ captures the class distribution of a 2D patch[1] surrounding the pixel, the penalty enforces the predicted logit distribution $\mathbf{l}$ to follow $\boldsymbol{\tau}$.

[1] More details about the priors and the enforced constraint in (Murugesan *et al.*, 2023a, 2025).

4.2.2 Class and region-wise penalties

The unconstrained formulation presented in Equation 4.2 employs a single uniform penalty. We argue that this scenario is suboptimal, as it disregards differences across individual categories, or even different regions with different uncertainty in the target object, which‘ may pose distinct inherent learning challenges. For example, annotations from a patch in the center of an organ typically have less uncertainty that labels in within the organ boundaries. A better, and more optimal strategy would integrate multiple penalty weights λ, one for each category and type of patch/region, leading to a set of penalty weights $\mathbf{\Lambda} \in \mathbb{R}_{+}^{K \times R}$, with R being the number of regions. For simplicity, in this work we will consider only two types of regions (i.e., $R = 2$, leading to $\mathbf{\Lambda} = \{\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1\}$), that we denote as *inner* and *outer* regions, and whose sets are defined as $\mathcal{I}$ and $\mathcal{O}$, respectively. More concretely, if the surrounding ground truth patch of a given pixel only contains one category, it will be considered as an *inner* patch, whereas otherwise it will be an *outer* patch. Thus, we can formally define our formulation as:

$$\min_{\theta} \quad \sum_{i \in \Omega} \mathcal{H}(\mathbf{y}^{(i)}, \mathbf{s}^{(i)}) + \sum_{i \in \mathcal{I}} \sum_{k \in K} \lambda_{k,0} |\tau_k^{(i)} - l_k^{(i)}| + \sum_{i \in \mathcal{O}} \sum_{k \in K} \lambda_{k,1} |\tau_k^{(i)} - l_k^{(i)}|, \tag{4.3}$$

where $\mathcal{H}(\mathbf{y}, \mathbf{s})$ is the standard cross-entropy loss. As stated in prior literature in constrained convolutional neural networks (Marquez Neila, Salzmann & Fua, 2017; Rony, Granger, Pedersoli & Ben Ayed, 2021; Liu *et al.*, 2023a; Silva-Rodriguez *et al.*, 2024), while $\mathbf{\Lambda}^* \in_{+}^{K \times R}$ are the Lagrange multipliers of the presented problem, and $\mathbf{\Lambda} = \mathbf{\Lambda}^*$ could be considered the best choice to solve (4.3), using $\mathbf{\Lambda}^*$ as the penalty weights may not feasible in practice. On the other hand, finding the optimal value for each penalty weight manually can pose optimization challenges, particularly for datasets with a large number of classes.

4.2.3 The proposed class and region adaptive solution

General Augmented Lagrangian. To alleviate the need of having to chose the penalty weights $\mathbf{\Lambda} \in_{+}^{K\times R}$, we propose to use an Augmented Lagrangian Multiplier (ALM) method. ALM approaches are optimization techniques that integrate penalties and primal-dual updates to efficiently tackle constrained optimization problems. These methods iteratively refine solutions by adjusting penalty terms based on Lagrange multipliers, effectively balancing between satisfying constraints, i.e., the penalties, and minimizing the main objective function, in our case the cross-entropy loss. ALM approaches are favoured due to their ability to handle complex constraints and their robust performance across various optimization scenarios, and enjoy widespread popularity in the general context of optimization (Bertsekas, 1996a; Nocedal & Wright, 2006). A general constrained optimization problem can be formally defined as:

$$\min_{x} \quad g(x) \quad \text{s.t.} \quad h_i(x) \leq 0, \quad i = 1, \ldots, n \tag{4.4}$$

with $g :^{d} \rightarrow$ the *objective function* and $h_i :^{d} \rightarrow, i = 1, \ldots, n$ being the *set of constraint functions*. Generally, this problem is tackled by solving a succession of $j \in \mathbb{N}$ unconstrained problems, each solved approximately w.r.t x:

$$\min_{x,\lambda} \quad \mathcal{L}^{(j)}(x) = g(x) + \sum_{i=1}^{n} P(h_i(x), \rho_i^{(j)}, \lambda_i^{(j)}), \tag{4.5}$$

where $P : \times_{++}\times_{++} \rightarrow$ is a *penalty-Lagrangian function*, whose derivative w.r.t. its first variable $P'(z, \rho, \lambda) \equiv \frac{\partial}{\partial z} P(z, \rho, \lambda)$ exists, is positive and continuous for all $z \in$ and $(\rho, \lambda) \in (_{++})^2$. In addition, we denote $\boldsymbol{\rho}^{(j)} = (\rho_i^{(j)})_{1\leq i\leq n} \in_{++}^{n}$ and $\boldsymbol{\lambda}^{(j)} = (\lambda_i^{(j)})_{1\leq i\leq n} \in_{++}^{n}$ as the penalty parameters and multipliers associated to the penalty P at the iteration j.

The ALM can be split into two iterations. First, in the *outer* iterations, which indexed by j, the *penalty multipliers* $\boldsymbol{\lambda}$ and the *penalty parameters* $\boldsymbol{\rho}$ are updated. Then, during the *inner* iterations, the objective $\mathcal{L}^{(j)}$ (Eq 4.5) is minimized using the previous solution as initialization

to this problem. Particularly, the penalty multipliers $\boldsymbol{\lambda}^{(j)}$ are updated to the derivative of P w.r.t. to the solution obtained during the last *inner* step:

$$\lambda_i^{(j+1)} = P'(h_i(x), \rho_i^{(j)}, \lambda_i^{(j)}). \tag{4.6}$$

This approach increases the value of the penalty multipliers when the constraint is violated, and decreases their value otherwise. Thus, integrating an ALM during optimization enables an *adaptive* and *learnable* strategy to determine an optimal value for the penalty weights.

Our global learning objective. Based on the benefits detailed above, we propose to solve the problem in Eq. 4.3 by using an ALM approach. More concretely, we reformulate this problem by integrating a penalty function P, which is parameterized by $(\boldsymbol{\rho}, \boldsymbol{\lambda}) \in_{++}^{K} \times_{++}^{K}$:

$$\begin{aligned} \min_{\theta, \boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1} \quad & \sum_{i \in \Omega} \mathcal{H}(\mathbf{y}^{(i)}, \mathbf{s}^{(i)}) + \sum_{i \in \mathcal{I}} \sum_{k \in K} P(\tau_k^{(i)} - l_k^{(i)}, \rho_{k,0}, \lambda_{k,0}) \\ & + \sum_{i \in O} \sum_{k \in K} P(\tau_k^{(i)} - l_k^{(i)}, \rho_{k,1}, \lambda_{k,1}). \end{aligned} \tag{4.7}$$

To obtain an accurate estimate of the penalty multipliers at each epoch, we compute the satisfaction of the constraint on the validation set, following standard practices in machine learning. In this work, we consider that a single training epoch approximately minimizes the loss function. Then, we compute the average penalty multiplier on the validation set. This means that, after a training epoch j, the penalty multipliers for all $k = 1, ..., K$ and each region r at epoch $j + 1$ can be computed as:

$$\lambda_{k,r}^{(j+1)} = \frac{1}{|\mathcal{D}_{val}|} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_{val}} P' \left(\tau_k - l_k, \rho_{k,r}^{(j)}, \lambda_{k,r}^{(j)} \right). \tag{4.8}$$

Furthermore, ρ is updated as:

$$\rho_{k,r}^{(j+1)} = \begin{cases} \gamma \rho_{k,r}^{(j)} & |\tau_k^{(j)} - l_k^{(j)}| > \mu * |\tau_k^{(j-1)} - l_k^{(j-1)}|; \\ \rho_{k,r}^{(j)} & \text{otherwise,} \end{cases} \tag{4.9}$$

where μ is a constant factor that determines the amount of the update. Last, following prior works on ALM in the context of constrained CNNs (Liu *et al.*, 2023a; Rony *et al.*, 2021; Silva-Rodriguez *et al.*, 2024), we employ PHR as the penalty, which is defined as:

$$\text{PHR}(z, \rho, \lambda) = \begin{cases} \lambda z + \frac{1}{2}\rho z^2 & \text{if} \quad \lambda + \rho z \geq 0; \\ -\frac{\lambda^2}{2\rho} & \text{otherwise.} \end{cases} \tag{4.10}$$

4.3 Experiments

Datasets. Following NACL (Murugesan *et al.*, 2023a), we use the ACDC and FLARE datasets with its setting. **ACDC** (Bernard *et al.*, 2018) contains 100 patient exams with cardiac MR volumes and their respective pixel-wise annotations. We follow the standard practices on this dataset, and extract 2D slices from the volumes, which are resized to 224×224. Furthermore, **FLARE** (Ma *et al.*, 2021b) includes 360 volumes of multiple organs in abdominal CTs, together with their corresponding segmentation masks, which are resampled to a common space and cropped to 192×192×30.

Baselines. We compare to relevant calibration losses, as well as to state-of-the-art methods for calibration in medical imaging segmentation: Focal Loss (FL) (Mukhoti *et al.*, 2020b), penalizing low-entropies (ECP) (Pereyra *et al.*, 2017), Label smoothing (LS) (Szegedy *et al.*, 2016), SVLS (Islam & Glocker, 2021), MbLS (Liu *et al.*, 2022b), NACL (Murugesan *et al.*, 2023a) and BWCR (Karani, Dey & Golland, 2023). As segmentation backbones, we have selected two well-known and popular networks, UNet (Ronneberger, Fischer & Brox, 2015a) and nnUNet (Isensee *et al.*, 2021).

Implementation details. For most of the compared methods, we use the hyperparameters values reported in (Murugesan *et al.*, 2023a): FL ($\gamma = 3$), LS ($\alpha = 0.1$), ECP ($\lambda = 0.1$), MbLS ($\lambda = 0.1$ and $m = 10$), SVLS ($\sigma = 2$) and NACL ($\lambda = 0.1$). Furthermore, for BWCR, the impact of the logit consistency is controlled by $\lambda_{min} = 0.01$, and $\lambda_{max} = 1$. Regarding the prior used in NACL and our method CRaC, we use the one proposed in (Murugesan *et al.*, 2023a), which is defined as $\tau_k = \sum_{i=1}^{d} y_i^k$, which is computed over a 3×3 patch. We train all the models during 100 epochs, with ADAM (Kingma & Ba, 2015) as optimizer and a batch size fixed to 16. The learning rate is set to 10^{-3} for the first 50 epochs, and reduced to 10^{-4} afterwards. Following (Murugesan *et al.*, 2023a), the models are trained on 2D slices, and the evaluation is performed over 3D volumes.

Evaluation. **Segmentation:** we employ common segmentation metrics in the medical domain, such as the DICE coefficient (DSC) and the 95% Hausdorff Distance (HD). **Calibration:** following recent works (Murugesan *et al.*, 2023a, 2025) we resort to the expected calibration error (ECE) (Naeini *et al.*, 2015a) on foreground classes, as in (Islam & Glocker, 2021), and Thresholded Adaptive Calibration Error (TACE) (threshold of 10^{-3}) (Nixon, Dusenberry, Zhang, Jerfel & Tran, 2019b). We further compute the Friedman rank (Friedman, 1937), to fairly compare the performance of different algorithms in various settings.

Table 4.1 **Quantitative performance.** Discriminative (DSC ↑, HD ↓) and calibration (ECE ↓, TACE ↓) metrics, using UNet as segmentation backbone. The best method is highlighted in bold, whereas the second best is underlined

	ACDC				FLARE				Friedman	Final
	DSC	HD	ECE	TACE	DSC	HD	ECE	TACE	Rank_F	Rank
FL ($\gamma = 3$)	0.620	7.30	0.153	0.224	0.834	6.65	0.053	0.145	7.88	8
ECP ($\lambda = 0.1$)	0.782	4.44	0.130	0.151	0.860	<u>5.30</u>	0.037	0.134	5.38	7
LS ($\alpha = 0.1$)	0.809	3.30	0.083	0.093	0.860	5.33	0.055	0.050	4.88	4
SVLS $_{\text{IPMI'21}}$	0.824	2.81	0.091	0.138	0.857	5.72	0.039	0.144	5.25	5
MbLS $_{\text{CVPR'22}}$	0.827	2.99	0.103	0.081	0.836	5.75	0.046	0.041	5.25	5
NACL $_{\text{MICCAI'23}}$	<u>0.854</u>	2.93	0.068	0.073	<u>0.868</u>	**5.12**	0.033	**0.031**	2.25	2
BWCR $_{\text{MICCAI'23}}$	0.841	<u>2.69</u>	**0.051**	0.075	0.848	5.39	**0.029**	0.059	3.13	3
CRaC (Ours)	**0.877**	**1.72**	<u>0.057</u>	**0.058**	**0.876**	5.52	**0.029**	<u>0.033</u>	1.75	1

Comparison to state-of-the-art calibration approaches. In Table 4.1 and 4.2, we present the quantitative results of our approach compared to a list of relevant state-of-the-art calibration approaches, when using UNet and nnUNet as segmentation backbones, respectively. In terms of **segmentation performance**, our proposed CRaC brings very competitive performance, typically ranking as best, or second best approach, regardless of the segmentation backbone employed.

Regarding **calibration**, the trend observed is similar, with CRaC providing well-calibrated models, either improving or at par with state-of-the-art for calibration. Furthermore, as it is common in evaluating many methods in multiple settings, we assess the **overall performance** with a multi-criteria analysis, the Friedman Rank. The results from this metric, which are reported at the right-most columns of both Tables 4.1 and 4.2, show that CRaC ranks at the first position, outperforming existing methods when a trade-off between calibration and segmentation performance is considered. Furthermore, the first rank position is maintained even when employing a more powerful backbone, i.e., nnUNet, consistently delivering the better segmentation-calibration compromise.

Table 4.2 **Quantitative performance.** Discriminative (DSC ↑, HD ↓) and calibration (ECE ↓, TACE ↓) using nnUNet (Isensee *et al.*, 2021) as segmentation backbone. The best method is highlighted in bold, whereas the second best is underlined

	ACDC				FLARE				Friedman	Final
	DSC	HD	ECE	TACE	DSC	HD	ECE	TACE	Rank_F	Rank
FL ($\gamma = 3$)	0.874	1.60	0.134	0.136	0.893	3.93	0.039	0.061	6.00	6
ECP ($\lambda = 0.1$)	0.889	<u>1.44</u>	0.067	0.112	0.873	5.85	0.046	0.131	6.00	6
LS ($\alpha = 0.1$)	**0.891**	**1.35**	0.067	0.066	0.891	3.61	0.062	0.047	4.00	4
SVLS $_{\text{IPMI'21}}$	0.883	1.69	0.059	0.111	0.894	4.02	0.026	0.115	5.13	5
MbLS $_{\text{CVPR'22}}$	0.886	1.46	0.057	0.052	0.891	3.65	0.031	0.031	3.50	3
NACL $_{\text{MICCAI'23}}$	0.884	1.52	<u>0.056</u>	0.059	**0.896**	<u>3.34</u>	**0.025**	**0.026**	2.50	2
$\text{BWCR}_{\text{MICCAI'23}}$ 22	0.864	1.82	0.063	0.079	0.868	4.47	0.041	0.099	6.63	8
CRaC (Ours)	**0.891**	1.48	**0.052**	**0.051**	<u>0.895</u>	**3.24**	0.029	<u>0.029</u>	1.88	1

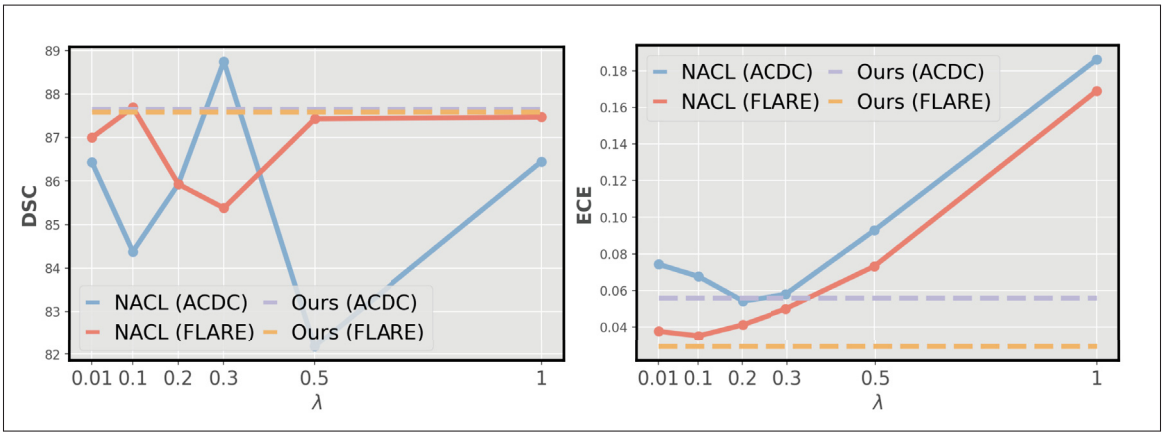

Figure 4.1 **Instability of NACL fine-tuning**. Discriminative (*left*) *vs.* calibration performance (*right*) as a function of λ in NACL (Murugesan *et al.*, 2023a)

Benefits compared to NACL. In this section we compare the sensitivity of NACL (Murugesan *et al.*, 2023a) to the choice of its λ value in Eq. 4.1, as our approach improves NACL by incorporating a mechanism to learn and adapt the class and region-wise penalty terms λ_{kr} in Eq. 4.3. We found that, despite performing at par in some settings, the performance of NACL significantly varies with the value of its penalty weight which, in addition, is dataset-dependent (Figure 4.1). For example, the left plot demonstrates that while setting $\lambda = 0.3$ in NACL yields the best discriminative performance in ACDC, it is substantially deteriorated in the FLARE dataset. Furthermore, the λ value that optimizes the discriminative performance may not be the same that minimizes the miscalibration issue. Thus, while one may argue that by fine-tuning λ in NACL can lead to improvements over CRaC (and only in certain settings), we advocate that performing a validation search in a dataset-basis is impractical for real-world problems, making of our approach an appealing solution.

4.4 Conclusion

We presented a novel approach to calibrate segmentation networks, which accounts for the inherent difficulties of different classes and regions. Results demonstrate that our approach outperforms existing approaches, becoming an excellent alternative to deliver high-performing and robust models.

CHAPTER 5

ROBUST CALIBRATION OF LARGE VISION-LANGUAGE ADAPTERS

Balamurali Murugesan[a] , Julio Silva-Rodriguez[a] , Ismail Ben Ayed[b] , Jose Dolz[a]

[a] Department of Software Engineering, École de Technologie Supérieure,
[b] Department of System Engineering, École de Technologie Supérieure,
1100 Notre-Dame West, Montreal, Quebec, Canada H3C 1K3
[c]

Paper published in European Conference on Computer Vision, November 2024

Presentation

This chapter presents the article *"Robust Calibration of large Vision-Language Adapters"* (Murugesan, Silva-Rodríguez, Ayed & Dolz, 2024b) published in European Conference on Computer Vision (**ECCV**), Pages. 147–165, 2024.

Abstract

This paper addresses the critical issue of miscalibration in CLIP-based model adaptation, particularly in the challenging scenario of out-of-distribution (OOD) samples, which has been overlooked in the existing literature on CLIP adaptation. We empirically demonstrate that popular CLIP adaptation approaches, such as Adapters, Prompt Learning, and Test-Time Adaptation, substantially degrade the calibration capabilities of the zero-shot baseline in the presence of distributional drift. We identify the increase in logit ranges as the underlying cause of miscalibration of CLIP adaptation methods, contrasting with previous work on calibrating fully-supervised models. Motivated by these observations, we present a simple and model-agnostic solution to mitigate miscalibration, by scaling the logit range of each sample to its zero-shot prediction logits. We explore three different alternatives to achieve this, which can be either integrated during adaptation or directly used at inference time. Comprehensive experiments on popular OOD classification benchmarks demonstrate the effectiveness of the proposed approaches in mitigating miscalibration while maintaining discriminative performance, whose

improvements are consistent across the three families of these increasingly popular approaches. The code is publicly available at: https://github.com/Bala93/CLIPCalib

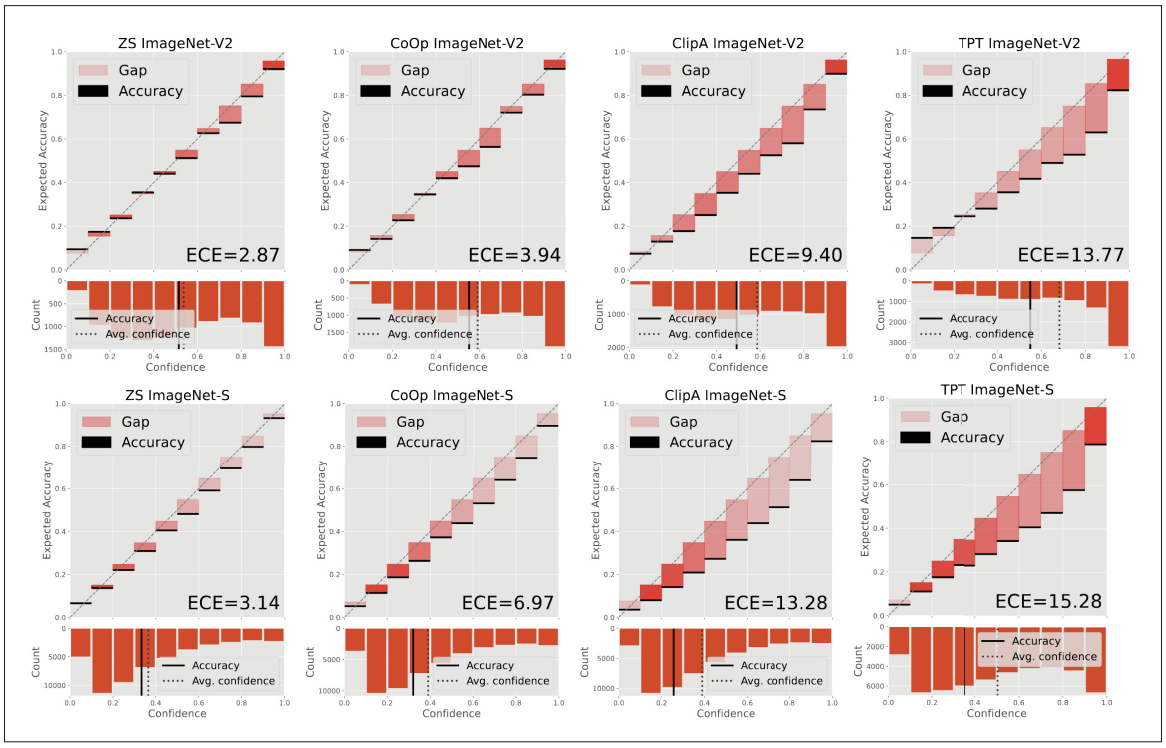

Figure 5.1 **CLIP-based adaptation methods are severely miscalibrated on Out-of-distribution (OOD) samples.** Three families of popular approaches to adapt CLIP under different scenarios, i.e., Prompt Learning (CoOp (Zhou *et al.*, 2022c)), Adapters (Clip-Ad (Gao *et al.*, 2024)) and Test-time prompt tuning (TPT (Shu *et al.*, 2022)), significantly degrade the miscalibration of the zero-shot baseline, despite improving its discriminative performance

5.1 Introduction

Deep learning is undergoing a paradigm shift with the emergence of pre-training large-scale language-vision models, such as CLIP (Radford *et al.*, 2021). These models, and more particularly the variants integrating vision transformers, have demonstrated impressive generalization capabilities in visual recognition tasks, yielding exceptional zero-shot and few-shot performance. Nevertheless, in a dynamic and evolving open world, machine learning applications inevitably encounter the challenge of out-of-distribution (OOD) data, which typically hinders the scalability

of these models to new domains. Existing literature based on CLIP faces this scenario with different solutions to exacerbate robustness. In particular, freezing the entire vision backbone to re-use these generalizable features has been a popular choice, especially in the low-data regime (Gao *et al.*, 2024; Zhou *et al.*, 2022c). Thus, CLIP adaptation during training typically resorts to Adapters (Gao *et al.*, 2024; Zhang *et al.*, 2022b) or Prompt Learning (Zhou *et al.*, 2022c,a) strategies, which leverage a few labeled samples to adapt the model with the hope that it will generalize properly to unseen related-domains. Furthermore, a more challenging scenario consists of adapting the model during inference without any access to labeled data, where a prevalent method is Test-Time Prompt Tuning (TPT) (Shu *et al.*, 2022).

While these strategies have further improved the discriminative performance of a zero-shot baseline, we have observed that the accuracy of the uncertainty estimates of the predictions, i.e., calibration, is significantly degraded (see 5.1), regardless of the family of adaptation models or setting. Thus, after adaptation, model predictions are often over-confident, even if they are wrong. This represents a major concern, as inaccurate uncertainty estimates can carry serious implications in safety-critical applications, such as healthcare, where CLIP is emerging as a popular strategy (Liu *et al.*, 2023b; Liang *et al.*, 2022b). Nevertheless, despite its importance, the miscalibration issue has been overlooked in the CLIP adaptation literature.

Motivated by these observations, this paper addresses this critical issue, which has been disregarded in current literature. Indeed, few-shot adaptation strategies, notably Prompt Learning and Adapters, are attracting wide attention recently, with an unprecedented surge in the number of methods proposed (Liu *et al.*, 2023b; Zhang *et al.*, 2022b; Yu *et al.*, 2023b; Silva-Rodriguez *et al.*, 2024; Hu *et al.*, 2022a; Zhou *et al.*, 2022c,a; Hantao Yao, 2023; Khattak *et al.*, 2023), albeit being a relatively recent research topic. Nevertheless, the main focus of this growing literature has been on improving the discriminative power of adapted models. Thus, given their increasing popularity, and quick adoption in real-world safety-critical problems, we believe that assessing the calibration performance of CLIP adaptation strategies in OOD scenarios is of paramount importance to deploy not only high-performing but also reliable models. We can summarize our contributions as follows:

1. We empirically demonstrate that popular CLIP adaptation strategies, such as Adapters, Prompt Learning, and Test-Time Prompt Tuning, substantially degrade the calibration capabilities of the zero-shot baseline in the presence of distributional drift.
2. For these adaptation strategies, we expose that the underlying cause of miscalibration is, in fact, the increase of the logit ranges. This contrasts with recent work in calibrating fully-supervised models (Wei *et al.*, 2022), which suggests that the inherent cause of miscalibration is the increase of its norm instead, due to the standard cross-entropy loss used for training.
3. Based on these observations, we present a simple, and model-agnostic solution, which consists in scaling the logit range of each sample based on the zero-shot logits. We further present several alternatives to accommodate our solution, which can be implemented either at training or inference time.
4. Comprehensive experiments on popular OOD classification benchmarks empirically demonstrate the effectiveness of our approaches to reduce the miscalibration error, while keeping the discriminative performance.

5.2 Related Work

5.2.1 Vision language models

Text-driven pre-training of image representation, so-called vision-language models (VLMs) is revolutionizing the paradigm of transfer learning. These models can integrate massive web-scrabbled data sources thus learning robust feature representations. In particular, models such as CLIP (Radford *et al.*, 2021) or ALIGN (Jia *et al.*, 2021) train joint multi-modal embedding spaces via contrastive learning of paired images and text, using dual encoder architectures. Such strong vision-language alignment has demonstrated robust open-vocabulary zero-shot generalization capabilities (Radford *et al.*, 2021; Zhai *et al.*, 2022). Given such potential, transferring pre-trained VLMs to a wide variety of tasks is gaining increasing popularity. Nevertheless, this process faces particular challenges. First, large-scale pre-training usually involves also scaling network

sizes, which is a computational bottleneck for low-resource adaptation scenarios. Second, recent attempts to fine-tune VLMs have demonstrated a deterioration of their robustness against domain drifts (Kumar *et al.*, 2022; Wortsman *et al.*, 2022), especially when available data is limited. Thus, an emerging core of recent literature is focusing on novel alternatives to overcome these limitations. More concretely, freezing the pre-trained backbone, and reusing such features by training a small set of parameters, via Prompt Learning (Zhou *et al.*, 2022c,a; Hantao Yao, 2023; Zhu *et al.*, 2023; Khattak *et al.*, 2023), or black-box Adapters (Gao *et al.*, 2024; Zhang *et al.*, 2022b; Yu *et al.*, 2023b; Silva-Rodriguez *et al.*, 2024; Ouali *et al.*, 2023; Li *et al.*, 2024; Zhang *et al.*, 2023), is getting increasing attention.

5.2.2 Prompt based learning

CLIP models have shown encouraging results by hand-crafting personalized text descriptions of the target visual representation (Menon & Vondrick, 2023). The automatizing of this cumbersome process raises the concept of Prompt Learning (PL) (Zhou *et al.*, 2022c), a family of methods to adapt CLIP that inserts a set of continuous learnable tokens in the original text prompt at the input of the VLM language encoder. While the CLIP model remains frozen, PL optimizes the most discriminative text input, given a few-shot support set (Zhou *et al.*, 2022c,a; Khattak *et al.*, 2023; Zhu *et al.*, 2023). CoOP (Zhou *et al.*, 2022c) represents one of the initial attempts to study the effect of prompt tuning on different tasks, and proposed to learn the prompt's context words. CoCoOP (Zhou *et al.*, 2022a), on the other hand, designed a simple network to predict the input text prompt through image features, as CoOP failed to match the zero-shot performance on generic tasks. TPT (Shu *et al.*, 2022) extends PL to address time-test adaptation scenarios by updating the prompt for a batch with original and augmented samples through entropy minimization.

5.2.3 Black-box Adapters

Prompt Learning involves using the CLIP's encoder throughout the entire training process as the backpropagation of the gradient has to pass through it to update the prompts, which results in large

computational constraints (Gao *et al.*, 2024). Adapter-based techniques provide an alternative to Prompt Learning for aligning to downstream tasks, leveraging uniquely pre-computed features with minimal additional parameters. A base version of such methods involves training a linear classifier via logistic regression, typically referred to as Linear Probing (Radford *et al.*, 2021). Nevertheless, leveraging only the vision features does not fully exploit the potential of VLMs. To this end, several methods have proposed enhanced Adapters, which further rely on zero-shot text-driven class-wise prototypes. In particular, Clip-Adapter (Gao *et al.*, 2024) introduced additional fully connected layers and operated on the vision or language branch through residual style feature combination. Training-free methods such as Tip-Adapter (Zhang *et al.*, 2022b) resorted to a key-value cache model based on the available few-shot supports. Likewise, TaskRes (Yu *et al.*, 2023b) introduced additional learning parameters and applied a residual modification of the text representation, which led to a better initialization when learning from few-shot supervision. More recently, (Silva-Rodriguez *et al.*, 2024) provided a wider look at the coupling of vision and text features in such Adapters, by pointing out that these methods largely build up their improved performance on initializing the logistic classifier weights with the zero-shot prototypes, proposing a simple solution, coined CLAP, for a better distillation of such prototypes.

5.2.4 Model calibration

Calibrating the confidence of deep learning models is paramount in developing reliable solutions, as the confidence is expected to correlate with correctness. Given the importance and the potential impact of miscalibration, a growing literature to address this issue has emerged in the last years. Post-processing techniques have been widely used to achieve calibration, wherein a linear transformation (Guo *et al.*, 2017b; Tomani, Cremers & Buettner, 2022; Joy, Pinto, Lim, Torr & Dokania, 2023) is applied to the predicted logits before converting it to softmax. Nevertheless, an important limitation is that these transformations are obtained using held-out validation data, which is assumed to follow the same distribution as the test data, limiting their applicability in the presence of domain drifts (Ovadia *et al.*, 2019). A popular alternative consists in calibrating the networks at training time. This can be achieved by incorporating

explicit penalties that either penalize overconfident softmax predictions (Pereyra *et al.*, 2017; Larrazabal, Martinez, Dolz & Ferrante, 2023a; Cheng & Vasconcelos, 2022; Park, Noh, Oh, Baek & Ham, 2023) or encourage small logit differences (Murugesan *et al.*, 2023a; Liu *et al.*, 2022b, 2023a). Furthermore, (Müller *et al.*, 2019b; Mukhoti *et al.*, 2020b) demonstrated that popular classification losses, such as Focal Loss (Lin *et al.*, 2017) or Label Smoothing (Szegedy *et al.*, 2016), integrate an implicit term that maximizes the entropy of the network predictions, thus favoring low-confidence models. Other works to improve the accuracy of the uncertainty estimates during training include the use of MixUp (Thulasidasan *et al.*, 2019b; Zhang, Deng, Kawaguchi & Zou, 2022a), or enforcing a constant vector norm on the logits (Wei *et al.*, 2022), among others. Nevertheless, all these methods have been proposed in the context of fully-supervised models, and the calibration of Prompt Learning and Adapter-based methods for CLIP remains unexplored in the literature.

5.3 Background

5.3.1 CLIP Zero-Shot Classification

CLIP (Radford *et al.*, 2021) is a large vision-language model, trained via contrastive learning to produce visual representations from images $\boldsymbol{x}$ paired with their associated text descriptions T. To do so, CLIP consists of an image encoder $\boldsymbol{\theta}$ and a text encoder $\boldsymbol{\phi}$. This generates the corresponding vision $\boldsymbol{z} \in \mathbb{R}^d$ and class text $\boldsymbol{w}_k \in \mathbb{R}^d$ embeddings, which are typically projected into an ℓ_2-normalized shared embedding space. Given a new task consisting in visually discriminating between K categories, the set containing all the text embeddings for all the K classes can be denoted as $\mathcal{W} = \{\boldsymbol{w}_k\}_k^K$, with $\boldsymbol{w}_k = \boldsymbol{\phi}(\text{“A photo of a [class}_k\text{]”})$. At inference, this learning paradigm enables zero-shot prediction. More concretely, for a given set of K classes, and an ensemble of N different prompts per category, we can generate the set of available prompts as $\mathcal{T} = \{\{T_{n,k}\}_{n=1}^N\}_{k=1}^K\}$. Then, a popular strategy (Radford *et al.*, 2021; Gao *et al.*, 2024; Wortsman *et al.*, 2022) consists in obtaining a class zero-shot prototype, which is computed as $\boldsymbol{t}_k = \frac{1}{N}\sum_{n=1}^N \boldsymbol{\phi}(T_{n,k})$. Then, for a given test image, the zero-shot prediction,

$\boldsymbol{p} = (p_k)_{1 \leq k \leq K}$, can be obtained as:

$$p_k = \frac{\exp\left(\boldsymbol{z}^\top \boldsymbol{t}_k / \tau\right)}{\sum_{j=1}^{K} \exp\left(\boldsymbol{z}^\top \boldsymbol{t}_j / \tau\right)} \tag{5.1}$$

where τ is a temperature hyperparameter, whose value is learned during training, and $\boldsymbol{z}^\top \boldsymbol{t}$ denotes the dot product operator[1].

5.3.2 Adaptation to novel tasks

Let us now consider a support set that contains a few labeled samples $\mathcal{S} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{S}$, with $\boldsymbol{y} \in \{0, 1\}^K$ the ground truth vector associated with $\boldsymbol{x}$. The vector of predicted logits of a given image i is defined as $\boldsymbol{l}_i = (l_{ik})_{1 \leq k \leq K}$. In Prompt Learning methods, such as CoOp (Zhou *et al.*, 2022c) or KgCoOp (Zhou *et al.*, 2022a), the adaptation is done by modeling the input text T_k of a given class k as learnable continuous vectors. Thus, in contrast with zero-shot inference, where the resulting text embeddings are obtained as the mean over the different pre-defined prompts, in Prompt Learning these are optimized. To generate the logits, the learnable prompts are combined with the fixed visual embedding from the test image i, such that $l_{ik} = \boldsymbol{z}_i^\top \boldsymbol{t}_k / \tau$, which can then be integrated into 5.1 to minimize the cross entropy loss over the few labeled shots. The family of methods commonly referred to as Adapters (Gao *et al.*, 2024; Zhang *et al.*, 2022b; Yu *et al.*, 2023b; Silva-Rodriguez *et al.*, 2024) proceeds differently, and learns transformations over the visual and text embeddings, yielding the following logits $l_{ik} = \boldsymbol{\theta}_a(\boldsymbol{z}_i, \boldsymbol{t}_k, \tau)$, where $\boldsymbol{\theta}_a$ is the set of learnable parameters of the Adapter. A more challenging scenario consists in adapting the text prompts at inference, which is commonly referred to as test-time prompt tuning (Shu *et al.*, 2022). As this setting does not include few-shot supports to adapt the prompts, the supervised cross-entropy objective is replaced by an unsupervised minimization of the Shannon entropy. Thus, regardless of the method selected, the objective is to optimize either $\boldsymbol{t}_k$ (Prompt Learning and test-time prompt tuning) or $\boldsymbol{\theta}_a$ (Adapters) to minimize either the CE over the

[1] As vectors are ℓ_2-normalized, the dot product between these two vectors is equivalent to their cosine similarity.

softmax predictions obtained from the few-shots, or the Shannon entropy on the test samples predictions at inference.

5.4 Constraining logits during adaptation

5.4.1 Impact of adaptation in logits

To understand the impact on calibration of using the cross-entropy (CE) loss to adapt CLIP, let us decompose the logit vector $\boldsymbol{l}$ into its Euclidean norm $\|\boldsymbol{l}\| = \sqrt{l_1^2 + ... + l_K^2}$, (*magnitude*) and its unit vector $\hat{\boldsymbol{l}}$ (*direction*), such that $\boldsymbol{l} = \|\boldsymbol{l}\|\hat{\boldsymbol{l}}$. Considering now the *magnitude* and *direction* of the logit vector, the general form of the cross-entropy loss over a given support sample, using the softmax probabilities in 5.1, can be formulated as:

$$- \log \frac{\exp\left(\|\boldsymbol{l}_i\|\hat{l}_{ik}\right)}{\sum_{j=1}^{K} \exp\left(\|\boldsymbol{l}_i\|\hat{l}_{ij}\right)} \tag{5.2}$$

This view of the cross-entropy implies that the direction of the logit vector $\hat{\boldsymbol{l}}_i$ determines the predicted class of the image i. Thus, if the predicted category is incorrect, $\hat{\boldsymbol{l}}_i$ will change to match the target class (k) provided in the one-hot encoded label. Once the network prediction is correct, i.e., $y_i = \arg\max_j(l_{ij})$, the direction of the vector will remain unchanged. Nevertheless, the nature of the cross-entropy loss will favor higher softmax probabilities for the predicted class. Recent literature (Wei *et al.*, 2022) suggests that this is achieved by increasing $\|\boldsymbol{l}_i\|$, indicating that the miscalibration issue originates from the augmentation of the logit norm. Nevertheless, in what follows we refute this argument and advocate for **the increase of the logits range as the potential cause of miscalibration.**

Proposition 1. *Let us consider the softmax cross entropy loss, where $\sigma(\cdot)$ denotes the softmax function. Assume that $\boldsymbol{l} \geq \mathbf{0}$. Then, for any scalar $a > 0$, $\sigma_k(\boldsymbol{l}) = \sigma_k(\boldsymbol{l} + a)\ \forall k$, and $\|\boldsymbol{l} + a\mathbf{1}\| > \|\boldsymbol{l}\|$, where $\mathbf{1}$ denotes the vector of ones.*

Prop. 1 demonstrates that adding a strictly positive constant value $a \in \mathbb{R}_{++}$ to all the logits increases the norm of the vector $\boldsymbol{l}$, but this does not necessarily lead to more confident predictions, whose probability scores remain unchanged.

Proposition 2. *Let* $R(\boldsymbol{l}) = \max(\boldsymbol{l}) - \min(\boldsymbol{l})$ *denotes the range of logit vector* $\boldsymbol{l}$*, where* $\max(\boldsymbol{l})$ *(respectively* $\min(\boldsymbol{l})$*) denotes the largest (respectively smallest) value among the elements of vector* $\boldsymbol{l}$*. Then, for any given scalar* $a > 1$*, and for* $k = \arg\max_j(l_j)$*, we have* $\sigma_k(a\boldsymbol{l}) > \sigma_k(\boldsymbol{l})$ *and* $R(a\boldsymbol{l}) > R(\boldsymbol{l})$*.*

From the above proposition we find that increasing the range of a given logit vector results in higher softmax probability values. Thus, contrary to the widely spread belief that increasing the logit norm hinders model calibration, we argue that this effect of *logit distance magnification*, which yields higher softmax distributions, is a potential source of miscalibration[2]. This explains why, even though adaptation of CLIP yields performance gains in terms of accuracy, adapted models are worse calibrated than a zero-shot baseline. Furthermore, this analysis is supported empirically by the observations depicted in 5.2, where can observe that, while calibration has been degraded in the adapted models, the logit norm of their predictions has substantially decreased.

5.4.2 Our solution

From our previous analysis and empirical observations, we can derive that: *i)* despite improving their classification performance, state-of-the-art strategies to adapt CLIP suffer from miscalibration, particularly compared to the original zero-shot predictions, and *ii)* one of the main causes arises from the logit magnification issue introduced by the cross-entropy loss used during adaptation.

In light of these findings, we propose a simple but effective solution that can alleviate the miscalibration issue in CLIP adaptation. More concretely, we propose to constraint the range of

[2] Note that the same reasoning applies to TPT (Shu *et al.*, 2022), whose learning objective to adapt the CLIP baseline is to minimize the Shannon entropy of the softmax distribution.

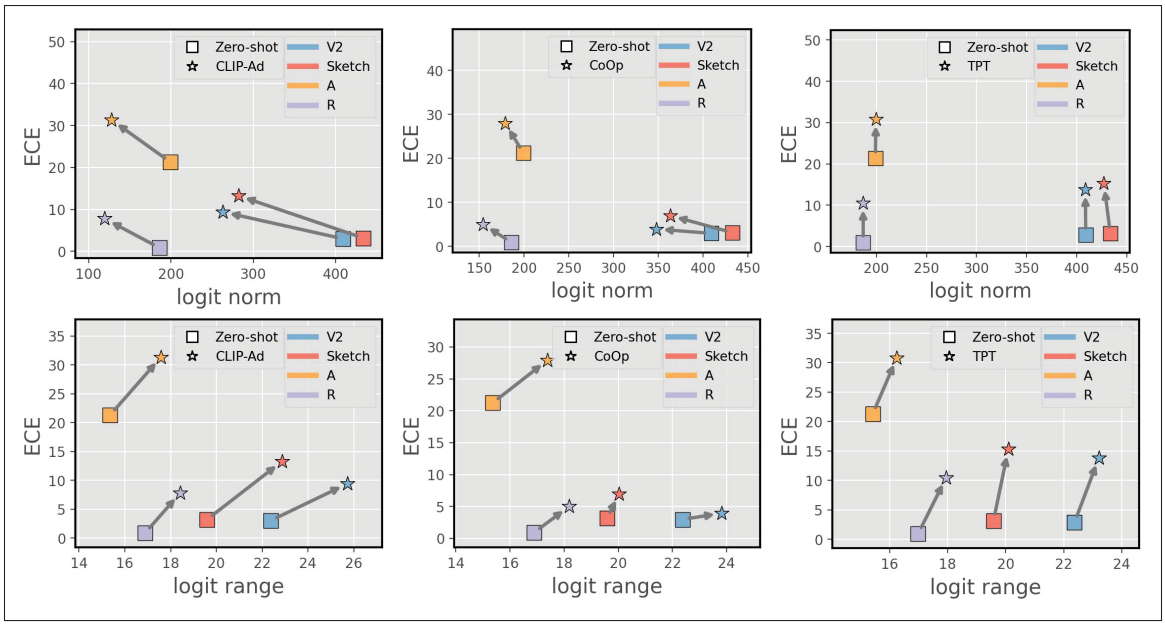

Figure 5.2 **Logit norm or logit range as the source of miscalibration?** These figures clearly show that when the calibration of the zero-shot (ZS) model is degraded, the logit norm of its predictions is reduced (*top*), which discards an increase of the logit norm as the main cause for miscalibration. In contrast, there exists a correlation between the increase of the logit ranges and miscalibration (*bottom*)

the logits during the training of a main objective $\mathcal{H}$, which results in the following constrained problem:

$$\begin{aligned} &\text{minimize} && \mathcal{H}(\boldsymbol{Y}, \boldsymbol{P}) \\ &\text{subject to} && l_i^{\text{ZS-min}}\mathbf{1} \leq \boldsymbol{l}_i \leq l_i^{\text{ZS-max}}\mathbf{1} \qquad \forall i \in \mathcal{D}, \end{aligned} \tag{5.3}$$

where $\boldsymbol{Y}$ and $\boldsymbol{P}$ are matrices containing the sample-wise ground-truth and softmax-prediction vectors for all the samples involved in the training, $l_i^{\text{ZS-min}}$ and $l_i^{\text{ZS-max}}$ are the min and max logit magnitudes of the zero-shot prediction for sample $\boldsymbol{x}_i$. $\mathcal{D}$ denotes a given set of available samples. Furthermore, in the test-time prompt tuning setting, we simply need to replace $\boldsymbol{Y}$ by $\boldsymbol{P}$ in 5.3. Directly solving the constrained problem in 5.3 in the context of deep models is

not trivial (Marquez Neila *et al.*, 2017), and Lagrangian-dual optimization has been typically avoided in modern deep networks involving millions of parameters. To address this issue, we propose several alternatives to approximate the constrained problem presented in 5.3, which are detailed below.

5.4.3 Zero-shot logit normalization during training (ZS-Norm)

The constraint in the presented problem, i.e., $l_i^{\text{ZS-min}}\mathbf{1} \leq \boldsymbol{l}_i \leq l_i^{\text{ZS-max}}\mathbf{1}, \forall i \in \mathcal{D}$, can be integrated into the main objective by transforming the logits before computing the CE loss over the support set samples (here $\mathcal{D} = \mathcal{S}$). More concretely, the modified learning objective can be defined as:

$$\mathcal{H}(\boldsymbol{Y}, \boldsymbol{P}) = -\sum_{i \in \mathcal{S}} \sum_{k=1}^{K} y_{ik} \log \frac{\exp\left(l'_{ik}\right)}{\sum_{j=1}^{K} \exp\left(l'_{ij}\right)}, \tag{5.4}$$

where $\boldsymbol{l}'_i$ denotes the zero-shot normalized logit vector of $\boldsymbol{x}_i$, obtained as:

$$\boldsymbol{l}'_i = \frac{(l_i^{\text{ZS-max}} - l_i^{\text{ZS-min}})}{(l_i^{\text{max}} - l_i^{\text{min}})}(\boldsymbol{l}_i - l_i^{\text{min}}\mathbf{1}) + l_i^{\text{ZS-min}}\mathbf{1}, \tag{5.5}$$

with $l_i^{\text{max}} = \max_j(l_{ij})$ and $l_i^{\text{min}} = \min_j(l_{ij})$, respectively. While the calibration strategy formalized in Eq. 5.4 forces the *direction* of the logit vector to match the correct category encoded in the one-hot label, its *magnitude* is normalized according to the ZS logit range of image $\boldsymbol{x}_i$. Note that this is different from the solution presented in (Wei *et al.*, 2022), as the logit values are normalized by the logit norm, which does not guarantee that the logit values will be in a certain range.

5.4.4 Integrating explicit constraints in the learning objective (Penalty)

The problem in 5.3 can also be approximated by an unconstrained problem, for example by transforming the enforced inequality constraints into penalties, which are implemented with the ReLU function. The resulting learning objective can be formally defined as:

$$\min_{\boldsymbol{\theta}} \quad \mathcal{H}(\boldsymbol{Y}, \boldsymbol{P}) + \lambda \sum_{i \in \mathcal{S}} \sum_{k=1}^{K} (\text{ReLU}(l_{ik} - l_i^{\text{ZS-max}}) + \text{ReLU}(l_i^{\text{ZS-min}} - l_{ik})), \tag{5.6}$$

where λ controls the trade-off between the main loss and the penalties. The intuition behind the penalties is that when the constraint in Eq. (5.3) is not satisfied, i.e., there exist logit magnitudes outside the zero-shot logit range, the value of the penalty term increases, backpropagating gradients to modify the logit values according to the enforced constraint. We would like to stress that a natural solution to tackle the constrained problem in 5.6 would be the use of Lagrangian multipliers. Nevertheless, as stated earlier, in the context of deep learning, these methods suffer from several well-known limitations, which include training instability and non-convergence due to the difficulty of convexifying loss functions (Sangalli, Erdil, Hötker, Donati & Konukoglu, 2021; Birgin, Castillo & Martínez, 2005; Bertsekas, 1996b). Thus, despite its simplicity, the use of penalties has proven to be effective in constraining deep models on a myriad of problems, such as image segmentation (Kervadec *et al.*, 2019c), adversarial attacks (Rony *et al.*, 2021), or modeling thermal dynamics (Drgoňa, Tuor, Chandan & Vrabie, 2021).

5.4.5 Sample-adaptive logit scaling (SaLS)

Last, we explore a simple but efficient solution that is closely related to temperature scaling (TS) (Guo *et al.*, 2017b). In particular, TS is a single-parameter variant of Platt scaling (Platt *et al.*, 1999), which consists in learning the scaling hyperparameter τ in 5.1. While this strategy has led to very competitive results, it requires an external validation set to fine-tune the value of τ, which limits its use to learning scenarios with abundant labeled data and absence of distributional drifts (Ovadia *et al.*, 2019). Furthermore, τ is fixed for a whole dataset, which is suboptimal from a

sample-wise standpoint. To alleviate these issues, we propose to use the logit normalization defined in 5.5 at inference time to obtain the final softmax probability in 5.1. More concretely, for each sample *i* to be classified, we compute its zero-shot prediction, whose *min* and *max* logit values are used in 5.5 to scale the logit distribution of that sample *i* provided by the adaptation method selected. This can be viewed as an *unsupervised sample-wise temperature scaling* during testing, which does not require additional validation samples to fix its value, and adapts to the specificity of each sample, regardless of distributional drifts.

5.5 Experiments

5.5.1 Setup

5.5.1.1 Datasets

We use popular datasets for benchmark few-shot (Zhou *et al.*, 2022c; Yu *et al.*, 2023b) and test-time (Shu *et al.*, 2022) CLIP adaptation. **Domain Generalization:** the adaptation *robustness* to domain shifts is evaluated using ImageNet (Deng *et al.*, 2009) distributions. Concretely, we sample a 16-shot training subset from ImageNet's training partition which is directly evaluated on out-of-distribution test data from ImageNetV2 (Recht, Roelofs, Schmidt, & VaishaalShankar, 2019), ImageNet-Sketch (Wang, Ge, Lipton & Xing, 2019c), ImageNet-A (Hendrycks, Zhao, Basart, Steinhardt & Song, 2019), and ImageNet-R (Hendrycks *et al.*, 2021). **Fine-grained tasks:** calibration during test-time adaptation is assessed on an assembly of 11 datasets that include heterogeneous discriminative tasks. These include Imagenet (Deng *et al.*, 2009), Caltech101 (Fei-Fei, Fergus & Perona, 2004), OxfordPets (Parkhi, Vedaldi, Zisserman & Jawahar, 2012), StanfordCars (Krause, Stark, Deng & Fei-Fei, 2012), Flowers102 (Nilsback & Zisserman, 2008), Food101 (Bossard, Guillaumin & Van Gool, 2014), FGVCAircraft (Maji, Kannala, Rahtu, Blaschko & Vedaldi, 2013), SUN397 (Xiao, Hays, Ehinger, Oliva & Torralba, 2010), DTD (Cimpoi, Maji, Kokkinos, Mohamed & Vedaldi, 2014), EuroSAT (Helber, Bischke,

Dengel & Borth, 2018), and UCF101 (Soomro, Zamir & Shah, 2012) datasets. Note that for test time adaptation we uniquely employed their corresponding test partitions.

5.5.1.2 Selected methods

Our proposed calibration framework is agnostic to any adaptation strategy of zero-shot models. We evaluate its performance across different popular settings and state-of-the-art methods for CLIP adaptation. **Prompt Learning (PL):** CoOp (Zhou *et al.*, 2022c), CoCoOp (Zhou *et al.*, 2022a), ProGrad (Zhu *et al.*, 2023) and MaPLe (Khattak *et al.*, 2023) are considered as the baselines. **Adapters:** CLIP-Adapter (Gao *et al.*, 2024), TIP-Adapter (Zhang *et al.*, 2022b), and TaskRes(Yu *et al.*, 2023b) are used. **Test-Time Adaptation:** TPT (Shu *et al.*, 2022) is selected as the primary method for test time prompt tuning, together with C-TPT (Yoon *et al.*, 2024), a concurrent method recently proposed for calibrating TPT.

5.5.1.3 CLIP adaptation

We now describe the experimental details for training the selected adaptation methods. **Backbones:** All experiments build upon CLIP (Radford *et al.*, 2021), using its ResNet-50 (He *et al.*, 2016b) and ViT-B/16 (Dosovitskiy *et al.*, 2021a) pre-trained weights. **Text prompts:** The textual descriptions for zero-shot representation of the target concepts used are the hand-crafted text prompts used in CoOp (Zhou *et al.*, 2022c). **Image augmentations:** For few-shot adaptation, we applied random zoom, crops, and flips, following (Zhou *et al.*, 2022a; Yu *et al.*, 2023b). Regarding Prompt Learning methods, these transformations are applied continuously during training, while for Adapters, since feature representations are pre-computed, the number of augmentations per support sample is set to 20, following (Silva-Rodriguez *et al.*, 2024). Finally, regarding test-time prompt tuning (TPT), we employed AugMix (Hendrycks *et al.*, 2020) as in (Shu *et al.*, 2022) to form a 64-image batch from each original image. **Training details:** Adapters are trained following the recent benchmark in (Silva-Rodriguez *et al.*, 2024). We optimized the Adapters for 300 epochs, using SGD optimizer with a Momentum of 0.9 and an initial learning rate of 0.1. In the case of PL, we set the context length of the prompt to 4

and trained CoOp and CoCoOp for 50 and 10 epochs, respectively. We set the same training schedule, optimizer, and learning as in (Zhou *et al.*, 2022c). For ProGrad and MaPLe, we follow the training settings considered for domain generalization reported in theirs respective works (Zhu *et al.*, 2023; Khattak *et al.*, 2023). Likewise, for TPT, we optimized the learned prompt by doing a single step with AdamW optimizer, with the learning rate set to 0.005, as in (Shu *et al.*, 2022).

5.5.1.4 Evaluation metrics

To measure the discriminative performance of the different methods, we use classification accuracy (ACC). In terms of calibration, we follow the standard literature and resort to the Expected Calibration Error (ECE). In particular, with N samples grouped into M bins $\{b_1, b_2, \ldots, b_K\}$, the ECE is calculated as: $\sum_{m=1}^{M} \frac{|b_m|}{N} |\text{acc}(b_m) - \text{conf}(b_m)|$, where acc $(\cdot)$ and conf $(\cdot)$ denote the average accuracy and confidence in bin b_m.

5.5.1.5 Calibration details

We introduced three different alternatives to alleviate the miscalibration of adapted models (5.4.2). For **ZS-Norm** and **Penalty**, we incorporated such modifications during training (i.e adaptation), and kept all implementation details previously presented. Furthermore, the penalty-based calibration weight λ in Eq. 5.6 is set to 10 and remains fixed across all settings.

5.5.2 Results

5.5.2.1 Task 1: Few-shot domain generalization

Table 5.1 introduces the average few-shot generalization (OOD) results using black-box Adapters, whereas Table 5.2 presents the same for PL approaches. First, results consistently show a miscalibration phenomenon when CLIP models are adapted, regardless of the CLIP backbone used, or the transferability approach. **Few-shot Adapters calibration:** We find that

miscalibration is especially occurring in few-shot black-box Adapters. For example, Clip-Ad or TaskRes in 5.1 (a) show ECE increments of +8.3 and +4.0 respectively. This is further magnified when using the popular TIP-Adapter method. **Few-shot PL calibration:** PL approaches are relatively more robust in this setting (e.g. +3.8 CoOp in 5.2 (a)). **On the impact of logit range regularization:** Results show the potential of logit range scaling among its different proposed variants, *improving calibration for all Prompt Learning approaches*, and most of the used Adapters. **Impact of different strategies to adjust logit range:**. The only strategy that does not allow for consistent performance gains is **ZS-Norm**, which deteriorates performance in some Adapters (see Clip-Ad in Table 5.1). We believe that the re-parameterization in Eq 5.4 might not properly prevent logit range de-adjustment before normalization, and thus overfit to the few support samples. In contrast, **Penalty** constraint directly regularizes such values, showing consistent ECE decreases for both Adapters (e.g. -22.0 for TIP-Ad(f) using ViTs, or -4.3 for CLIP-Ad using RN50) and PL (e.g. -2.9 for CoOp using RN50, or -0.94 for CoCoOp using ViTs). Interestingly, as a side effect, we also observed accuracy improvements for domain generalization for several methods. Nevertheless, the best calibration performance is provided by a simple, yet effective post-processing standardization, **SaLS**. This is especially relevant, since

Table 5.1 **Results for robust Adapters calibration.** The average over the four ImageNet OOD datasets is reported. In brackets, we highlight the difference with respect to each baseline, to stress the impact of the proposed methods (**ZS-Norm**, **Penalty**, and **SaLS**)

(a) ResNet-50

Method	Avg. OOD	
	ACC	ECE
Zero-Shot	40.62	7.18
CLIP-Ad	34.07	15.45
w/ **ZS-Norm**	30.06$_{(-4.01)}\downarrow$	21.27$_{(+5.82)}\uparrow$
w/ **Penalty**	**35.20**$_{(+1.13)}\uparrow$	11.22$_{(-4.23)}\downarrow$
w/ **SaLS**	34.07	**8.95**$_{(-6.50)}\downarrow$
TIP-Ad(f)	41.45	19.04
w/ **ZS-Norm**	41.73$_{(+0.28)}\uparrow$	19.80$_{(+0.76)}\uparrow$
w/ **Penalty**	**43.73**$_{(+2.28)}\uparrow$	12.18$_{(-6.86)}\downarrow$
w/ **SaLS**	41.45	**8.13**$_{(-10.91)}\downarrow$
TaskRes	41.18	11.25
w/ **ZS-Norm**	**41.30**$_{(+0.12)}\uparrow$	9.07$_{(-2.18)}\downarrow$
w/ **Penalty**	41.29$_{(+0.11)}\uparrow$	10.62$_{(-0.63)}\downarrow$
w/ **SaLS**	41.18	**9.03**$_{(-2.22)}\downarrow$

(b) ViT-B/16

Method	Avg. OOD	
	ACC	ECE
Zero-Shot	57.15	4.78
CLIP-Ad	50.61	7.82
w/ **ZS-Norm**	49.73$_{(+0.88)}\downarrow$	12.53$_{(+4.71)}\uparrow$
w/ **Penalty**	**51.59**$_{(+0.98)}\uparrow$	6.38$_{(-1.44)}\downarrow$
w/ **SaLS**	50.61	**4.38**$_{(-3.44)}\downarrow$
TIP-Ad(f)	25.86	63.63
w/ **ZS-Norm**	41.64$_{(+15.78)}\uparrow$	58.27$_{(-5.36)}\downarrow$
w/ **Penalty**	**49.23**$_{(+23.37)}\uparrow$	**40.98**$_{(-22.65)}\downarrow$
w/ **SaLS**	25.86	44.37$_{(-19.26)}\downarrow$
TaskRes	58.01	7.52
w/ **ZS-Norm**	**58.41**$_{(+0.40)}\uparrow$	**5.72**$_{(-1.80)}\downarrow$
w/ **Penalty**	58.31$_{(+0.30)}\uparrow$	6.65$_{(-0.87)}\downarrow$
w/ **SaLS**	58.01	6.21$_{(-1.31)}\downarrow$

Table 5.2 **Results for robust Prompt Learning calibration.** The average over the four ImageNet OOD datasets is reported. In brackets, we highlight the difference with respect to each baseline, to stress the impact of the proposed methods (**ZS-Norm**, **Penalty** and **SaLS**)

(a) ResNet-50

Method	Avg. OOD	
	ACC	ECE
Zero-Shot	40.62	7.18
CoOp	40.86	10.97
w/ **ZS-Norm**	41.59$_{(+0.73)}\uparrow$	10.19$_{(-0.78)}\downarrow$
w/ **Penalty**	**41.87**$_{(+1.01)}\uparrow$	8.06$_{(-2.91)}\downarrow$
w/ **SaLS**	40.86	**7.82**$_{(-3.15)}\downarrow$
CoCoOp	43.36	7.69
w/ **ZS-Norm**	43.70$_{(+0.34)}\uparrow$	7.12$_{(-0.57)}\downarrow$
w/ **Penalty**	**43.86** $_{(+0.50)}\uparrow$	**6.15** $_{(-1.54)}\downarrow$
w/ **SaLS**	43.36	6.82$_{(-1.87)}\downarrow$
ProGrad	42.32	7.66
w/ **ZS-Norm**	42.21 $_{(+0.11)}\uparrow$	7.98 $_{(+0.32)}\uparrow$
w/ **Penalty**	**42.57**$_{(+0.25)}\uparrow$	**6.84**$_{(-0.82)}\downarrow$
w/ **SaLS**	42.32	6.90$_{(-0.76)}\downarrow$

(b) ViT-B/16

Method	Avg. OOD	
	ACC	ECE
Zero-Shot	57.15	4.78
CoOp	58.41	6.61
w/ **ZS-Norm**	58.75$_{(+0.34)}\uparrow$	**4.35**$_{(-2.26)}\downarrow$
w/ **Penalty**	**59.18**$_{(+0.77)}\uparrow$	4.91$_{(-1.70)}\downarrow$
w/ **SaLS**	58.41	4.90$_{(-1.71)}\downarrow$
CoCoOp	59.74	4.83
w/ **ZS-Norm**	59.90$_{(+0.16)}\uparrow$	3.94$_{(-0.89)}\downarrow$
w/ **Penalty**	**60.20** $_{(+0.46)}\uparrow$	**3.89** $_{(-0.94)}\downarrow$
w/ **SaLS**	59.74	4.81$_{(-0.00)}\sim$
MaPLe	60.07	4.13
w/ **ZS-Norm**	60.09$_{(+0.02)}\uparrow$	**3.59**$_{(-0.14)}\downarrow$
w/ **Penalty**	**60.62** $_{(+0.55)}\uparrow$	3.78 $_{(-0.35)}\downarrow$
w/ **SaLS**	60.07	4.38$_{(+0.25)}\uparrow$

this method *does not require any modification of the adaptation strategy*, and can be potentially applied to the output of any few-shot model.

5.5.2.2 Task 2: Test Time Adaptation (TTA)

We report in Table 5.3 the performance for test-time prompt tuning across 11 fine-grained adaptation datasets for ResNet-50 backbone. Our results show that compared to zero-shot prediction, TPT largely deteriorates the calibration. Despite this degradation is somehow alleviated by C-TPT, further integrating our approaches show promising potential for better calibration of such methods, with consistent improvements for both strategies (e.g.,-2.0 and -0.9 in ECE for TPT and C-TPT with **SaLS**).

5.5.2.3 Further constraining the logit range to smaller values

ZS predictions are well calibrated. Nevertheless, during adaptation, the model improves its discriminative performance at the cost of degrading its calibration capabilities. While in this work we advocate for increases of the logit range as a cause of miscalibration, decreasing this range should be done with care. In particular, further decreasing the logit range approaches a

Table 5.3 **Test-time Prompt Learning calibration.** Results for the popular TPT, as well as the concurrent work in (Yoon *et al.*, 2024), with ResNet-50 backbone, where our three solutions are implemented

		Avg.	INet	CAL	PET	CAR	FLW	FOO	AIR	SUN	DTD	SAT	UCF
ACC	Zero-shot	56.03	58.17	85.68	83.62	55.75	61.67	73.96	15.69	58.82	40.43	23.69	58.90
	TPT	58.03	60.74	87.22	84.49	58.36	62.81	74.97	17.58	61.17	42.08	28.40	60.61
	w/ **ZS-Norm**	57.94	60.69	87.38	84.41	58.45	62.12	75.01	17.13	61.09	41.96	28.53	60.59
	w/ **Penalty**	57.69	60.74	87.06	84.30	58.13	61.84	75.17	17.22	61.11	42.02	26.60	60.35
	w/ **SaLS**	**58.03**	60.74	87.22	84.49	58.36	62.81	74.97	17.58	61.17	42.08	28.40	60.61
	C-TPT	57.54	60.02	87.18	83.65	56.41	64.80	74.89	16.62	60.72	41.55	27.06	60.01
	w/ **ZS-Norm**	**57.63**	60.00	87.06	83.65	56.57	65.04	74.82	16.86	60.58	41.61	27.51	60.27
	w/ **Penalty**	57.52	60.06	86.94	83.51	56.78	64.76	74.88	16.29	60.67	41.90	26.63	60.32
	w/ **SaLS**	57.54	60.02	87.18	83.65	56.41	64.80	74.89	16.62	60.72	41.55	27.06	60.01
ECE	Zero-shot	5.04	1.90	3.56	5.64	4.17	2.10	2.35	6.31	3.79	8.60	14.40	2.66
	TPT	11.27	11.34	4.10	3.78	3.70	13.66	5.18	15.57	9.20	25.29	21.00	11.20
	w/ **ZS-Norm**	10.57	10.81	4.29	3.71	3.62	13.29	4.73	15.28	8.50	23.95	17.61	10.49
	w/ **Penalty**	9.58	11.31	3.99	1.57	2.26	13.94	4.27	14.51	8.88	23.10	11.82	9.78
	w/ **SaLS**	**9.26**	9.81	4.45	2.90	2.50	12.01	3.91	15.23	8.64	21.09	12.31	9.05
	C-TPT	6.33	3.05	2.60	2.46	0.87	3.91	1.62	11.30	2.73	21.38	13.58	2.88
	w/ **ZS-Norm**	5.74	2.85	2.29	2.69	0.78	3.53	1.61	10.94	2.72	20.94	12.17	2.65
	w/ **Penalty**	**3.14**	5.93	2.26	2.66	0.81	3.79	1.64	11.58	2.74	20.49	10.83	2.51
	w/ **SaLS**	5.22	2.21	3.41	3.94	2.55	1.75	1.78	10.15	2.58	12.92	10.41	2.71

scenario of maximum entropy, where the predicted probabilities are semantically meaningless, leading to worse discrimination performance. This reasoning is empirically supported in Table 5.4, where we can see that, regardless of the learning paradigm, significantly decreasing the logit range yields higher ECE scores, i.e., miscalibration is magnified.

Table 5.4 **What if the logit range is further decreased?** ECE scores on ImageNet shifts (V2, S, A and R) for representative methods when reducing the original ZS logit range (denoted as **1**) to half (**1/2**) and one quarter (**1/4**) in **SaLS**

	CLIP-Ad			CoOp			TPT		
ZS-Range	**1**	**1/2**	**1/4**	**1**	**1/2**	**1/4**	**1**	**1/2**	**1/4**
RN50	8.95	21.31	31.51	7.82	24.44	37.72	16.74	25.15	40.7

5.5.2.4 Effect on logits

Following one of our main observations (5.2), we argued that the source of miscalibration in CLIP adaptation models is the increase of the logit range of their predictions, and not the logit norm. To empirically validate this hypothesis, we depict in 5.3 both the logit norm and logit ranges for a relevant method of each category, as well as the version improved with our **SaLS** solution, across the four OOD datasets of ImageNet. We can observe that, indeed, applying our

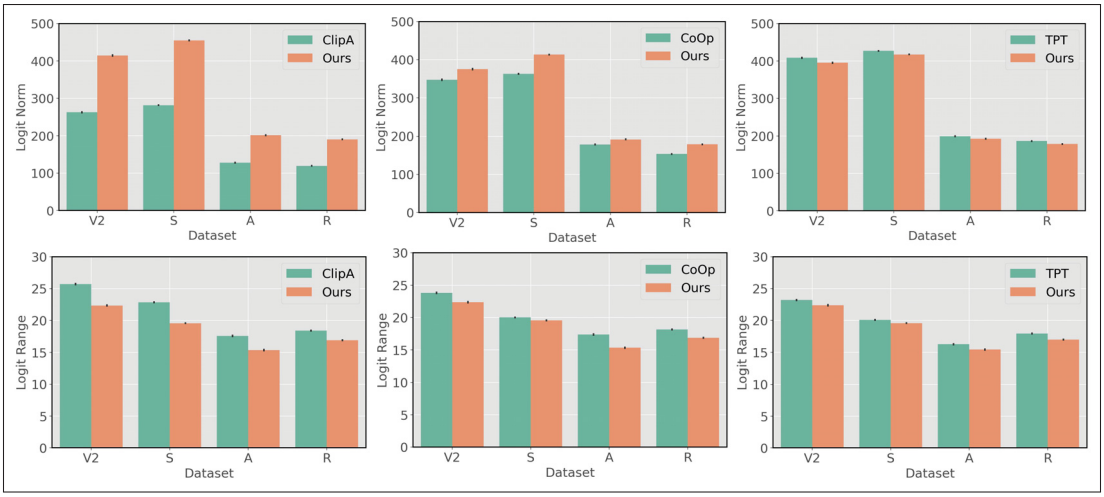

Figure 5.3 **Effect of calibrating adapted CLIP models.** Mean of the distribution of logit norms (*top*) and logit ranges (*bottom*) across the four ImageNet OOD datasets for a relevant Adapter-based (CLIP-Ad), Prompt Learning (CoOp) and TPT approach

approach (which improves calibration) leads to reduced logit ranges (*bottom*), whereas the logit norm (*top*) typically increases.

5.6 Conclusions

We have investigated the miscalibration issue of popular CLIP adaptation approaches on the challenging task of few-shot and zero-shot adaptation under distributional drifts. We have analyzed the source of this issue and demonstrated that, in contrast to existing evidence pointing to the logit norm, increases in the range of predicted logits might be a potential cause of miscalibration on the adapted models. To overcome this issue, we have presented three simple solutions, which consist in constraining the logit ranges to the values of the zero-shot predictions, either at training or test time. Extensive experiments on multiple models from the three categories, and popular OOD benchmarks, demonstrate that incorporating our simple solution to existing CLIP adaptation approaches considerably enhances their calibration performance, without sacrificing model accuracy. The proposed approach is model-agnostic, and demonstrate superior performance regardless of the family of approaches or setting, making of our model an

appealing yet simple solution for zero-shot and few-shot CLIP adaptation, particularly in the challenging scenario of out-of-distribution data.

CONCLUSION AND RECOMMENDATIONS

In this thesis, we predominantly focus on improving the calibration of state-of-the-art deep neural networks in least explored domains including medical image segmentation, and vision-language adapters. We begin by introducing model calibration specific to classification tasks, and provide background on the design choices causing miscalibration. From that, we mainly addressed the miscalibration caused by cross entropy through application specific regularizers. In the case of medical image segmentation, we started with a benchmark study to compare existing calibration methods with a focus on margin based logit smoothing, followed it with spatial-aware regularizer, and further improved it with class and region adaptive weights. For vision-language models, the logits of the out-of-distribution predictions are refined based on its zero-shot results.

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

proposed loss term should be preferred to train models that provide higher discriminative performance, while delivering accurate uncertainty estimates.

In the second contribution, we observe that 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 effectively. 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.

In the third contribution, we proposed a class and region-wise constraint approach to tackle the miscalibration issue in semantic segmentation models. In particular, we formulated a solution that considers the specificities of each category and different regions by introducing independent class and region-wise penalty weights. This contrasts with the second contribution, where a uniform scalar penalty weight is employed, regardless of categories or regions. Furthermore, we transferred the constrained problem to its dual unconstrained optimization counterpart by using an Augmented Lagrangian method (ALM). This alleviates the need for manually adjusting each penalty weight and allows, through a series of iterative *inner* and *outer* steps, to find the optimal value of each penalty weight, which can be learned in an adaptive manner. Comprehensive experiments on two popular segmentation benchmarks, and with two well-known segmentation backbones, demonstrate the superiority of our approach over a set of relevant recent calibration approaches.

In the fourth and final contribution, we have investigated the miscalibration issue of popular CLIP adaptation approaches on the challenging task of few-shot and zero-shot adaptation under distributional drifts. We have analyzed the source of this issue and demonstrated that, in contrast to existing evidence pointing to the logit norm, increases in the range of predicted logits might be a potential cause of miscalibration on the adapted models. To overcome this issue, we have presented three simple solutions, which consist in constraining the logit ranges to the values of the zero-shot predictions, either at training or test time. Extensive experiments on multiple models from the three categories, and popular OOD benchmarks, demonstrate that incorporating our simple solution to existing CLIP adaptation approaches considerably enhances their calibration performance, without sacrificing model accuracy. The proposed approach is model-agnostic, and demonstrates superior performance regardless of the family of approaches or setting, making our model an appealing yet simple solution for zero-shot and few-shot CLIP adaptation, particularly in the challenging scenario of out-of-distribution data.

To summarize, as a major part of this thesis we have provided regularizers to overcome the miscalibration problem prevalent with use of cross entropy in medical image segmentation. While the proposed solutions offered superior performance to existing approaches, there exist multiple avenues which are worth exploring. For example, one of the limitations of our approaches is that they disregard 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 formulations. Another important direction concerns the effect of class imbalance on calibration (Zhong, Cui, Liu & Jia, 2021b; Liu *et al.*, 2023a). In medical image segmentation, where long-tailed distributions are common (Ma *et al.*, 2021a), majority classes often yield overconfident predictions, while minority classes tend to be underconfident. This aspect has received little attention so far, and future calibration-based objective functions should be designed to explicitly address such frequency disparities in order to improve reliability. Given the scarcity of medical imaging datasets, training-time regularization may not be the most scalable approach; therefore, post-hoc calibration is often preferred (Hwang, Kim & Whang, 2025). While this thesis did not extensively study post-hoc methods for segmentation, the findings on spatial awareness are highly relevant to test-time calibration (Ding *et al.*, 2021; Wang *et al.*, 2023).

The other part of the thesis introduced the miscalibration in vision language foundation models like CLIP and proposed solutions to handle them both during training and inference time. Firstly, the work only evaluated the proposed normalization strategy with preliminary domain shifts and could be investigated with popular domain generalization benchmarks in classification (Koh *et al.*, 2021). Besides, recent works like KgCoOp (Hantao Yao, 2023) in prompt learning and

CLAP (Silva-Rodriguez *et al.*, 2024) in adapters have integrated zero-shot predictions during few shot training to enhance the generalization ability. Hence, it would be interesting to observe the logits from these lines of work to verify whether the range is lesser than the previous methods and how it compares to zero-shot predictions. Interestingly, in our experiments with test-time prompt tuning, we noticed that there were few challenging datasets Shu *et al.* (2022), for which the zero-shot performance was not satisfactory. Thus, it is not recommended to assume the zero-shot logits are always better calibrated, important to also consider the overall accuracy. Lastly, this work hasn't studied open vocabulary classification (Wu *et al.*, 2024), as reliable predictions for both the base and novel classes are expected for real-time deployment. Our initial experiments on understanding the model calibration of the same revealed that, base class predictions are underconfident, while novel classes are mostly overconfident. Similar observations along with solutions have been presented in Distance-Aware Calibration (DAC) (Wang *et al.*, 2024b), and Dynamic Outlier Regularization (DOR) (Wang *et al.*, 2024a). DAC showed that scaling the temperature value based on distance between the predicted text label and base classes, fine-tuned CLIP tends to give a more reliable confidence level for new classes. To further improve the base class calibration, DOR showed that utilizing relevant but non-overlapped outliers regularizes the textual distribution.

In our contributions, we have only focused on the calibration specific to medical image segmentation. But, there are other key tasks in the medical image analysis pipeline, which also require calibration (Litjens *et al.*, 2017). For instance, detecting tumors require reliable object detector. Lately, there have been few attempts to define and propose solutions for calibration in object detection (Munir, Khan, Sarfraz & Ali, 2022; Munir, Khan, Khan & Khan, 2023a; Munir, Khan, Khan, Ali & Shahbaz Khan, 2023b; Pathiraja, Gunawardhana & Khan, 2023). Cal-DETR (Munir *et al.*, 2023b) is the one which is close to the solutions proposed in this thesis. Like how we have constrained the logits based on the spatial information, Cal-DETR modulates the logits based on the uncertainty specifically obtained from the transformer based

architecture. As another example, survival risk prediction (Xu *et al.*, 2022b) is key in staging cancer from pathological images, and uncertainty intervals associated with the survival score is essential. This particular kind of problems which require ordinal regression (Wang *et al.*, 2025) could be approached by regression-by-classification approaches (Pintea, Lin, Dijkstra & van Gemert, 2023). As our approaches in this thesis, calibrates both the maximum confidence score, and the scores of remaining classes, it is highly desirable for categorizing the severity of the condition. Another possible direction would be to obtain reliable uncertainty estimates for tasks like image reconstruction (Zou *et al.*, 2023), as hallucinated structures are problematic for downstream tasks. Most of the works in this literature (Murugesan, Vijaya Raghavan, Sarveswaran, Ram & Sivaprakasam, 2019; Fischer, Thomas & Baumgartner, 2023; Zhang, Li & Chen, 2024) have attempted to improve the uncertainty estimate through architecture designs, but our findings on constraining the output space based on the neighboring information could be translated, restricting trivial irregularities. Furthermore, the metrics used for measuring calibration in segmentation and detection are a direct extension of classification (Lane, 2025), hence there is scope for developing metrics (Kuzucu, Oksuz, Sadeghi & Dokania, 2024) specific to the respective tasks.

Though the proposed models in thesis provide desired calibration, they do not provide formal uncertainty guarantee, which can lead to critical errors. Conformal Prediction (CP) framework (Vovk, Gammerman & Saunders, 1999; Sadinle, Lei & Wasserman, 2019) has been designed to provide uncertainty with desired guarantee under exchangeability assumption. The key idea behind conformal inference is to assess the "nonconformity" scores of the test data compared to the observed calibration data, and provide valid prediction sets. One of the popular conformal baselines is to prepare the set by including classes from highest to lowest probability until their sum just exceeds the threshold. However, this does not necessarily guarantee coverage, and lately, Adaptive Prediction Sets (APS) (Romano, Sesia & Candes, 2020), and Regularized Adaptive Prediction Sets (RAPS) (Angelopoulos, Bates, Jordan & Malik, 2021) have become

the standard approaches. Recently, attempts have been made to bring split conformal inference to the segmentation task. To begin with, direct extension of solutions proposed in classification literature have been adapted (Angelopoulos & Bates, 2021; Mossina, Dalmau & Andéol, 2024), which is not desirable as the uncertainty of a particular pixel should be decided by its neighbouring pixel. Second, based on classwise and clustered conformal inference (Ding, Angelopoulos, Bates, Jordan & Tibshirani, 2023), Kandinsky Conformal (Brunekreef, Marcus, Sheombarsing, Sonke & Teuwen, 2024) partially incorporated the spatial information through pixel grouping, and have threshold specific to each cluster. However, finding Kandinsky clusters depends deeply on the prior characteristics of the data and may not be able to handle all possible geometric priors. Recent advancements have extended conformal inference to object detectors (Andéol, Fel, De Grancey & Mossina, 2023), CLIP models (Silva-Rodríguez, Ben Ayed & Dolz, 2025; Morales-Álvarez, Christodoulidis, Vakalopoulou, Piantanida & Dolz, 2024; Fillioux *et al.*, 2024), language models (Quach *et al.*, 2024), and time series (Xu & Xie, 2023).

APPENDIX I

PROMPTING CLASSES: EXPLORING THE POWER OF PROMPT CLASS LEARNING IN WEAKLY SUPERVISED SEMANTIC SEGMENTATION

Balamurali Murugesan[a] , Rukhshanda Hussain[c] , Rajarshi Bhattacharya[c] , Ismail Ben Ayed[b] , Jose Dolz[a]

[a] Department of Software Engineering, École de Technologie Supérieure,
[b] Department of System Engineering, École de Technologie Supérieure,
1100 Notre-Dame West, Montreal, Quebec, Canada H3C 1K3
[c] Jadavpur University

Paper published in Winter Conference on Applications of Computer Vision, April 2024

Presentation

This chapter presents the article *"Prompting classes: Exploring the Power of Prompt Class Learning in Weakly Supervised Semantic Segmentation"* (Murugesan *et al.*, 2024b) published in Winter Conference on Applications of Computer Vision (**WACV**), Pages. 291–302, 2024.

Abstract

Recently, CLIP-based approaches have exhibited remarkable performance on generalization and few-shot learning tasks, fueled by the power of contrastive language-vision pre-training. In particular, prompt tuning has emerged as an effective strategy to adapt the pre-trained language-vision models to downstream tasks by employing task-related textual tokens. Motivated by this progress, in this work we question whether other fundamental problems, such as weakly supervised semantic segmentation (WSSS), can benefit from prompt tuning. Our findings reveal two interesting observations that shed light on the impact of prompt tuning on WSSS. First, modifying only the class token of the text prompt results in a greater impact on the Class Activation Map (CAM), compared to arguably more complex strategies that optimize the context. And second, the class token associated with the image ground truth does not necessarily correspond to the category that yields the best CAM. Motivated by these observations, we

introduce a novel approach based on a **P**r**O**mpt c**L**ass l**E**arning (**POLE**) strategy. Through extensive experiments we demonstrate that our simple, yet efficient approach achieves SOTA performance in a well-known WSSS benchmark. These results highlight not only the benefits of language-vision models in WSSS but also the potential of prompt learning for this problem. The code is available at https://github.com/Ruxie189/WSS_POLE.

3. Introduction

Image semantic segmentation is a fundamental problem in computer vision, as it serves as a precursor of many tasks, such as medical image analysis or autonomous driving. Fueled by the advances in deep learning, semantic segmentation has experienced a tremendous progress. Nevertheless, obtaining precise pixel-wise annotations is a labor-intensive and time-consuming task.

To alleviate the annotation burden, weakly supervised semantic segmentation (WSSS) has emerged as an appealing alternative, where labels typically come in the form of image tags Fan, Zhang, Song & Tan (2020a); Kolesnikov & Lampert (2016); Hou, Jiang, Wei & Cheng (2018); Lee, Kim, Lee, Lee & Yoon (2019), bounding boxes Song, Huang, Ouyang & Wang (2019); Khoreva, Benenson, Hosang, Hein & Schiele (2017), scribbles Lin, Dai, Jia, He & Sun (2016); Tang *et al.* (2018b) or global constraints Pathak, Krahenbuhl & Darrell (2015); Kervadec *et al.* (2019c), among others. In particular, image-level WSSS has received significant attention, as it offers a cost-effective alternative to pixel-level annotations (e.g., 20 seconds reported in Bearman, Russakovsky, Ferrari & Fei-Fei (2016)). Under this setting, WSSS commonly leverages class activation maps (CAMs) Zhou, Khosla, Lapedriza, Oliva & Torralba (2016) obtained from image classification networks to localize objects. Specifically, these maps are later used as pixel-wise pseudo-labels to train a segmentation model, mimicking full supervision. However, CAMs tend to highlight discriminative regions, while ignoring other useful cues, which results in suboptimal pseudo-labels that do not cover the whole extent of the target objects. Narrowing down the existing gap between classification and segmentation tasks is therefore crucial for the progress of WSSS models. To solve this issue, existing approaches intend to complete

generated CAMs by forcing the network to focus on more non-discriminative regions, which can be achieved by region mining strategies Kweon, Yoon, Kim, Park & Yoon (2021); Hou *et al.* (2018), or integrating iterative processes Ahn & Kwak (2018b). Despite employing complex CAM refinement strategies, sometimes involving multiple training steps, existing approaches still exhibit suboptimal performance in terms of both completeness of the target objects and segmentation accuracy.

This motivates the exploration of complementary learning strategies that can further improve the segmentation performance of these models. Vision-language pre-training (VLP) models, such as the recently introduced Contrastive Language-Image Pre-training (CLIP) Radford *et al.* (2021) strategy, have the potential to bring WSSS approaches to the next level, as it can associate much wider visual concepts in an image with their corresponding text labels in an open-world scenario. This contrasts with standard WSSS settings, where the fixed set of predetermined object categories limits the quality of generated CAMs due to unnecessary background activations from class-related background pixels. For example, CLIMS Xie, Hou, Ye & Shen (2022) exposed these issues showing that background pixels related to the class *'railroad'* contributed to the prediction of the CAM associated to the category *'train'*, leading to over-segmented CAMs.

With the rise of these powerful vision-language pre-training models, recent evidence has highlighted the importance of their text input, typically referred to as *prompt*, in adapting these models to downstream tasks and datasets. For instance, Zhou *et al.* Zhou, Yang, Loy & Liu (2022d) empirically demonstrated that the use of *'a* [CLS]' or *'a photo of a* [CLS]' as a prompt led to substantial differences in the classification performance of the model. Following these findings, recent literature has focused on tuning the context of these prompts, typically as continuous learnable vectors Zhou *et al.* (2022d); Zhou, Yang, Loy & Liu (2022b); Ju, Han, Zheng, Zhang & Xie (2022). Despite its potential importance, the impact of modifying the [CLS] token has been largely overlooked in the context of prompt learning. Additionally, while prompt learning has shown promising results in fine-tuning and classification tasks such as zero-shot image recognition, its effectiveness on other visual recognition problems is not well-understood.

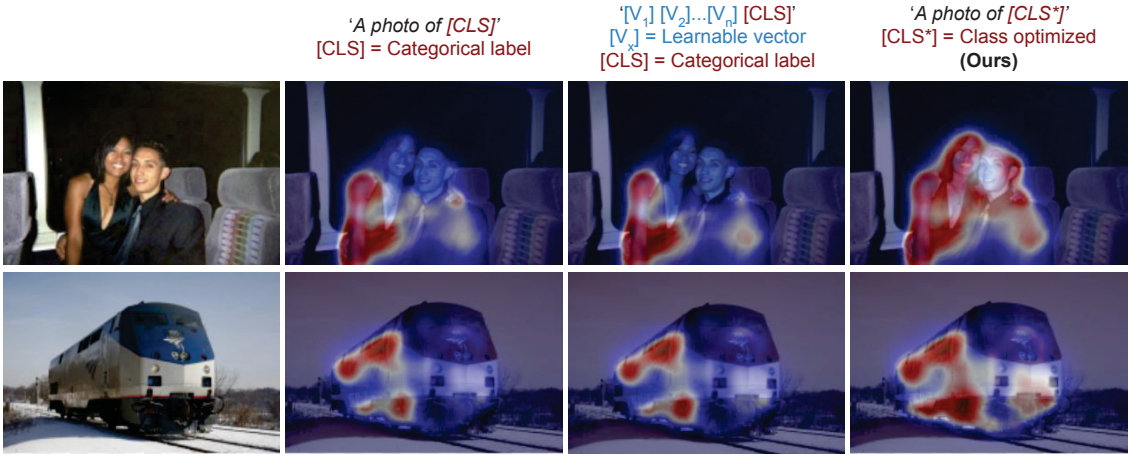

Figure-A I-1 **Impact of the input text prompt on the generation of class activation maps (CAMs).** Employing the ground truth categorical label as [CLS] token (*second column*) does not necessarily result in the best initial CAMs. Furthermore, even though complex techniques to optimize the [CTX] tokens, such as CoOp Zhou *et al.* (2022b) (*third column*) may improve the CAMs, we have observed that simply modifying the ground truth class in the [CLS] token by a higher correlated synonym leads to improvements in the identified class-related regions (*fourth column*).

Based on these observations, we explore in this work how vision-language pre-training can be further leveraged to improve the performance of WSSS models. In particular, we want to address the following questions: ① *Is prompt learning useful in weakly supervised segmentation?*, ② *Which parts of the prompt have a greater impact on the generated CAMs?* ③ *Can we devise a simple yet effective alternative to improve the segmentation performance under the weakly-supervised learning paradigm?*

Our contributions can be summarized as follows:

- We provide empirical evidence that modifying the input prompt in VLP models has a direct impact on the generated CAMs in a weakly supervised segmentation scenario, which in turn affects the performance of the segmentation network.
- More interestingly, our findings reveal that replacing the [CLS] token in the input prompt has a greater impact on the performance than modifying the prompt context, which contrasts with recent observations in classification problems (See Fig. **??**). Furthermore, the [CLS]

token associated with the actual image ground truth does not necessarily correspond to the category that yields the best CAM, and the performance varies considerably across closely related categories. These insights shed light on the importance of careful prompt design in optimizing the performance of segmentation models trained under the weakly-supervised paradigm.

- Based on these observations, we propose a simple yet efficient strategy to leverage language driven models in the challenging task of WSSS. The resulting model, based on a **Pr**ompt c**L**ass l**E**arning (**POLE**) approach, learns the category name that produces the highest correlation between the image and a corresponding text prompt, and uses it to further leverage the segmentation performance.
- Following the literature, we conduct extensive experiments on PASCAL VOC 2012 to well demonstrate the superiority of our method over other state-of-the-art methods for WSSS.

4. Related Work

Weakly supervised semantic segmentation. Due to its low annotation cost, WSSS based on image-level labels has gained increasing popularity. These methods rely on class activation maps (CAMs) to identify target object regions by discovering informative pixels for the classification task. As discovered regions are typically highly discriminative and fail to cover the whole context of the target objects, recent literature focuses on generating high-quality CAMs by refining initial estimations from simple models. A common strategy is to mine or erase regions at either image Wei *et al.* (2017); Zhang, Gu, Zhang & Dai (2021b); Kweon *et al.* (2021) or features level Hou *et al.* (2018); Lee *et al.* (2019), and can be seen as a way of preventing a classifier from focusing exclusively in highly discriminative areas. Other works have instead exploited sub-categories dependencies Chang *et al.* (2020), cross-image semantics Fan, Zhang, Tan, Song & Xiao (2020b); Sun, Wang, Dai & Van Gool (2020), attention mechanisms Wu *et al.* (2021), equivariant constraints Wang, Zhang, Kan, Shan & Chen (2020b); Patel & Dolz (2022) and pairwise semantic affinities Ahn & Kwak (2018b); Wang, Liu, Ma & Yang (2020a). Furthermore, additional supervision, such as saliency maps can be also integrated to provide additional hints about the location of the target object Lee, Lee, Lee & Shim (2021c); Jiang,

Yang, Hou & Wei (2022). More recently, visual transformers (ViT) Dosovitskiy *et al.* (2020) have been also leveraged to improve original CAMs Xu, Ouyang, Bennamoun, Boussaid & Xu (2022a); Li *et al.* (2022b), demonstrating superior performance than their CNNs counterparts.

Contrastive Language-Image Pre-training (CLIP) based semantic segmentation. Very recently, large-scale VLP models, such as CLIP Radford *et al.* (2021), have demonstrated to improve significantly the performance of vision models on classic recognition tasks, such as zero-shot object detection Gu, Lin, Kuo & Cui (2021), few-shot learning Hu, Li, Stühmer, Kim & Hospedales (2022c) and zero-shot semantic segmentation Li, Weinberger, Belongie, Koltun & Ranftl (2021); Zhou, Lei, Zhang, Liu & Liu (2023b). Closely related to our work, CLIMS Xie *et al.* (2022) integrates CLIP in the context of weakly-supervised segmentation, which enhances the initial CAMs by highlighting more comprehensive object regions, while suppressing closely-related background areas. Inspired by the improvement observed in the robustness and generability of visual recognition models driven by language assistance, our work delves deeper into understudied factors, particularly in the weakly-supervised scenario. We stress that our work is different from Xie *et al.* (2022). In particular, CLIMS Xie *et al.* (2022) proposes to leverage standard CLIP in WSSS, whereas we further explore the effect of the given prompt on this task. As we will show in our empirical validation, properly designing the input prompt results in significant improvements over the standard text prompts. More surprisingly, *using the class ground truth as categorical name in the input text prompt does not necessarily yields the best segmentation results.*

Prompt learning in visual recognition problems is a rapidly growing research direction, whose popularity stems from the promising results observed in NLP tasks Lester, Al-Rfou & Constant (2021); Li & Liang (2021); Liu *et al.* (2023c); Raffel *et al.* (2020). For example, recent works in prompt learning Du *et al.* (2022b); Rao *et al.* (2022); Zhou *et al.* (2022b,d); Wang *et al.* (2022b); Nayak, Yu & Bach (2023) have achieved promising results on several vision-language tasks, notably in classification. In addition to tackle a different task, i.e., classification *vs.* weakly supervised segmentation, the main differences with our work is that these approaches mostly study prompt learning from a context perspective. In particular, existing literature considers the

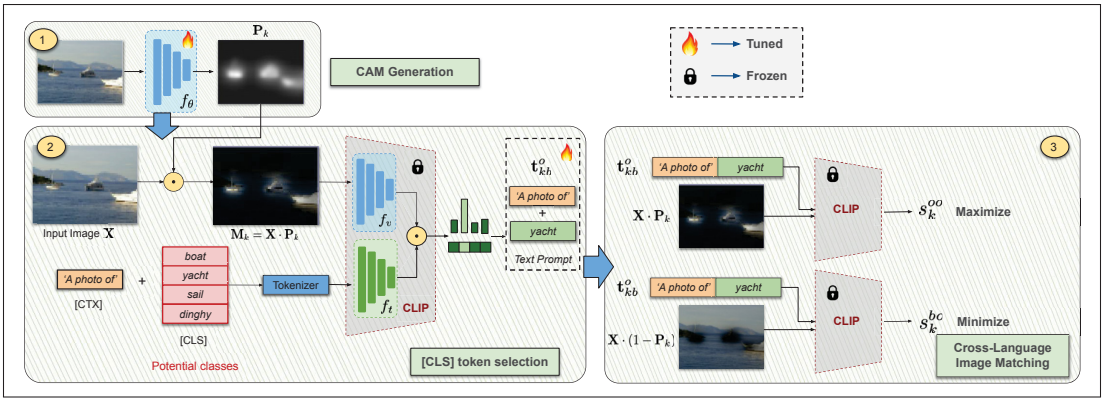

Figure-A I-2 **Proposed Weakly Supervised Segmentation approach. 1)** Class activation maps are generated for an input image $\mathbf{X}$. **2)** CLIP pre-trained visual and text encoders (f_θ and f_θ) are leveraged to find the category name [CLS] presenting the highest correlation with the image $\mathbf{M}_k$, the result of multiplying the input image $\mathbf{X}$ and its corresponding CAM $\mathbf{P}_k$. **3)** With the [CLS] token selected, we generate the input text prompt $\mathbf{t}^o_{kb}$ to the Cross-Language Image Matching (CLIMS) learning framework

class token [CLS] as a fixed word embedding, while optimizing the context Du *et al.* (2022b); Rao *et al.* (2022); Zhou *et al.* (2022b,d); Wang *et al.* (2022b) or attributes Nayak *et al.* (2023). In most cases, the context tokens are represented by learnable continuous vectors, which yield to text embeddings lacking semantic knowledge Zhou *et al.* (2022d). In contrast, our approach performs a selection on a finite set of potential synonyms, which facilitates both the search and the interpretation of the selected token.

5. Methodology

5.1 Problem setting.

Let us denote $\mathcal{D} = \{(\boldsymbol{X}_i, \mathbf{y}_i)\}_{i=1}^N$ as a weakly labeled dataset, where $\boldsymbol{X}_i \in \mathbb{R}^{\Omega_i}$ is an input image, Ω_i denotes its spatial dimensionality, $\mathbf{y}_i \in \{0,1\}^K$ its associated one-hot encoded image label[1], and K indicates the number of categories. Thus, the goal of weakly supervised semantic

[1] In PascalVOC, multiple classes can be present in the same image, where $\mathbf{y}$ becomes a multi-class one-hot encoded vector.

segmentation is to provide pixel-wise predictions from an input image $\mathbf{X}_i$ given its corresponding image-level label $\mathbf{y}_i$.

5.2 Our framework.

Preliminaries: Class Activation Maps. We first revisit the generation of class activation maps (CAM) from the image-level labels, a popular strategy in WSSS. Let us first define a feature extractor $f_\theta(\cdot)$, which can be represented by a deep neural network parameterized by θ. Thus, for a given image $\mathbf{X}$, the feature extractor provides a representation $\mathbf{Z} \in \mathbb{R}^{C\times H'\times W'}$, where C is the number of channels and H' and W' represent the dimensionality of the feature map. To provide CAMs, a global average pooling (GAP) layer, followed by a 1×1 convolutional layer $\boldsymbol{W} \in \mathbb{R}^{C\times K}$ is applied to the learned features $\mathbf{Z}$ from an image $\mathbf{X}$. Then, the resulting logits are mapped into probabilities $\hat{\mathbf{y}} \in [0, 1]^K$ by applying a sigmoid function. To train the neural network, we follow the literature Xu *et al.* (2021b); Xu *et al.* (2022a); Xie *et al.* (2022) and use the multi-label soft-margin loss as the classification function:

$$\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) = -\frac{1}{K}\sum_{k=1}^{K}(y^k \log \hat{y}^k + (1 - y_k) \cdot \log (1 - \hat{y}_k)). \tag{A I-1}$$

Once the backbone network is trained, the initial CAMs can be obtained as follows:

$$\boldsymbol{P}_k(h, w) = \boldsymbol{W}_k^\top \boldsymbol{Z}(h, w), \tag{A I-2}$$

where $\boldsymbol{P}_k$ is the activation map for a given category k.

Learning objectives. The framework used for this work shares the same overall structure as the recent CLIMS Xie *et al.* (2022), as it represents the first weakly-supervised segmentation approach integrating CLIP text embeddings. Nevertheless, we stress that in our work we study how we can leverage prompt learning to further improve the performance of weakly supervised segmentation models. In particular, the standard GAP layer employed to generate CAMs is

replaced by a sigmoid function $\sigma(\cdot)$, resulting in $\boldsymbol{P}_k(h,w) = \sigma(\boldsymbol{W}_k^\top \boldsymbol{Z}(h,w))$. The pretrained CLIP model Radford *et al.* (2021) uses a visual and a text encoder which we denote as: f_v and f_t, respectively. Instead of passing the raw input image $\mathbf{X}$ to the CLIP image encoder, it is multiplied by $\boldsymbol{P}_k$, with the goal of focusing only on the class highlighted by its corresponding CAM. Moreover, to avoid interferences from related background regions, the input image is also multiplied by $(1 - \boldsymbol{P}_k)$. Hence, we can create two different visual embeddings: the embedding of the target category ($\boldsymbol{v}_k^{io}$), and its background ($\boldsymbol{v}_k^{ib}$), which are formally given by:

$$\boldsymbol{v}_k^{io} = f_v(\boldsymbol{X} \cdot \boldsymbol{P}_k), \quad \boldsymbol{v}_k^{ib} = f_v(\boldsymbol{X} \cdot (1 - \boldsymbol{P}_k)). \tag{A I-3}$$

Now, for all the potential object classes k and their corresponding text inputs $\boldsymbol{t}_k^o$, we obtain their text embeddings, which are referred to as $\boldsymbol{v}_k^{to}$:

$$\boldsymbol{v}_k^{to} = f_t(\boldsymbol{t}_k^o). \tag{A I-4}$$

Note that to generate these embeddings, we just need to provide the different text inputs to the trained CLIP text encoder $f_t(\cdot)$. Following the reasoning behind the training of CLIP, the foreground image embedding $\boldsymbol{v}_k^{io}$ should be highly correlated to the text embedding $\boldsymbol{v}_k^{to}$ of that particular class. In contrast, the background image embedding $\boldsymbol{v}_k^{ib}$ should have a much lower correlation with the object classes. This can be modeled by using the following objective function:

$$\mathcal{L}_{Cont} = -\alpha \sum_{k=1}^{K} y_k \cdot \log(s_k^{oo}) - \beta \sum_{k=1}^{K} y_k \cdot \log(1 - s_k^{bo}), \tag{A I-5}$$

where $s_k^{oo} = \text{sim}(\boldsymbol{v}_k^{io}, \boldsymbol{v}_k^{to})$ and $s_k^{bo} = \text{sim}(\boldsymbol{v}_k^{ib}, \boldsymbol{v}_k^{to})$ represent the *object-to-object* and *background-to-object* similarities between visual and text embeddings, computed as a cosine similarity. Both terms in Eq. A I-5 act together to ensure that the activation map $\boldsymbol{P}_k$ covers the maximum possible region of the target, while excluding related background.

5.3 Category and image-driven prompt generation

Finding potential category related embeddings. To generate the object text representation $\boldsymbol{v}_k^{to}$ for a class k, the standard input prompt given to the text encoder has the following format: a context token [CTX] followed by the class name token [CLS] and ended by a punctuator ('.'). While the literature in prompt learning for large-scale visual language pre-trained models focuses on learning the context [CTX] Zhou *et al.* (2022b), CLIMS Xie *et al.* (2022) uses a fixed prompt, where the [CLS] token corresponds to the categorical label of the image. Contrary to these works, we hypothesize that modifying the input text prompts by optimizing only the [CLS] token has a greater impact on the generated CAMs. Indeed, as we will show empirically in the results section, using the ground truth class as a [CLS] token does not necessarily always results in the best segmentation performance.

Let us suppose that we take an input image $\mathbf{X}$ with its corresponding image class label $\mathbf{y}$, which indicates the k categories present on the image. For each category k in $\mathbf{y}$, we obtain a set of similar words, in terms of closeness in the semantic space, using chatGPT cha. More concretely, we provide the following query as input to chatGPT *"Give me m semantically similar words for [CLS] and also print the cosine similarity scores based on CLIP model"*, where [CLS] is a class name. This returns a list of m words along with their similarity scores for that particular [CLS]. This means that for each class [CLS], we can derive a set of m closest words, denoted as $\mathcal{S}$ =[CLS_1,CLS_2,...,CLS_m]. With this set of related categories, we can create a set of $m+1$ potential text prompts $\mathcal{T} = [\boldsymbol{t}_{k0}^o, \boldsymbol{t}_{k1}^o, \boldsymbol{t}_{k2}^o, \dots \boldsymbol{t}_{km}^o]$, where $\boldsymbol{t}_{k0}^o$ is the text prompt containing the categorical ground truth label for class k, i.e., [CLS] followed by a fixed [CTX] token, and $\boldsymbol{t}_{k1}^o$, ..., $\boldsymbol{t}_{km}^o$ are composed of the fixed [CTX] followed by a variable [CLS] token chosen from set $\mathcal{S}$. Now, we can extract an embedding for each of the prompts in $\mathcal{T}$ from the CLIP text encoder, resulting in $\mathcal{V} = [\boldsymbol{v}_{k0}^{to}, \boldsymbol{v}_{k1}^{to}, \boldsymbol{v}_{k2}^{to}, \dots \boldsymbol{v}_{km}^o]$.

How to select the best [CLS]? Given the input image $\mathbf{X}$ and its generated class activation map $\boldsymbol{P}_k$ for class k, we can obtain an image focusing on discriminative regions for that category by simply doing $\boldsymbol{M}_k = \boldsymbol{X} \cdot \boldsymbol{P}_k$ The resulting image, $\boldsymbol{M}_k$, is given to the CLIP image encoder to get a

compressed representation of $\mathbf{X}$, i.e., $\boldsymbol{v}_k^{io}$. similar to the steps described in Section 5.2. In contrast, we now obtain a correlation between each of the closest words in set $\mathcal{S}$ for the class k and the CAM activated image $\boldsymbol{M}_k$. In particular, this correlation is found by computing a similarity score between the visual and text embeddings: $\boldsymbol{v}_k^{io}$ and every $\boldsymbol{v}_{kj}^{to}$ for $j \in \{0, 1, 2, 3, \ldots, m\}$, given as:

$$\text{sim}(\boldsymbol{v}_k^{io}, \boldsymbol{v}_{kj}^{to}) = \frac{\boldsymbol{v}_k^{io} \cdot \boldsymbol{v}_{kj}^{to}}{|\boldsymbol{v}_k^{io}||\boldsymbol{v}_{kj}^{to}|}, \tag{A I-6}$$

which generates a vector containing the similarities between the visual encoding and each of the text encodings $[s_{k0}, s_{k1}, \ldots, s_{km}]$. From this similarity vector, we select the most correlated [CLS] token, which corresponds to the text embedding $\boldsymbol{v}_{kj}^{to}$ with the highest similarity with $\boldsymbol{v}_k^{io}$, computed with the argmax operator.

5.4 Weakly supervised adaptors

Following the success of adaptors in pre-trained language-vision models for classification tasks Rao *et al.* (2022); Zhang *et al.* (2022c); Gao *et al.* (2024), we propose to further improve our segmentation network by integrating image and text adaptors. In particular, and similar to Gao *et al.* (2024) in classification, we introduce two MLP layers $A_v(\cdot)$ and $A_t(\cdot)$ to transform the embeddings in the image side and text space, respectively, which is formulated as:

$$\boldsymbol{v}_k^{io*} = \boldsymbol{r}_v \cdot A_v(\boldsymbol{v}_k^{io}) + (1 - \boldsymbol{r}_v) \cdot \boldsymbol{v}_k^{io} \tag{A I-7}$$

$$\boldsymbol{v}_k^{to*} = \boldsymbol{r}_t \cdot A_t(\boldsymbol{v}_k^{to}) + (1 - \boldsymbol{r}_t) \cdot \boldsymbol{v}_k^{to}, \tag{A I-8}$$

with $A_v(\boldsymbol{v}_k^{io}) = ReLU(\boldsymbol{v}_k^{io}\mathbf{W}_1^v)\mathbf{W}_2^v$ and $A_t(\boldsymbol{v}_k^{to}) = ReLU(\boldsymbol{v}_k^{to}\mathbf{W}_1^t)\mathbf{W}_2^t$, where $\mathbf{W}$ represents the learnable parameters of the MLP layers. Furthermore, the parameters $\boldsymbol{r}_v$ and $\boldsymbol{r}_t$ are learnable vectors of the same shape as the original embeddings, and are used to selectively suppress or amplify the refinement of each feature through the MLP layers. Note that this contrasts

with standard adapters, i.e., Rao *et al.* (2022); Zhang *et al.* (2022c); Gao *et al.* (2024), whose balancing weight is a fixed hyperparameter. Our hypothesis is that using a fixed scalar to control the importance of each embedding is suboptimal, as the features refinement process may differ across images as well as depend on the class variation of the dataset.

6. Experiments

6.1 Experimental Settings

Dataset and evaluation protocol. Following CLIMS Xie *et al.* (2022), as well as other recent WSSS works Lee, Kim & Yoon (2021b), we conduct experiments on the popular PASCAL VOC 2012 Everingham, Van Gool, Williams, Winn & Zisserman (2009) benchmark. This dataset contains images with 20 foreground classes, which are split into 1,464 for training, 1,449 for validation and 1,456 for testing. The training set is augmented with 10,582 images and their associated image-level annotations from SBD Hariharan, Arbeláez, Bourdev, Maji & Malik (2011). To evaluate the performance of the proposed method, we resort to the mean intersection over union (mIoU). Last, while the results reported in the ablation studies are obtained on the training set, the results for the validation and testing sets of PASCAL VOC are obtained from the official evaluation server.

Implementation Details. We followed the default settings of CLIMS Xie *et al.* (2022) for training. In particular, input images are randomly rescaled and then augmented by random cropping to 512×512, as well as by horizontal flipping. We use SGD as the default optimizer, with a cosine annealing policy for scheduling the learning rate, and a batch size of 16 images. The model is trained for 10 epochs, with an initial learning rate of 0.00025 and a weight decay of 0.0001. We follow Ahn & Kwak (2018a) to adopt ResNet-50 He, Zhang, Ren & Sun (2016a) as backbone network for the generation of initial CAMs. All models are implemented in PyTorch and trained on NVIDIA A100 GPU with 40 GB memory. Furthermore, as the initial CAMs coarsely covers the target object, we further perform a refinement step with IRNet Ahn, Cho & Kwak (2019), to improve their quality before using them as pseudo ground-truth masks, a common practice in WSSS.

6.2 Results

How effective is prompt learning for weakly supervised segmentation? In this section we assess whether modifying the input text prompt leads to performance differences, which corresponds to our first question. To do this, we first manually selected two popular prompts: "*A photo of* [CLS]." , and "*An image of* [CLS]." , which are employed over all the images of the whole dataset. Note that the original CLIMS Xie *et al.* (2022) used the first prompt, and it is therefore considered as the baseline model. Furthermore, inspired by the recent advances on prompt learning, we also evaluate the impact different strategies, which model the different tokens with learnable continuous vectors: CoOp Zhou *et al.* (2022d) ([CTX] tokens), *target optimization* baseline in Wang *et al.* (2022b) ([CLS] tokens) and a modified version of DeFo Wang *et al.* (2022b). These results, which are reported in Table I-1, reveal that the text input, i.e., prompt, of the pre-trained vision-language model plays an important role on the segmentation performance. Indeed, we can observe that depending on the prompt employed, performance differences may diverge up to 3%, particularly on the final generated CAM (*last column*).

Table-A I-1 **Does prompt learning improve the performance of weakly supervised segmentation?** Comparison of the quality of initial CAMs and refined pseudo ground-truth masks obtained by different prompt learning strategies (with R50 as backbone), where either [CTX] or [CLS] tokens are modified. Evaluation is reported on the *train* set of PASCAL VOC2012, and refinement of the pseudo-masks is performed using **RW** (IRN Ahn *et al.* (2019)). [CTX] and [CLS] are used to indicate which part of the prompt is optimized in each approach. Furthermore, in the approaches optimizing the [CLS] token, 'V' indicates a continuous learnable vector, whereas CLS* represents the class selected among a set of potential classes

Method	[CTX]	[CLS]	Prompt	CAMs	+**RW**
Manual Xie *et al.* (2022)	✓	✗	"*A photo of* [CLS]."	56.6	70.5
Manual	✓	✗	"*An image of* [CLS]."	56.5	71.0
CoOp Zhou *et al.* (2022b)	✓	✗	$[V]_1[V]_2 \dots [V]_N$[CLS].	57.6	73.1
DeFo Wang *et al.* (2022b)	✓	✓	$[V]_1[V]_2 \dots [V]_N[V_{CLS}]$	56.6	73.2
Target optimization	✗	✓	"*A photo of* $[V_{CLS}]$."	56.8	73.1
Ours	✗	✓	"*A photo of* [CLS*]."	**58.7**	**73.6**

Table-A I-2 Comparison of the quality of initial CAMs and refined pseudo ground-truth masks using **RW** (PSA Ahn & Kwak (2018a)) on PASCAL VOC2012. The mIoU values here are reported on the *train* set. Backbone denotes the backbone network to generate CAMs. Best approaches (CAM and refined CAM) highlighted in bold

Method	Backbone	CAMs	+**RW**
PSA $_{\text{CVPR'2018}}$ Ahn & Kwak (2018b)	WR38	48.0	61.0
SC-CAM $_{\text{CVPR'2020}}$ Chang *et al.* (2020)	WR38	50.9	63.4
SEAM $_{\text{CVPR'2020}}$ Wang *et al.* (2020b)	WR38	55.4	63.6
AdvCAM $_{\text{CVPR'2021}}$ Lee *et al.* (2021b)	R50	55.6	68.0
MCTformer $_{\text{CVPR'2022}}$ Xu *et al.* (2022a)	ViT	**61.7**	69.1
SIPE $_{\text{CVPR'2022}}$ Chen, Yang, Lai & Xie (2022a)	R50	58.6	64.7
RECAM $_{\text{CVPR'2022}}$ Chen *et al.* (2022b)	R50	54.8	70.5
AdvCAM+W-OoD $_{\text{CVPR'2022}}$ Lee *et al.* (2022a)	R50	59.1	72.1
AFA $_{\text{CVPR'2022}}$ Ru, Zhan, Yu & Du (2022b)	MiT		68.7
CLIMS $_{\text{CVPR'2022}}$ Xie *et al.* (2022)	R50	56.6	70.5
VWL $_{\text{IJCV'2022}}$ Ru, Du, Zhan & Wu (2022a)	R101	57.3	71.4
AEFT $_{\text{ECCV'2022}}$ Yoon, Kweon, Cho, Kim & Yoon (2022)	WR38	56.0	71.0
ViT-PCM $_{\text{ECCV'2022}}$ Rossetti, Zappia, Sanzari, Schaerf & Pirri (2022)	ViT-B/16	–	71.4
ESOL $_{\text{NeurIPS'2022}}$ Li, Jie, Wang, Wei & Ma (2022a)	R50	53.6	68.7
POLE (Ours)	R50	59.0	**74.2**

Context *vs.* category in prompt learning. Once we have observed empirically that modifying input prompts results in performance differences, one question that naturally arises is which component of the prompt must be changed. Table I-1 shows that replacing a standard [CTX] token (i.e., *'A photo of'*) by a similar sequence (i.e., *'An image of'*) brings 0.5% difference. Nevertheless, if the [CTX] token is optimized for the whole dataset, e.g., CoOp Zhou *et al.* (2022b), these differences are further increased, with similar results if [CLS] token is optimized as a continuous vector (e.g., DeFo Wang *et al.* (2022b) and *Target Optimization*). Last, we can observe that only optimizing the [CLS] prompt, based on a set of pre-defined closely-related categories, actually provides the best performance across all the methods. These findings align with recent observations in Natural Language Inference Logan IV *et al.* (2022), which suggest that hand-crafted prompts conveying meaningful instructions outperform automatically optimized prompts.

Comparison to state-of-the-art. Previous experiments empirically demonstrated that modifying the input prompt can significantly improve the performance of weakly supervised segmentation

models. These observations motivated the proposed approach, which we benchmark against state-of-the-art models to show its superiority. Note that in what follows, the proposed approach, POLE, is composed of the [CLS] token selection process and the adaptor with learnable weights (eq. A I-7). First, Table I-2 reports the results of state-of-the-art methods in the generation of pseudo-masks. We can observe that, even compared to very recent models, the proposed approach brings substantial improvements, ranging from 2 to 6%. We interpret that the performance gain observed comes from the highest correlation between the selected category name and the content of the CAM-activated region in the image., which may capture larger discriminant areas of semantic objects. More interestingly, the proposed approach also outperforms recent methods that use additional information, for example in the form of extra saliency maps, which are typically trained on a supervised foreground-background detection dataset. Thus, based on these results we can argue that our approach represents an effective alternative to generate initial CAMs.

While the quality of the initial CAMs was evaluated in the previous section, we now assess their impact on the semantic segmentation task. In particular, Table I-3 reports the segmentation performance of the proposed approach compare to state-of-the-art methods in the validation and testing sets of PASCAL VOC2012 dataset. Compared to approaches that resort to the same supervision, our method provides very satisfactory results, ranking second if all the approaches are considered. Nevertheless, it has been found recently that vision transformers provide much better quality CAMs than conventional convolutional neural networks. Thus, if we just consider the approaches that leverage CNNs for the CAM generation, our approach achieves the best performance, with improvement gains ranging from 0.7% to 3% compared to very recent methods (e.g., RECAM, SIPE or CLIMS Xie *et al.* (2022)). Furthermore, even benchmarking POLE against recent approaches that use additional supervision, e.g., saliency maps, it yields very competitive performance.

How many synonyms are sufficient? And from which Corpus? Previous results have demonstrated empirically that the proposed approach brings substantial improvements by just replacing the categorical name on the ground truth by a closely related synonym. Thus, we

Table-A I-3 Evaluation results on PASCAL VOC2012 *val* and *test* sets. The best results are in **bold**. Sup. denotes the weak supervision type. $\mathcal{F}$ denotes full supervision. $\mathcal{S}$ denotes saliency map supervision. $\mathcal{I}$ denotes image-level supervision. Seg. denotes the segmentation network. Bac. denotes the backbone network for CAMs generation. V1: DeepLabV1. V2: DeepLabV2. V16: VGG-16 Simonyan & Zisserman (2014). R50: ResNet-50 He *et al.* (2016a). WR38: WideResNet38 Wu *et al.* (2019b). Segmentation network pretrained with ImageNet otherwise using MS COCO dataset (‡). Approaches based on visual transformers (*last section*) and convolutional neural networks (*before-last section*) are separated, and best method in each is highlighted in **bold**

Sup.	Method	Seg.	Bac.	*val*	*test*
$\mathcal{F}$	DeepLabV2 $_{\text{TPAMI'18}}$ Chen, Papandreou, Kokkinos, Murphy & Yuille (2018)	-	-	77.6	79.7
	WideResNet38 $_{\text{PR'19}}$ Wu, Shen & van den Hengel (2019a)	-	-	80.8	82.5
$\mathcal{I}+\mathcal{S}$	NSROM $_{\text{CVPR'21}}$ Yao *et al.* (2021b)	V2‡-R101	V16	68.3	68.5
	DRS $_{\text{AAAI'21}}$ Kim, Han & Kim (2021)	V2‡	V16	70.4	70.7
	EPS $_{\text{CVPR'21}}$ Lee *et al.* (2021c)	V2‡-R101	WR38	70.9	70.8
	EDAM $_{\text{CVPR'21}}$ Wu *et al.* (2021)	V2‡-R101	WR38	70.9	70.6
	AuxSegNet $_{\text{ICCV'21}}$ Xu *et al.* (2021a)	WR38	-	69.0	68.6
	SANCE $_{\text{CVPR'22}}$ Xu *et al.* (2021a)	V2-R101	R50	72.0	72.9
$\mathcal{I}$	SEAM $_{\text{CVPR'20}}$ Wang *et al.* (2020b)	V3-R38	WR38	64.5	65.7
	BES $_{\text{ECCV'20}}$ Chen, Wu, Fu, Han & Zhang (2020a)	V2-R101	R50	65.7	66.6
	SC-CAM $_{\text{CVPR'20}}$ Chang *et al.* (2020)	V2-R101	WR38	66.1	65.9
	A^2GNN $_{\text{TPAMI'21}}$ Zhang, Xiao, Jiao, Wei & Zhao (2021a)	V2-R101	WR38	66.8	67.4
	VWE $_{\text{IJCAI'21}}$ Ru, Du & Wu (2021)	V2-R101	R50	67.2	67.3
	AdvCAM$_{\text{CVPR'21}}$ Lee *et al.* (2021b)	V2-R101	R50	68.1	68.0
	VWL$_{\text{IJCV'22}}$ Ru *et al.* (2022a)	V2‡-R101	R101	70.6	70.7
	SIPE $_{\text{CVPR'22}}$ Chen *et al.* (2022a)	V2-R38	R50	68.2	69.5
	CLIMS $_{\text{CVPR'22}}$ Xie *et al.* (2022)	V2‡-R101	R50	70.4	70.0
	AdvCAM+W-OoD$_{\text{CVPR'22}}$ Lee *et al.* (2022a)	V2-R101	R50	69.8	69.9
	SIPE$_{\text{CVPR'22}}$ Chen *et al.* (2022a)	V2-R101	R50	68.8	69.7
	RECAM$_{\text{CVPR'22}}$ Chen *et al.* (2022b)	V2-R101	R50	68.5	68.4
	AMN$_{\text{CVPR'22}}$ Lee, Kim & Shim (2022b)	V2‡R101	R50	70.7	70.6
	Spatial-BCE$_{\text{ECCV'22}}$ Wu *et al.* (2022)	V2-R101	R38	68.5	69.7
	ESOL$_{\text{NeurIPS'22}}$ Li *et al.* (2022a)	V2-R101	R50	69.9	69.3
	POLE (ours)	V2‡-R101	R50	**71.5**	**71.4**
$\mathcal{I}$	AFA$_{\text{CVPR'22}}$ Ru *et al.* (2022b)	MiT-B1		66.0	66.3
	MCTformer$_{\text{CVPR'22}}$ Xu *et al.* (2022a)	V1-R38	DeiT-S	**71.9**	**71.6**
	ViT-PCM$_{\text{ECCV'22}}$ Rossetti *et al.* (2022)	V2-R101	ViT-B/16	70.3	70.9

now want to evaluate the impact of the corpus selected, as well as the amount of synonyms needed to achieve satisfactory results. In particular, we evaluate the performance of the proposed approach (Fig I-3), without adapters, i.e., 'A photo of [CLS*].', when the optimal [CLS] token is selected from a set of potential synonyms extracted from different corpus: British National Corpus Consortium *et al.* (2007), Google News Kutuzov, Fares, Oepen & Velldal (2017) and English Wikipedia. The first observation is that, while the use of different corpus increase the performance over the baseline (*'A photo of [CLS].'* in Table I-1), synonyms from ChatGPT yield

the best performance, regardless on the number of names requested. This may be explained by the larger and richer body of text, from a variety of sources, used to train ChatGPT. Next, we can observe that the quality of the generated CAMs typically augments with the number of synonyms (e.g., with ChatGPT we obtain a mIoU of 72.2 *vs* 73.6, from 2 and 4 synonyms, respectively). These results showcase how the most semantically related words, from a natural language standpoint, do not always yield the best performance. Indeed, as the performance increases with the number of synonyms included in our method, one can easily deduce that the synonyms newly added (less correlated than the first ones) may provide better supervisory signals for certain images. Additionally, we investigate the frequency that the actual ground truth category is selected as the [CLS] token, whose results are depicted in Fig. I-4. The findings from this radar plots reveal that, surprisingly, the ground truth associated with most instances does not correspond to the most correlated category, which may explain the performance gains observed in our approach. Further exploration on the choice of the synonyms across corpus can be found in Supplemental Material.

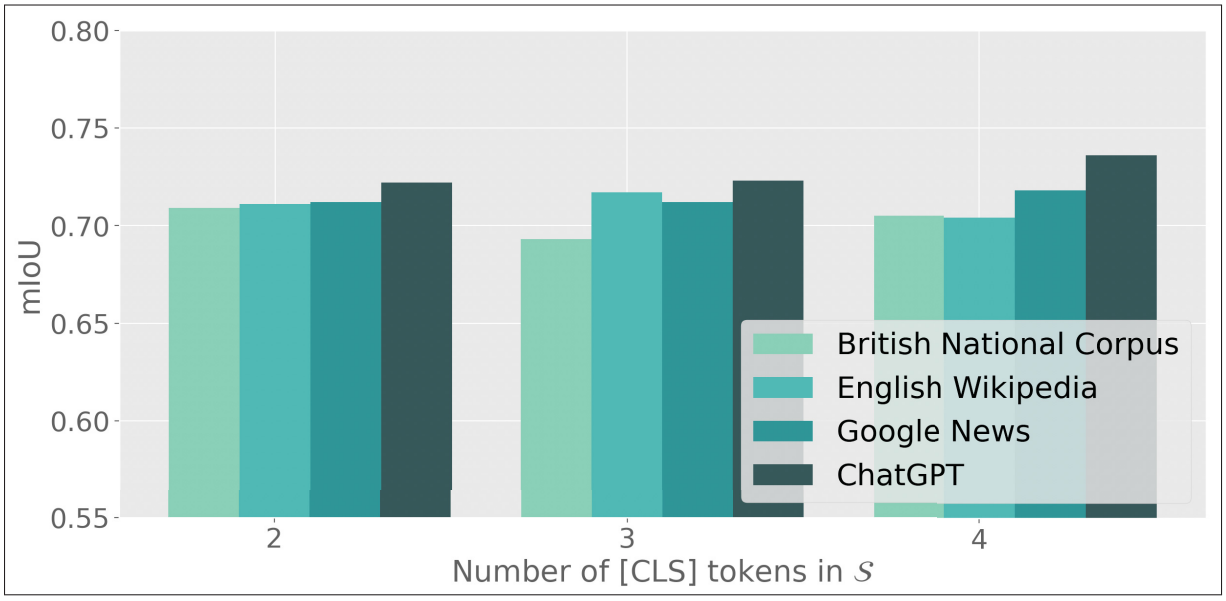

Figure-A I-3 **Impact of the Corpus choice and number of synonyms selected.** ChatGPT offers the richer variety of synonyms, yielding the best results across other corpus. Furthermore, increasing the number of synonyms (from 2 up to 4) further improves the results. Note that the number of synonyms includes the categorical name from the ground truth and the requested close synonyms

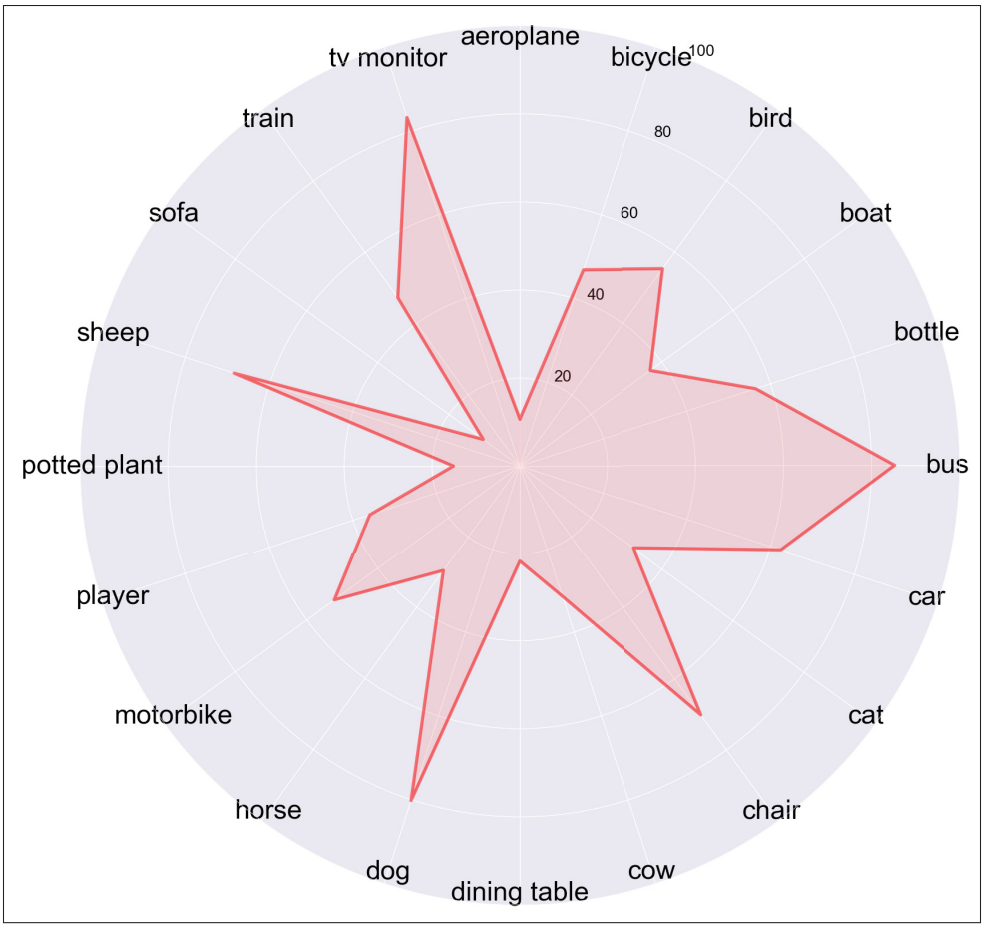

Figure-A I-4 **What does CLIP think about the best [CLS]? Is the ground truth category chosen everytime? How likely is it that CLIP will select something different?** The plot summarises the percentage of cases where the ground truth category was chosen for an instance of that class. Thus, an inward point on the radial plot indicates that the number of instances where the ground truth category was chosen as the best [CLS] token is considerably low

On the impact of the different components. We observed in Table I-1 that the proposed yet simple strategy to optimize the category name achieves better performance than arguably more complex techniques that attempt to optimize the whole text prompt. In this section, we empirically motivate the use of the proposed adapters, as well as the choice of adding learnable parameters to control the importance of each term in Eq. A I-7, unlike the fixed hyperparameter used in the existing literature. In particular, table I-4 reports these results, where the first observation is that adding the proposed adaptors results in slight improvement compared to the

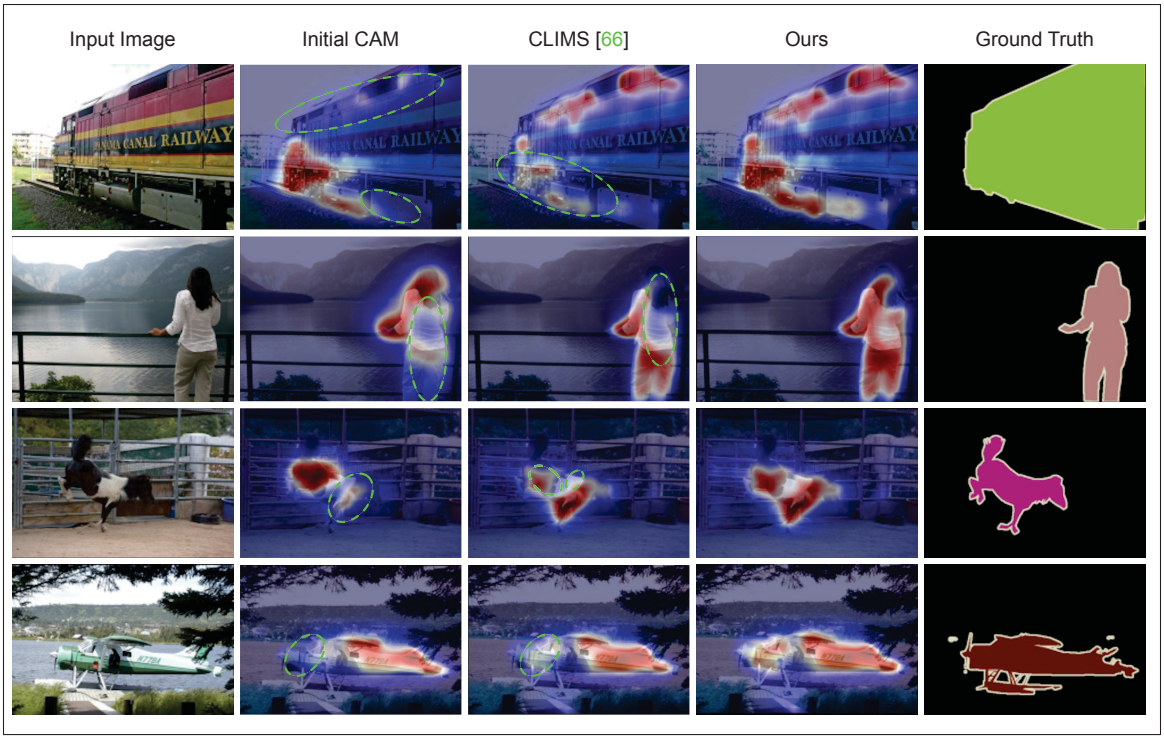

Figure-A I-5 **Qualitative results of the initial class activation maps.** Green dotted lines ellipses are used to indicate missed regions by previous approaches (original CAMs and CLIMS Xie *et al.* (2022)) compared to the proposed method. No refinement on the obtained CAMs is done (e.g., **RW**) to better illustrate the impact of our approach

model without them. In contrast, replacing the fixed vectors by learnable ones, the performance is further improved by 0.4%. Thus, the negligible increase in model complexity due to the adapters, and the performance gain observed, support the choices behind our approach.

Table-A I-4 **Ablation on the main components**. Empirical results that validate the different components of the proposed methodology. A) Image label as [CLS], i.e., CLIMS Xie *et al.* (2022); B) Optimal [CLS] selected; C) Fixed adaptor; D) Learnable adaptor. Results are performed on the *train* set of PASCAL VOC2012

A	B	C	D	mIoU (%)
✓				70.5
	✓			$73.6_{(+3.1)}$
	✓	✓		$73.8_{(+3.3)}$
	✓		✓	$74.2_{(+3.7)}$

Qualitative results of the obtained CAMs, compared to standard CAM generation and the related CLIMS Xie *et al.* (2022) are depicted in Figure I-5. We can observe that despite CLIMS somehow alleviates the under-segmentation problem in conventional CAMs, it still fails to cover larger target regions (see for example first and second columns). In contrast, POLE typically identifies better larger semantic regions related to the target class, resulting in more complete CAMs compared to related approaches.

7. Conclusions

In this work, we have investigated the potential of prompt tuning, an emerging strategy to adapt large pre-trained language-vision models, in the challenging task of weakly supervised semantic segmentation. Our empirical observations have demonstrated that simply replacing the text-token associated with the category name yields better segmentation performance than more complex prompt learning strategies focusing on optimizing the context, which dominate the literature in adapting models. More interestingly, we have observed that employing the corresponding image-level ground truth does not always lead to the best segmentation performance, and closely-related synonyms can indeed result in further performance gains. In light of these findings, we have introduced a simple yet efficient approach, POLE, that selects the most correlated class for a given image in order to generate a better text prompt. Comprehensive experiments have shown that the proposed approach can generate high quality pseudo-labels for WSSS, and achieve state-of-the-art performance in a popular WSSS benchmark.

BIBLIOGRAPHY

[Generated text should be treated as such]. Information provided by ChatGPT, an artificial intelligence language model developed by OpenAI. Retrieved from: Accessedon[March23rd,2023].

(2012). Practical Bayesian Optimization of Machine Learning Algorithms. *Adv. Neural Inf. Process. Syst. 25*, 1–9. doi: 2012arXiv1206.2944S.

(2019). Calibration: the Achilles heel of predictive analytics. *BMC medicine*, 17(1), 230.

Adiga, S., Dolz, J. & Lombaert, H. (2024). Anatomically-aware uncertainty for semi-supervised image segmentation. *Medical Image Analysis*, 91, 103011.

Ahn, J. & Kwak, S. (2018a). Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4981–4990.

Ahn, J. & Kwak, S. (2018b). Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4981–4990.

Ahn, J., Cho, S. & Kwak, S. (2019). Weakly supervised learning of instance segmentation with inter-pixel relations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2209–2218.

Al-Dhabyani, W., Gomaa, M., Khaled, H. & Fahmy, A. (2020). Dataset of breast ultrasound images. *Data in brief*, 28, 104863.

Alpher, F. (2002). Frobnication. *PAMI*, 12(1), 234–778.

Alpher, F. & Fotheringham-Smythe, F. (2003). Frobnication revisited. *Journal of Foo*, 13(1), 234–778.

Alpher, F. & Gamow, F. (2005). Can a computer frobnicate? *CVPR*, pp. 234–778.

Alpher, F., Fotheringham-Smythe, F. & Gamow, F. (2004). Can a machine frobnicate? *Journal of Foo*, 14(1), 234–778.

Amini, A., Schwarting, W., Soleimany, A. & Rus, D. (2020). Deep evidential regression. *Advances in neural information processing systems*, 33, 14927–14937.

Andéol, L., Fel, T., De Grancey, F. & Mossina, L. (2023). Confident object detection via conformal prediction and conformal risk control: an application to railway signaling. *Conformal and Probabilistic Prediction with Applications*, pp. 36–55.

Angelopoulos, A. N. & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Angelopoulos, A. N., Bates, S., Jordan, M. & Malik, J. (2020). Uncertainty Sets for Image Classifiers using Conformal Prediction. *International Conference on Learning Representations*.

Angelopoulos, A. N., Bates, S., Jordan, M. & Malik, J. (2021). Uncertainty Sets for Image Classifiers using Conformal Prediction. *International Conference on Learning Representations*. Retrieved from: https://openreview.net/forum?id=eNdiU_DbM9.

Anonymous. [ECCV submission ID 00324, supplied as supplemental material `00324.pdf`]. (2024a). The frobnicatable foo filter.

Anonymous. [Supplied as supplemental material `tr.pdf`]. (2024b). Frobnication tutorial.

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M. et al. (2022). The medical segmentation decathlon. *Nature communications*, 13(1), 4128.

Avidan, S., Brostow, G., Cissé, M., Farinella, G. M. & Hassner, T. (Eds.). (2022). *Computer Vision – ECCV 2022*. Springer. doi: 10.1007/978-3-031-19769-7.

Badrinarayanan, V., Kendall, A. & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12), 2481–2495.

Bahnsen, A. C., Stojanovic, A., Aouada, D. & Ottersten, B. (2014). Improving credit card fraud detection with calibrated probabilities. *Proceedings of the 2014 SIAM international conference on data mining*, pp. 677–685.

Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P. M. & Rueckert, D. (2017). Semi-supervised learning for network-based cardiac MR image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 253–260.

Bai, Y., Mei, J., Yuille, A. L. & Xie, C. (2021). Are transformers more robust than cnns? *Advances in neural information processing systems*, 34, 26831–26843.

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

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

Bar, A., Huger, F., Schlicht, P. & Fingscheidt, T. (2019). On the Robustness of Redundant Teacher-Student Frameworks for Semantic Segmentation. *CVPRW*.

Barfoot, T., Garcia Peraza Herrera, L. C., Glocker, B. & Vercauteren, T. (2024). Average calibration error: A differentiable loss for improved reliability in image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 139–149.

Bateson, M., Kervadec, H., Dolz, J., Lombaert, H. & Ben Ayed, I. (2020). Source-Relaxed Domain Adaptation for Image Segmentation. *MICCAI*.

Baur, C., Albarqouni, S. & Navab, N. (2017). Semi-supervised deep learning for fully convolutional networks. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 311–319.

Bearman, A., Russakovsky, O., Ferrari, V. & Fei-Fei, L. (2016). What's the point: Semantic segmentation with point supervision. *Proceedings of the European Conference on Computer Vision*, pp. 549–565.

Bensch, R. & Ronneberger, O. (2015). Cell segmentation and tracking in phase contrast images using graph cut with asymmetric boundary costs. *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pp. 1220–1223.

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Ballester, M. A. G. et al. (2018). Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE TMI*, 37(11), 2514–2525.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A. & Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32.

Bertsekas, D. P. (1996a). *Constrained Optimization and Lagrange Multiplier Methods* (ed. 1st). Athena Scientific.

Bertsekas, D. P. (1996b). *Constrained Optimization and Lagrange Multiplier Methods (Optimization and Neural Computation Series)* (ed. 1). Athena Scientific.

Bertsekas, D. (1995). *Nonlinear Programming*. Athena Scientific, Belmont, MA.

Bhalgat, Y., Shah, M. & Awate, S. (2018). Annotation-cost minimization for medical image segmentation using suggestive mixed supervision fully convolutional networks. *Medical Imaging meets NeurIPS Workshop*.

Bhattacharyya, A. (1946). On some analogues of the amount of information and their use in statistical estimation. *Sankhyā: The Indian Journal of Statistics*, 1–14.

Birgin, E. G., Castillo, R. A. & Martínez, J. M. (2005). Numerical comparison of augmented Lagrangian algorithms for nonconvex problems. *Computational Optimization and Applications*, 31(1), 31–55.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag.

Blasco, T., Sánchez, J. S. & García, V. (2024). A survey on uncertainty quantification in deep learning for financial time series prediction. *Neurocomputing*, 576, 127339.

Blundell, C., Cornebise, J., Kavukcuoglu, K. & Wierstra, D. (2015). Weight uncertainty in neural network. *International conference on machine learning*, pp. 1613–1622.

Bohdal, O., Yang, Y. & Hospedales, T. (2021). Meta-calibration: Meta-learning of model calibration using differentiable expected calibration error. *arXiv preprint arXiv:2106.09613*.

Bohdal, O., Yang, Y. & Hospedales, T. (2023). Meta-Calibration: Learning of Model Calibration Using Differentiable Expected Calibration Error. *Transactions on Machine Learning Research*.

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J. et al. (2016). End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*.

Bommasani, R. et al. (2021). On the Opportunities and Risks of Foundation Models. *ArXiv*.

Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I. & de Bruijne, M. (2019). Semi-supervised Medical Image Segmentation via Learning Consistency Under Transformations. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 810–818.

Bossard, L., Guillaumin, M. & Van Gool, L. (2014). Food-101 – Mining Discriminative Components with Random Forests. *European Conference on Computer Vision (ECCV)*.

Boudiaf, M., Ziko, I., Rony, J., Dolz, J., Piantanida, P. & Ben Ayed, I. (2020). Information Maximization for Few-Shot Learning. *NeurIPS*, 33.

Boyd, S. & Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Brier, G. W. et al. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1), 1–3.

Brodley, C. E. & Friedl, M. A. (1999). Identifying mislabeled training data. *Journal of artificial intelligence research*, 11, 131–167.

Brostow, G. J., Shotton, J., Fauqueur, J. & Cipolla, R. (2008). Segmentation and recognition using structure from motion point clouds. *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I 10*, pp. 44–57.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 1877–1901.

Brunekreef, J., Marcus, E., Sheombarsing, R., Sonke, J.-J. & Teuwen, J. (2024). Kandinsky conformal prediction: efficient calibration of image segmentation algorithms. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4135–4143.

Büchel, P., Kratochwil, M., Nagl, M. & Rösch, D. (2022). Deep calibration of financial models: turning theory into practice. *Review of Derivatives Research*, 25(2), 109–136.

Bucher, M., Vu, T.-H., Cord, M. & Pérez, P. (2019). Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32.

Buolamwini, J. & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on fairness, accountability and transparency*, pp. 77–91.

Calegari, R., Castañé, G. G., Milano, M. & O'Sullivan, B. (2023). Assessing and enforcing fairness in the AI lifecycle.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q. & Wang, M. (2022). Swin-unet: Unet-like pure transformer for medical image segmentation. *European conference on computer vision*, pp. 205–218.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1721–1730.

Chaitanya, K., Karani, N., Baumgartner, C. F., Becker, A., Donati, O. & Konukoglu, E. (2019). Semi-supervised and task-driven data augmentation. *International Conference on Information Processing in Medical Imaging*, pp. 29–41.

Chang, Y.-T., Wang, Q., Hung, W.-C., Piramuthu, R., Tsai, Y.-H. & Yang, M.-H. (2020). Weakly-supervised semantic segmentation via sub-category exploration. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8991–9000.

Chapelle, O., Scholkopf, B. & Zien, A. (2009). Semi-supervised learning. *IEEE Transactions on Neural Networks*, 20(3), 542–542.

Charoenphakdee, N., Vongkulbhisal, J., Chairatanakul, N. & Sugiyama, M. (2021). On focal loss for class-posterior probability estimation: A theoretical perspective. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5202–5211.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321–357.

Chen, G., Yao, W., Song, X., Li, X., Rao, Y. & Zhang, K. (2023). Prompt Learning with Optimal Transport for Vision-Language Models. *International Conference on Learning Representations (ICLR)*.

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

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. (2015). Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *International Conference on Learning Representations*.

Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *CVPR*.

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. (2018). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *Transactions on Pattern Analysis and Machine Intelligence*, 40, 834–848.

Chen, L., Wu, W., Fu, C., Han, X. & Zhang, Y. (2020a). Weakly Supervised Semantic Segmentation with Boundary Exploration. *Proceedings of the European Conference on Computer Vision*, pp. 347–362.

Chen, L., Wu, W., Fu, C., Han, X. & Zhang, Y. (2020b). Weakly supervised semantic segmentation with boundary exploration. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pp. 347–362.

Chen, Q., Yang, L., Lai, J.-H. & Xie, X. (2022a). Self-supervised Image-specific Prototype Exploration for Weakly Supervised Semantic Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4288–4298.

Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. (2020c, 2). A simple framework for contrastive learning of visual representations. *International Conference on Machine Learning (ICML)*, pp. 1–11.

Chen, Z., Wang, T., Wu, X., Hua, X.-S., Zhang, H. & Sun, Q. (2022b). Class re-activation maps for weakly-supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 969–978.

Cheng, J. & Vasconcelos, N. (2022). Calibrating deep neural networks by pairwise constraints. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13709–13718.

Chidambaram, M. & Ge, R. (2024). On the Limitations of Temperature Scaling for Distributions with Overlaps. *The Twelfth International Conference on Learning Representations*. Retrieved from: https://openreview.net/forum?id=zavLQJ1XjB.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. (2016). 3D U-Net: learning dense volumetric segmentation from sparse annotation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pp. 424–432.

Cichocki, A. & Amari, S.-i. (2010). Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6), 1532–1568.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S. & Vedaldi, A. (2014). Describing Textures in the Wild. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3606–3613.

Coates, A., Ng, A. & Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223.

Consortium, B. et al. (2007). British national corpus. *Oxford Text Archive Core Collection*.

Corless, R. M., Gonnet, G. H., Hare, D. E., Jeffrey, D. J. & Knuth, D. E. (1996). On the LambertW function. *Advances in Computational mathematics*, 5(1), 329–359.

Cui, W., Liu, Y., Li, Y., Guo, M., Li, Y., Li, X., Wang, T., Zeng, X. & Ye, C. (2019). Semi-supervised brain lesion segmentation with an adapted mean teacher model. *IPMI*, pp. 554–565.

DeGroot, M. & Fienberg, S. (1983). The comparison and evaluation of forecasters. *The Statistician*.

Delaney, E., Greene, D. & Keane, M. T. (2021). Uncertainty estimation and out-of-distribution detection for counterfactual explanations: Pitfalls and solutions. *arXiv preprint arXiv:2107.09734*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255.

Desai, S. & Durrett, G. (2020). Calibration of Pre-trained Transformers. *EMNLP*, pp. 295–302.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT*.

Dhillon, G. S., Chaudhari, P., Ravichandran, A. & Soatto, S. (2019). A Baseline for Few-Shot Image Classification. *ICLR*.

Ding, T., Angelopoulos, A., Bates, S., Jordan, M. & Tibshirani, R. J. (2023). Class-conditional conformal prediction with many classes. *Advances in neural information processing systems*, 36, 64555–64576.

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

Dolz, J., Desrosiers, C. & Ayed, I. B. (2018). 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. *NeuroImage*, 170, 456–470.

Dolz, J., Desrosiers, C., Wang, L., Yuan, J., Shen, D. & Ben Ayed, I. (2020). Deep CNN ensembles and suggestive annotations for infant brain MRI segmentation. *Computerized Medical Imaging and Graphics*, 79, 101660.

Dolz, J., Desrosiers, C. & Ben Ayed, I. (2021). Teach me to segment with mixed supervision: Confident students become masters. *International Conference on Information Processing in Medical Imaging*, pp. 517–529.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. (2021a). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. (2021b). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations*. Retrieved from: https://openreview.net/forum?id=YicbFdNTTy.

Drgoňa, J., Tuor, A. R., Chandan, V. & Vrabie, D. L. (2021). Physics-constrained deep learning of multi-zone building thermal dynamics. *Energy and Buildings*, 243, 110992.

Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S. & Pal, C. (2016). The importance of skip connections in biomedical image segmentation. *International Workshop on Deep Learning in Medical Image Analysis*, pp. 179–187.

Du, Y., Fu, Z., Liu, Q. & Wang, Y. (2022a). Weakly supervised semantic segmentation by pixel-to-prototype contrast. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4320–4329.

Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y. & Li, G. (2022b). Learning to prompt for open-vocabulary object detection with vision-language model. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14084–14093.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J. & Zisserman, A. (2009). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88, 303–308.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J. & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88, 303–338.

Everingham, M., Eslami, S. M., Gool, L., Williams, C. K., Winn, J. & Zisserman, A. (2015). The Pascal Visual Object Classes Challenge: A Retrospective. *IJCV*, 111(1), 98–136. doi: 10.1007/s11263-014-0733-5.

Falkner, S., Klein, A. & Hutter, F. (2018). BOHB: Robust and Efficient Hyperparameter Optimization at Scale. *35th ICML, ICML 2018*, 4, 2323–2341. doi: 10.48550/arxiv.1807.01774.

Fan, J., Zhang, Z., Song, C. & Tan, T. (2020a). Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4283–4292.

Fan, J., Zhang, Z., Tan, T., Song, C. & Xiao, J. (2020b). CIAN: Cross-image affinity net for weakly supervised semantic segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07), 10762–10769.

Fei, Y., Hou, Y., Chen, Z. & Bosselut, A. (2023). Mitigating Label Biases for In-context Learning. *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Fei-Fei, L., Fergus, R. & Perona, P. (2004). Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Worskshops (CVPRW)*, pp. 178–178.

Feng, C.-M., Yu, K., Liu, Y., Khan, S. & Zuo, W. (2023). Diverse data augmentation with diffusions for effective test-time prompt tuning. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2704–2714.

Fernando, K. R. M. & Tsokos, C. P. (2021). Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*.

Fillioux, L., Silva-Rodríguez, J., Ayed, I. B., Cournède, P.-H., Vakalopoulou, M., Christodoulidis, S. & Dolz, J. (2024). Are foundation models for computer vision good conformal predictors? *arXiv preprint arXiv:2412.06082*.

Finn, C., Abbeel, P. & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *International conference on machine learning*, pp. 1126–1135.

Fischer, P., Thomas, K. & Baumgartner, C. F. (2023). Uncertainty estimation and propagation in accelerated mri reconstruction. *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pp. 84–94.

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

Fortunato, M., Blundell, C. & Vinyals, O. (2017). Bayesian recurrent neural networks. *arXiv preprint arXiv:1704.02798*.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200), 675–701.

Fu, H., Cheng, J., Xu, Y., Wong, D. W. K., Liu, J. & Cao, X. (2018). Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. *IEEE transactions on medical imaging*, 37(7), 1597–1605.

Fuchs, M., Gonzalez, C. & Mukhopadhyay, A. (2021). Practical uncertainty quantification for brain tumor segmentation. *Medical Imaging with Deep Learning*.

Furlanello, T., Lipton, Z., Tschannen, M., Itti, L. & Anandkumar, A. (2018). Born Again Neural Networks. *ICML*.

Gal, Y. & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *ICML'16*, pp. 1050–1059.

Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H. & Qiao, Y. (2024). CLIP-Adapter: Better Vision-Language Models with Feature Adapters. *International Journal of Computer Vision (IJCV)*.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R. et al. (2021). A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*.

Geng, S., Yuan, J., Tian, Y., Chen, Y. & Zhang, Y. (2023). HiCLIP: Contrastive Language-Image Pretraining with Hierarchy-aware Attention. *International Conference on Learning Representations (ICLR)*.

Ghosh, A., Schaaf, T. & Gormley, M. (2022). Adafocal: Calibration-aware adaptive focal loss. *Advances in Neural Information Processing Systems*, 35, 1583–1595.

Gong, Y., Lin, X., Yao, Y., Dietterich, T. G., Divakaran, A. & Gervasio, M. (2021). Confidence calibration for domain generalization under covariate shift. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8958–8967.

Goodfellow, I., Bengio, Y. & Courville, A. (2016a). *Deep Learning*. MIT Press.

Goodfellow, I. J., Bengio, Y. & Courville, A. (2016b). *Deep Learning*. Cambridge, MA, USA: MIT Press.

Goyal, S., Kumar, A., Garg, S., Kolter, Z. & Raghunathan, A. (2023). Finetune like you pretrain: Improved finetuning of zero-shot vision models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19338–19347.

Grandvalet, Y. & Bengio, Y. (2005). Semi-supervised learning by entropy minimization. *NeurIPS*.

Graves, A. & Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. *ICML*.

Graves, A., Mohamed, A.-r. & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 6645–6649.

Gu, X., Lin, T.-Y., Kuo, W. & Cui, Y. (2021). Zeroshot detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2(3), 4.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J. et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama*, 316(22), 2402–2410.

Gumbel, E. J. (1954). Statistical Theory of Extreme Values and Some Practical Applications. A Series of Lectures. *Number 33. US Govt. Print. Office*.

Guo et al. (2017a). On calibration of modern neural networks. *ICML*.

Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. (2017b). On calibration of modern neural networks. *International conference on machine learning (ICML)*, pp. 1321–1330.

Guo, M.-H., Lu, C.-Z., Hou, Q., Liu, Z., Cheng, M.-M. & Hu, S.-M. (2022). Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in neural information processing systems*, 35, 1140–1156.

Guoguo Chen,Shuzhou Chai, G. W. J. D. W.-Q. Z. C. W. D. S. D. P. J. T. J. Z. M. J. S. K. S. W. S. Z. W. Z. X. L. X. Y. Y. W. Y. W. Z. Y. Z. Y. (2021). GigaSpeech: An Evolving, Multi-domain ASR Corpus with 10,000 Hours of Transcribed Audio. *Proc. Interspeech 2021*.

Gupta, K., Rahimi, A., Ajanthan, T., Mensink, T., Sminchisescu, C. & Hartley, R. (2020). Calibration of Neural Networks using Splines. *ICLR*.

Habibpour, M., Gharoun, H., Mehdipour, M., Tajally, A., Asgharnezhad, H., Shamsi, A., Khosravi, A. & Nahavandi, S. (2023). Uncertainty-aware credit card fraud detection using deep learning. *Engineering Applications of Artificial Intelligence*, 123, 106248.

Han, Z., Hao, Y., Dong, L., Sun, Y. & Wei, F. (2023). Prototypical Calibration for Few-shot Learning of Language Models. *The Eleventh International Conference on Learning Representations*.

Hantao Yao,Rui Zhang, C. X. (2023). Visual-Language Prompt Tuning with Knowledge-guided Context Optimization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S. & Malik, J. (2011). Semantic contours from inverse detectors. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 991–998.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R. & Xu, D. (2022). Unetr: Transformers for 3d medical image segmentation. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 574–584.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep Residual Learning for Image Recognition. *CVPR*.

He, K., Zhang, X., Ren, S. & Sun, J. (2015). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December, 770–778. doi: 10.48550/arxiv.1512.03385.

He, K., Zhang, X., Ren, S. & Sun, J. (2016a). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

He, K., Zhang, X., Ren, S. & Sun, J. (2016b). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hebbalaguppe, R., Prakash, J., Madan, N. & Arora, C. (2022). A stitch in time saves nine: A train-time regularizing loss for improved neural network calibration. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16081–16090.

Helber, P., Bischke, B., Dengel, A. & Borth, D. (2018). Introducing Eurosat: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 3606–3613.

Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M. 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*.

Hendrycks, D. & Dietterich, T. (2018). Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *International Conference on Learning Representations*.

Hendrycks, D., Mazeika, M. & Dietterich, T. (2018). Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J. & Song, D. (2019). Natural adversarial examples. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15262–15271.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J. & Lakshminarayanan, B. (2020). AugMix: A Simple Method to Improve Robustness and Uncertainty under Data Shift. *International Conference on Learning Representations (ICLR)*.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J. & Gilmer, J. (2021). The many faces of robustness: A critical analysis of out-of-distribution generalization. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8340–8349.

Hernández-Lobato, J. M. & Adams, R. (2015). Probabilistic backpropagation for scalable learning of bayesian neural networks. *ICML*.

Hinton, G., Vinyals, O. & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735–1780.

Hong, S., Noh, H. & Han, B. (2015). Decoupled deep neural network for semi-supervised semantic segmentation. *NeurIPS*.

Hou, Q., Jiang, P., Wei, Y. & Cheng, M.-M. (2018). Self-erasing network for integral object attention. *Advances in Neural Information Processing Systems*, 31.

Howard, J. (2019, March). Imagewoof: a subset of 10 classes from Imagenet that aren't so easy to classify. GitHub. Retrieved from: https://github.com/fastai/imagenette#imagewoof.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. & Chen, W. (2022a). Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR)*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W. et al. (2022b). Lora: Low-rank adaptation of large language models. *ICLR*, 1(2), 3.

Hu, M., Maillard, M., Zhang, Y., Ciceri, T., La Barbera, G., Bloch, I. & Gori, P. (2020). Knowledge distillation from multi-modal to mono-modal segmentation networks. *MICCAI*.

Hu, S. X., Li, D., Stühmer, J., Kim, M. & Hospedales, T. M. (2022c). Pushing the Limits of Simple Pipelines for Few-Shot Learning: External Data and Fine-Tuning Make a Difference. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9068–9077.

Huang, G., Liu, Z., Maaten, L. V. D. & Weinberger, K. Q. (2016). Densely Connected Convolutional Networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January, 2261–2269. doi: 10.48550/arxiv.1608.06993.

Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708.

Huang, L., Ruan, S., Xing, Y. & Feng, M. (2024). A review of uncertainty quantification in medical image analysis: Probabilistic and non-probabilistic methods. *Medical Image Analysis*, 103223.

Hüllermeier, E. & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3), 457–506.

Huo, X., Xie, L., He, J., Yang, Z., Zhou, W., Li, H. & Tian, Q. (2021). ATSO: Asynchronous Teacher-Student Optimization for Semi-Supervised Image Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1235–1244.

Hwang, S.-H., Kim, M. & Whang, S. E. (2025). T-CIL: Temperature Scaling using Adversarial Perturbation for Calibration in Class-Incremental Learning. *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15339–15348.

Ioffe, S. & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International conference on machine learning*, pp. 448–456.

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2), 203–211.

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

Islyaev, S. & Date, P. (2015). Electricity futures price models: Calibration and forecasting. *European Journal of Operational Research*, 247(1), 144–154.

Jang, E., Gu, S. & Poole, B. (2017a). Categorical reparameterization with gumbel-softmax. *ICLR*.

Jang, E., Gu, S. & Poole, B. (2017b). Categorical reparameterization with gumbel-softmax. *ICLR'17*.

Jang, J., Kong, C., Jeon, D., Kim, S. & Kwak, N. (2023). Unifying vision-language representation space with single-tower transformer. *AAAI*, 37(1), 980–988.

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

Jha, D., Ali, S., Tomar, N. K., Johansen, H. D., Johansen, D., Rittscher, J., Riegler, M. A. & Halvorsen, P. (2021). Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. *Ieee Access*, 9, 40496–40510.

Ji, B., Jung, H., Yoon, J., Kim, K. et al. (2019). Bin-wise temperature scaling (BTS): Improvement in confidence calibration performance through simple scaling techniques. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 4190–4196.

Ji, W., Yu, S., Wu, J., Ma, K., Bian, C., Bi, Q., Li, J., Liu, H., Cheng, L. & Zheng, Y. (2021). Learning calibrated medical image segmentation via multi-rater agreement modeling. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12341–12351.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z. & Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. *International Conference on Machine Learning (ICML)*, pp. 4904–4916.

Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B. & Lim, S.-N. (2022). Visual prompt tuning. *European Conference on Computer Vision (ECCV)*, pp. 709–727.

Jiang, P.-T., Hou, Q., Cao, Y., Cheng, M.-M., Wei, Y. & Xiong, H.-K. (2019). Integral Object Mining via Online Attention Accumulation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2070–2079.

Jiang, P.-T., Yang, Y., Hou, Q. & Wei, Y. (2022). L2G: A Simple Local-to-Global Knowledge Transfer Framework for Weakly Supervised Semantic Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16886–16896.

Jiang, Z., Zhang, Y., Liu, C., Zhao, J. & Liu, K. (2023). Generative Calibration for In-context Learning. *arXiv preprint arXiv:2310.10266*.

Jin, L., Lazarow, J. & Tu, Z. (2017). Introspective classification with convolutional nets. *NeurIPS*, 30.

Jo, S. & Yu, I.-J. (2021). Puzzle-CAM: Improved localization via matching partial and full features. *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 639–643.

Joy, T., Pinto, F., Lim, S.-N., Torr, P. H. & Dokania, P. K. (2023). Sample-dependent adaptive temperature scaling for improved calibration. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37, 14919–14926.

Ju, C., Han, T., Zheng, K., Zhang, Y. & Xie, W. (2022). Prompting visual-language models for efficient video understanding. *Proceedings of the European Conference on Computer Vision*, pp. 105–124.

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

Kamiran, F. & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1), 1–33.

Karandikar et al. (2021a). Soft Calibration Objectives for Neural Networks. *NeurIPS*.

Karandikar, A., Cain, N., Tran, D., Lakshminarayanan, B., Shlens, J., Mozer, M. C. & Roelofs, B. (2021b). Soft calibration objectives for neural networks. *Advances in Neural Information Processing Systems*, 34, 29768–29779.

Karani, N., Dey, N. & Golland, P. (2023). Boundary-weighted logit consistency improves calibration of segmentation networks. *MICCAI*, pp. 367–377.

Karimi, D. & Gholipour, A. (2022). Improving Calibration and Out-of-Distribution Detection in Deep Models for Medical Image Segmentation. *IEEE Transactions on Artificial Intelligence*.

Kassapis, E., Dikov, G., Gupta, D. K. & Nugteren, C. (2021). Calibrated adversarial refinement for stochastic semantic segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7057–7067.

Kendall, A. & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Kennedy, M. C. & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3), 425–464.

Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J. & Ayed, I. B. (2019a). Boundary loss for highly unbalanced segmentation. *International conference on medical imaging with deep learning*, pp. 285–296.

Kervadec, H., Dolz, J., Granger, É. & Ben Ayed, I. (2019b). Curriculum semi-supervised segmentation. *MICCAI*.

Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y. & Ayed, I. B. (2019c). Constrained-CNN losses for weakly supervised segmentation. *Medical image analysis*, 54, 88–99.

Kervadec, H., Dolz, J., Wang, S., Granger, E. & Ben Ayed, I. (2020). Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision. *MIDL*.

Kervadec, H., Dolz, J., Yuan, J., Desrosiers, C., Granger, E. & Ayed, I. B. (2022). Constrained deep networks: Lagrangian optimization via log-barrier extensions. *2022 30th European Signal Processing Conference (EUSIPCO)*, pp. 962–966.

Khattak, M. U., Rasheed, H., Maaz, M., Khan, S. & Khan, F. S. (2023). Maple: Multi-modal prompt learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19113–19122.

Khened, M., Kori, A., Rajkumar, H., Krishnamurthi, G. & Srinivasan, B. (2021). A generalized deep learning framework for whole-slide image segmentation and analysis. *Scientific reports*, 11(1), 11579.

Khoreva, A., Benenson, R., Hosang, J., Hein, M. & Schiele, B. (2017). Simple does it: Weakly supervised instance and semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 876–885.

Kim, B., Han, S. & Kim, J. (2021). Discriminative Region Suppression for Weakly-Supervised Semantic Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1754–1761.

Kingma, D. P. & Ba, J. (2015). Adam: A method for stochastic optimization. *ICLR*.

Kingma, D. P., Salimans, T. & Welling, M. (2015). Variational dropout and the local reparameterization trick. *NeurIPS'15*, 28, 2575–2583.

Kock, F., Thielke, F., Chlebus, G. & Meine, H. (2021). Confidence Histograms for Model Reliability Analysis and Temperature Calibration. *Medical Imaging with Deep Learning*.

Koehn, P. & Knowles, R. (2017). Six Challenges for Neural Machine Translation. *First Workshop on Neural Machine Translation*, pp. 28–39.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I. et al. (2021). Wilds: A benchmark of in-the-wild distribution shifts. *International conference on machine learning*, pp. 5637–5664.

Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J. R., Maier-Hein, K., Eslami, S., Jimenez Rezende, D. & Ronneberger, O. (2018). A probabilistic u-net for segmentation of ambiguous images. *Advances in neural information processing systems*, 31.

Kolesnikov, A. & Lampert, C. H. (2016). Seed, expand and constrain: Three principles for weakly-supervised image segmentation. *Proceedings of the European Conference on Computer Vision*, pp. 695–711.

Krause, J., Stark, M., Deng, J. & Fei-Fei, L. (2012). 3d object representations for fine-grained categorization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3498–3505.

Krause, J., Stark, M., Deng, J. & Fei-Fei, L. (2013). 3d object representations for fine-grained categorization. *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561.

Krishnan, R. & Tickoo, O. (2020). Improving model calibration with accuracy versus uncertainty optimization. *Advances in Neural Information Processing Systems*, 33, 18237–18248.

Krizhevsky, A. & Hinton, G. (2009). *Learning multiple layers of features from tiny images*.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *NeurIPS*, 25.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90.

Kull, M., Filho, T. S. & Flach, P. (2017). Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 54, 623–631.

Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H. & Flach, P. (2019). Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *NeurIPS*, 32.

Kumar, A., Liang, P. S. & Ma, T. (2019). Verified uncertainty calibration. *Advances in neural information processing systems*, 32.

Kumar, A., Raghunathan, A., Jones, R., Ma, T. & Liang, P. (2022). Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. *International Conference on Learning Representations (ICLR)*, pp. 1–42.

Kumar, A. & Sarawagi, S. (2019). Calibration of encoder decoder models for neural machine translation. *arXiv preprint arXiv:1903.00802*.

Kumar, A., Sarawagi, S. & Jain, U. (2018a). Trainable calibration measures for neural networks from kernel mean embeddings. *ICML*, pp. 2805–2814.

Kumar, A. et al. (2018b). Trainable calibration measures for neural networks from kernel mean embeddings. *ICML*.

Kutuzov, A., Fares, M., Oepen, S. & Velldal, E. (2017). Word vectors, reuse, and replicability: Towards a community repository of large-text resources. *Proceedings of the 58th Conference on Simulation and Modelling*, pp. 271–276.

Kuzucu, S., Oksuz, K., Sadeghi, J. & Dokania, P. K. (2024). On calibration of object detectors: Pitfalls, evaluation and baselines. *European Conference on Computer Vision*, pp. 185–204.

Kweon, H., Yoon, S.-H., Kim, H., Park, D. & Yoon, K.-J. (2021). Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6994–7003.

Lakshminarayanan, B., Pritzel, A. & Blundell, C. (2017a). Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *NeurIPS*.

Lakshminarayanan, B., Pritzel, A. & Blundell, C. (2017b). Simple and scalable predictive uncertainty estimation using deep ensembles. *NeurIPS'17*.

Lane, R. O. (2025). A comprehensive review of classifier probability calibration metrics. *arXiv preprint arXiv:2504.18278*.

Lang, K. (1995). Newsweeder: Learning to filter netnews. In *Machine learning proceedings 1995* (pp. 331–339). Elsevier.

Larrazabal, A., Martinez, C., Dolz, J. & Ferrante, E. (2023a). Maximum Entropy on Erroneous Predictions (MEEP): Improving model calibration for medical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

Larrazabal, A. J., Martínez, C., Dolz, J. & Ferrante, E. (2021). Orthogonal Ensemble Networks for Biomedical Image Segmentation. *MICCAI*.

Larrazabal, A. J., Martínez, C., Dolz, J. & Ferrante, E. (2023b). Maximum entropy on erroneous predictions: Improving model calibration for medical image segmentation. *MICCAI*, pp. 273–283.

LastName, F. [Face and Gesture submission ID 324. Supplied as supplemental material `fg324.pdf`]. (2014a). The frobnicatable foo filter.

LastName, F. [Supplied as supplemental material `tr.pdf`]. (2014b). Frobnication tutorial.

Laves, M.-H., Ihler, S., Kortmann, K.-P. & Ortmaier, T. (2019). Well-calibrated model uncertainty with temperature scaling for dropout variational inference. *arXiv preprint arXiv:1909.13550*.

Le, Y. & Yang, X. S. (2015). Tiny ImageNet Visual Recognition Challenge.

LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.

Lee, D.-H. et al. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *Workshop on challenges in representation learning, ICML*, 3(2), 896.

Lee, J., Kim, E., Lee, S., Lee, J. & Yoon, S. (2019). Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5267–5276.

Lee, J., Choi, J., Mok, J. & Yoon, S. (2021a). Reducing information bottleneck for weakly supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 34, 27408–27421.

Lee, J., Kim, E. & Yoon, S. (2021b). Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4071–4080.

Lee, J., Oh, S. J., Yun, S., Choe, J., Kim, E. & Yoon, S. (2022a). Weakly supervised semantic segmentation using out-of-distribution data. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16897–16906.

Lee, M., Kim, D. & Shim, H. (2022b). Threshold matters in WSSS: manipulating the activation for the robust and accurate segmentation model against thresholds. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4330–4339.

Lee, S., Lee, M., Lee, J. & Shim, H. (2021c). Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5495–5505.

Lester, B., Al-Rfou, R. & Constant, N. (2021). The Power of Scale for Parameter-Efficient Prompt Tuning. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059.

LeVine, W., Pikus, B., Raja, P. & Gil, F. A. (2023). Enabling calibration in the zero-shot inference of large vision-language models. *arXiv preprint arXiv:2303.12748*.

Li, B., Weinberger, K. Q., Belongie, S., Koltun, V. & Ranftl, R. (2021). Language-driven Semantic Segmentation. *International Conference on Learning Representations*.

Li, J., Jie, Z., Wang, X., Wei, X. & Ma, L. (2022a). Expansion and Shrinkage of Localization for Weakly-Supervised Semantic Segmentation. *Advances in Neural Information Processing Systems*.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A. & Talwalkar, A. (2016). Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *Journal of Machine Learning Research*, 18, 1–52. doi: 10.48550/arxiv.1603.06560.

Li, R., Mai, Z., Trabelsi, C., Zhang, Z., Jang, J. & Sanner, S. (2022b). TransCAM: Transformer Attention-based CAM Refinement for Weakly Supervised Semantic Segmentation. *arXiv preprint arXiv:2203.07239*.

Li, X. L. & Liang, P. (2021). Prefix-Tuning: Optimizing Continuous Prompts for Generation. *ACL*, pp. 4582–4597.

Li, X. & Roth, D. (2002). Learning question classifiers. *COLING 2002: The 19th International Conference on Computational Linguistics*.

Li, X., Lian, D., Lu, Z., Bai, J., Chen, Z. & Wang, X. (2024). Graphadapter: Tuning vision-language models with dual knowledge graph. *Advances in Neural Information Processing Systems (NeurIPS)*, 36.

Li, Y., Duan, Y., Kuang, Z., Chen, Y., Zhang, W. & Li, X. (2022c). Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2), 1447–1455.

Liang, G., Zhang, Y., Wang, X. & Jacobs, N. (2020). Improved trainable calibration method for neural networks on medical imaging classification. *arXiv preprint arXiv:2009.04057*.

Liang, S., Li, Y. & Srikant, R. (2018). Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. *International Conference on Learning Representations*. Retrieved from: https://openreview.net/forum?id=H1VGkIxRZ.

Liang, W., Zhang, Y., Kwon, Y., Yeung, S. & Zou, J. (2022a). Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. *Advances in Neural Information Processing Systems (NeurIPS)*.

Liang, X., Wu, Y., Han, J., Xu, H., Xu, C. & Liang, X. (2022b). Effective adaptation in multi-task co-training for unified autonomous driving. *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 19645–19658.

Lin, D., Dai, J., Jia, J., He, K. & Sun, J. (2016). Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3159–3167.

Lin, M., Chen, Q. & Yan, S. (2014a). Network in network. *ICLR*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L. (2014b). Microsoft coco: Common objects in context. *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755.

Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. (2017). Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988.

Lin, Z., Yu, S., Kuang, Z., Pathak, D. & Ramanan, D. (2023). Multimodality Helps Unimodality: Cross-Modal Few-Shot Learning with Multimodal Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Linmans, J., Elfwing, S., van der Laak, J. & Litjens, G. (2023). Predictive uncertainty estimation for out-of-distribution detection in digital pathology. *Medical Image Analysis*, 83, 102655.

Litjens, G., Toth, R., van de Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J. et al. (2014). Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. *Medical image analysis*, 18(2), 359–373.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B. & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *MedIA*, 42, 60–88.

Liu, B., Ben Ayed, I., Galdran, A. & Dolz, J. (2022a). The Devil is in the Margin: Margin-based Label Smoothing for Network Calibration. *CVPR*.

Liu, B., Ben Ayed, I., Galdran, A. & Dolz, J. (2022b). The devil is in the margin: Margin-based label smoothing for network calibration. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 80–88.

Liu, B., Rony, J., Galdran, A., Dolz, J. & Ben Ayed, I. (2023a). Class Adaptive Network Calibration. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16070–16079.

Liu, J., Zhang, Y., Chen, J.-N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y. & Zhou, Z. (2023b). Clip-driven universal model for organ segmentation and tumor detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21152–21164.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H. & Neubig, G. (2023c). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), 1–35.

Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z. & Wang, J. (2019). Structured knowledge distillation for semantic segmentation. *CVPR*, pp. 2604–2613.

Liu, Y., Wu, Y.-H., Wen, P.-S., Shi, Y.-J., Qiu, Y. & Cheng, M.-M. (2020). Leveraging instance-, image-and dataset-level information for weakly supervised instance segmentation. *Transactions on Pattern Analysis and Machine Intelligence*.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T. & Xie, S. (2022c). A convnet for the 2020s. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986.

Logan IV, R., Balažević, I., Wallace, E., Petroni, F., Singh, S. & Riedel, S. (2022). Cutting Down on Prompts and Parameters: Simple Few-Shot Learning with Language Models. *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2824–2835.

Long, J., Shelhamer, E. & Darrell, T. (2015a). Fully convolutional networks for semantic segmentation. *CVPR*.

Long, J., Shelhamer, E. & Darrell, T. (2015b). Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440.

Loquercio, A., Segu, M. & Scaramuzza, D. (2020). A general framework for uncertainty estimation in deep learning. *IEEE Robotics and Automation Letters*, 5(2), 3153–3160.

Louizos, C. & Welling, M. (2016). Structured and efficient variational deep learning with matrix gaussian posteriors. *ICML*.

Lukasik, M., Bhojanapalli, S., Menon, A. & Kumar, S. (2020). Does label smoothing mitigate label noise? *ICML*.

Luketina, J., Berglund, M., Greff, K. & Raiko, T. (2015). Scalable Gradient-Based Tuning of Continuous Regularization Hyperparameters. *33rd ICML, ICML 2016*, 6, 4333–4341. doi: 10.48550/arxiv.1511.06727.

Luo, W. & Yang, M. (2020a). Learning Saliency-Free Model with Generic Features for Weakly-Supervised Semantic Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Luo, W. & Yang, M. (2020b). Semi-supervised semantic segmentation via strong-weak dual-branch network. *ECCV*.

Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X. & Martel, A. L. (2021a). Loss odyssey in medical image segmentation. *Medical image analysis*, 71, 102035.

Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., Cao, S., Zhang, Q., Liu, S., Wang, Y., Li, Y., He, J. & Yang, X. (2021b). AbdomenCT-1K: Is Abdominal Organ Segmentation A Solved Problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*. doi: 10.1109/TPAMI.2021.3100536.

Ma, J., He, Y., Li, F., Han, L., You, C. & Wang, B. (2024). Segment anything in medical images. *Nature Communications*, 15(1), 654.

Ma, X. & Blaschko, M. B. (2021). Meta-Cal: Well-controlled Post-hoc Calibration by Ranking. *ICML*.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y. & Potts, C. (2011, June). Learning Word Vectors for Sentiment Analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150. Retrieved from: http://www.aclweb.org/anthology/P11-1015.

Maddison, C. J., Mnih, A. & Teh, Y. W. (2017). The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. *5th International Conference on Learning Representations (ICLR)*.

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P. & Wilson, A. G. (2019). A simple baseline for bayesian uncertainty in deep learning. *NeurIPS*, 32.

Maier, O., Menze, B. H., von der Gablentz, J., Häni, L., Heinrich, M. P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., Christiaens, D., Dutil, F., Egger, K., Feng, C., Glocker, B., Götz, M., Haeck, T., Halme, H.-L., Havaei, M., Iftekharuddin, K. M., Jodoin, P.-M., Kamnitsas, K., Kellner, E., Korvenoja, A., Larochelle, H., Ledig, C., Lee, J.-H., Maes, F., Mahmood, Q., Maier-Hein, K. H., McKinley, R., Muschelli, J., Pal, C., Pei, L., Rangarajan, J. R., Reza, S. M., Robben, D., Rueckert, D., Salli, E., Suetens, P., Wang, C.-W., Wilms, M., Kirschke, J. S., Krämer, U. M., Münte, T. F., Schramm, P., Wiest, R., Handels, H. & Reyes, M. (2017). ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. *Medical Image Analysis*, 35, 250–269. doi: https://doi.org/10.1016/j.media.2016.07.009.

Maji, S., Kannala, J., Rahtu, E., Blaschko, M. & Vedaldi, A. (2013). Fine-Grained Visual Classification of Aircraft. *ArXiv Preprint*.

Malinin, A., Mlodozeniec, B. & Gales, M. (2019). Ensemble Distribution Distillation. *International Conference on Learning Representations*.

Marcus, M., Santorini, B. & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2), 313–330.

Marquez Neila, P., Salzmann, M. & Fua, P. (2017). Imposing Hard Constraints on Deep Networks: Promises and Limitations. *CVPR Workshop on Negative Results in Computer Vision*, (CONF).

Maška, M., Ulman, V., Svoboda, D., Matula, P., Matula, P., Ederra, C., Urbiola, A., España, T., Venkatesan, S., Balak, D. M. et al. (2014). A benchmark for comparison of cell tracking algorithms. *Bioinformatics*, 30(11), 1609–1617.

Maška, M., Ulman, V., Delgado-Rodriguez, P., Gómez-de Mariscal, E., Nečasová, T., Guerrero Peña, F. A., Ren, T. I., Meyerowitz, E. M., Scherr, T., Löffler, K. et al. (2023). The cell tracking challenge: 10 years of objective benchmarking. *Nature Methods*, 20(7), 1010–1020.

Mehrtash, A., Wells, W. M., Tempany, C. M., Abolmaesumi, P. & Kapur, T. (2020). Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE transactions on medical imaging*, 39(12), 3868–3878.

Mena, J., Pujol, O. & Vitria, J. (2021). A survey on uncertainty estimation in deep learning classification systems from a Bayesian perspective. *ACM Computing Surveys (CSUR)*, 54(9), 1–35.

Mendrik, A. M., Vincken, K. L., Kuijf, H. J., Breeuwer, M., Bouvy, W. H., De Bresser, J., Alansary, A., De Bruijne, M., Carass, A., El-Baz, A. et al. (2015a). MRBrainS challenge: online evaluation framework for brain image segmentation in 3T MRI scans. *Computational intelligence and neuroscience*, 2015.

Mendrik, A. M., Vincken, K. L., Kuijf, H. J., Breeuwer, M., Bouvy, W. H., De Bresser, J., Alansary, A., De Bruijne, M., Carass, A., El-Baz, A. et al. (2015b). MRBrainS challenge: online evaluation framework for brain image segmentation in 3T MRI scans. *Comput. Intell. Neurosci.*, 2015, 1.

Menon, S. & Vondrick, C. (2023). Visual Classification via description from large language models. *International Conference on Learning Representations (ICLR)*, pp. 1–17.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.-A., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., Corso, J. J., Criminisi, A., Das, T., Delingette, H., Demiralp, , Durst, C. R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K. M., Jena, R., John, N. M., Konukoglu, E., Lashkari, D., Mariz, J. A., Meier, R., Pereira, S., Precup, D., Price, S. J., Raviv, T. R., Reza, S. M. S., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H.-C., Shotton, J., Silva, C. A., Sousa, N., Subbanna, N. K., Szekely, G., Taylor, T. J., Thomas, O. M., Tustison, N. J., Unal, G., Vasseur, F., Wintermark, M., Ye, D. H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M. & Van Leemput, K. (2015). The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10), 1993–2024. doi: 10.1109/TMI.2014.2377694.

Merity, S., Xiong, C., Bradbury, J. & Socher, R. (2016). Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26.

Milletari, F., Navab, N. & Ahmadi, S. (2016a). V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *3DV*.

Milletari, F., Navab, N. & Ahmadi, S.-A. (2016b). V-net: Fully convolutional neural networks for volumetric medical image segmentation. *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571.

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D. & Lucic, M. (2021). Revisiting the calibration of modern neural networks. *Advances in neural information processing systems*, 34, 15682–15694.

Mlynarski, P., Delingette, H., Criminisi, A. & Ayache, N. (2019). Deep learning with mixed supervision for brain tumor segmentation. *J. Med. Imaging*, 6, 034002.

Monteiro, M., Le Folgoc, L., Coelho de Castro, D., Pawlowski, N., Marques, B., Kamnitsas, K., van der Wilk, M. & Glocker, B. (2020). Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *Advances in neural information processing systems*, 33, 12756–12767.

Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6), 47–60.

Morales-Álvarez, P., Christodoulidis, S., Vakalopoulou, M., Piantanida, P. & Dolz, J. (2024). BayesAdapter: enhanced uncertainty estimation in CLIP few-shot adaptation. *arXiv preprint arXiv:2412.09718*.

Mossina, L., Dalmau, J. & Andéol, L. (2024). Conformal semantic image segmentation: Post-hoc quantification of predictive uncertainty. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3574–3584.

Mozafari, A. S., Gomes, H. S., Leão, W., Janny, S. & Gagné, C. (2018). Attended temperature scaling: a practical approach for calibrating deep neural networks. *arXiv preprint arXiv:1810.11586*.

Muehleisen, R. T. & Bergerson, J. (2016). Bayesian calibration-what, why and how.

Mukhoti et al. (2020a). Calibrating deep neural networks using focal loss. *NeurIPS*.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P. & Dokania, P. (2020b). Calibrating deep neural networks using focal loss. *NeurIPS*, 33, 15288–15299.

Mukhoti, J., van Amersfoort, J., Torr, P. H. & Gal, Y. (2021). Deep deterministic uncertainty for semantic segmentation. *arXiv preprint arXiv:2111.00079*.

Mukhoti, J., Gal, Y., Torr, P. H. S. & Dokania, P. K. (2023). Fine-tuning can cripple your foundation model; preserving features may be the solution. *ArXiv*.

Müller et al. (2019a). When does label smoothing help? *NeurIPS*.

Müller, R., Kornblith, S. & Hinton, G. E. (2019b). When does label smoothing help? *Advances in neural information processing systems (NeurIPS)*, 32.

Munir, M. A., Khan, M. H., Sarfraz, M. & Ali, M. (2022). Towards improving calibration in object detection under domain shift. *Advances in Neural Information Processing Systems*, 35, 38706–38718.

Munir, M. A., Khan, M. H., Khan, S. & Khan, F. S. (2023a). Bridging precision and confidence: A train-time loss for calibrating object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11474–11483.

Munir, M. A., Khan, S. H., Khan, M. H., Ali, M. & Shahbaz Khan, F. (2023b). Cal-DETR: calibrated detection transformer. *Advances in neural information processing systems*, 36, 71619–71631.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Murugesan, B., Vijaya Raghavan, S., Sarveswaran, K., Ram, K. & Sivaprakasam, M. (2019). Recon-GLGAN: a global-local context based generative adversarial network for MRI reconstruction. *International workshop on machine learning for medical image reconstruction*, pp. 3–15.

Murugesan, B., Adiga Vasudeva, S., Liu, B., Lombaert, H., Ben Ayed, I. & Dolz, J. (2023a). Trust your neighbours: Penalty-based constraints for model calibration. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 572–581.

Murugesan, B., Liu, B., Galdran, A., Ayed, I. B. & Dolz, J. (2023b). Calibrating segmentation networks with margin-based label smoothing. *Medical Image Analysis*, 87, 102826.

Murugesan, B., Hussain, R., Bhattacharya, R., Ben Ayed, I. & Dolz, J. (2024a). Prompting classes: exploring the power of prompt class learning in weakly supervised semantic segmentation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 291–302.

Murugesan, B., Silva-Rodríguez, J., Ayed, I. B. & Dolz, J. (2024b). Robust calibration of large vision-language adapters. *European Conference on Computer Vision*, pp. 147–165.

Murugesan, B., Silva-Rodriguez, J., Ben Ayed, I. & Dolz, J. (2024c). Class and region-adaptive constraints for network calibration. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 57–67.

Murugesan, B., Vasudeva, S. A., Liu, B., Lombaert, H., Ayed, I. B. & Dolz, J. (2025). Neighbor-aware calibration of segmentation networks with penalty-based constraints. *Medical Image Analysis*, 103501.

Naeini, M. P., Cooper, G. & Hauskrecht, M. (2015a). Obtaining well calibrated probabilities using bayesian binning. *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Naeini, M. P., Cooper, G. F. & Hauskrecht, M. (2015b). Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*.

Naeini, M. P., Cooper, G. F. & Hauskrecht, M. (2015c). Obtaining Well Calibrated Probabilities Using Bayesian Binning. *AAAI*.

Nayak, N. V., Yu, P. & Bach, S. H. (2023). Learning to compose soft prompts for compositional zero-shot learning. *International Conference on Learning Representations*.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A. Y. et al. (2011). Reading digits in natural images with unsupervised feature learning. *NIPS workshop on deep learning and unsupervised feature learning*, 2011(5), 7.

Nguyen, K. & O'Connor, B. (2015). Posterior calibration and exploratory analysis for natural language processing models. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1587–1598.

Niculescu-Mizil, A. & Caruana, R. (2005a). Predicting Good Probabilities with Supervised Learning. *Proceedings of the 22nd ICML (ICML)*.

Niculescu-Mizil, A. & Caruana, R. (2005b). Predicting Good Probabilities with Supervised Learning. *ICML*.

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

Nilsback, M.-E. & Zisserman, A. (2008). Automated Flower Classification over a Large Number of Classes. *Indian Conference on Computer Vision, Graphics and Image Processing*.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G. & Tran, D. (2019a, June). Measuring Calibration in Deep Learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G. & Tran, D. (2019b). Measuring Calibration in Deep Learning. *CVPR workshops*, 2(7).

Nocedal, J. & Wright, S. J. (2006). *Numerical Optimization* (ed. 2nd). New York, NY, USA: Springer.

Oh, C., Lim, H., Kim, M., Han, D., Yun, S., Choo, J., Hauptmann, A., Cheng, Z.-Q. & Song, K. (2024). Towards calibrated robust fine-tuning of vision-language models. *Advances in Neural Information Processing Systems*, 37, 12677–12707.

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

Ouali, Y., Bulat, A., Martinez, B. & Tzimiropoulos, G. (2023). Black Box Few-Shot Adaptation for Vision-Language models. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

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

Panayotov, V., Chen, G., Povey, D. & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210. doi: 10.1109/ICASSP.2015.7178964.

Pang, B. & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv preprint cs/0506075*.

Papandreou, G., Chen, L.-C., Murphy, K. P. & Yuille, A. L. (2015). Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. *ICCV*.

Park, H., Noh, J., Oh, Y., Baek, D. & Ham, B. (2023). Acls: Adaptive and conditional label smoothing for network calibration. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3936–3945.

Park, S., Bastani, O., Weimer, J. & Lee, I. (2020). Calibrated prediction with covariate shift via unsupervised domain adaptation. *International Conference on Artificial Intelligence and Statistics*, pp. 3219–3229.

Park, S. Y. & Caragea, C. (2022). On the Calibration of Pre-trained Language Models using Mixup Guided by Area Under the Margin and Saliency. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5364–5374.

Parkhi, O. M., Vedaldi, A., Zisserman, A. & Jawahar, C. (2012). Cats and dogs. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3498–3505.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. & Chintala, S. (2019a). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. & Garnett, R. (Eds.), *NeurIPS 32* (pp. 8024–8035).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. et al. (2019b). Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32.

Patel, G. & Dolz, J. (2022). Weakly supervised segmentation with cross-modality equivariant constraints. *Medical Image Analysis*, 77, 102374.

Pathak, D., Krahenbuhl, P. & Darrell, T. (2015). Constrained convolutional neural networks for weakly supervised segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1796–1804.

Pathiraja, B., Gunawardhana, M. & Khan, M. H. (2023). Multiclass confidence and localization calibration for object detection. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 19734–19743.

Patra, R., Hebbalaguppe, R., Dash, T., Shroff, G. & Vig, L. (2023). Calibrating deep neural networks using explicit regularisation and dynamic data pruning. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1541–1549.

Pawłowska, A., Ćwierz-Pieńkowska, A., Domalik, A., Jaguś, D., Kasprzak, P., Matkowski, R., Fura, Ł., Nowicki, A. & Żołek, N. (2024). Curated benchmark dataset for ultrasound based breast lesion analysis. *Scientific Data*, 11(1), 148.

Pei, J., Wang, C. & Szarvas, G. (2022). Transformer uncertainty estimation with hierarchical stochastic attention. *AAAI*, 36(10), 11147–11155.

Peng, J., Estrada, G., Pedersoli, M. & Desrosiers, C. (2020a). Deep co-training for semi-supervised image segmentation. *Pattern Recognition*, 107, 107269.

Peng, J., Kervadec, H., Dolz, J., Ben Ayed, I., Pedersoli, M. & Desrosiers, C. (2020b). Discretely-constrained deep network for weakly supervised segmentation. *Neural Networks*, 130, 297–308.

Peng, J., Pedersoli, M. & Desrosiers, C. (2020c). Mutual information deep regularization for semi-supervised segmentation. *MIDL*.

Peng, J., Pedersoli, M. & Desrosiers, C. (2021a). Boosting Semi-supervised Image Segmentation with Global and Local Mutual Information Regularization. *arXiv preprint arXiv:2103.04813*.

Peng, L., Wang, H. & Li, J. (2021b). Uncertainty evaluation of object detection algorithms for autonomous vehicles. *Automotive Innovation*, 4(3), 241–252.

Pennington, J., Socher, R. & Manning, C. D. (2014). Glove: Global Vectors for Word Representation. *EMNLP*.

Penso, C., Frenkel, L. & Goldberger, J. (2024). Confidence Calibration of a Medical Imaging Classification System that is Robust to Label Noise. *IEEE Transactions on Medical Imaging*, 1–1. doi: 10.1109/TMI.2024.3353762.

Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł. & Hinton, G. (2017). Regularizing neural networks by penalizing confident output distributions. *ICLR*.

Perone, C. S. & Cohen-Adad, J. (2018). Deep semi-supervised segmentation with weight-averaged consistency targets. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 12–19). Springer.

Pintea, S. L., Lin, Y., Dijkstra, J. & van Gemert, J. C. (2023). A step towards understanding why classification helps regression. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19972–19981.

Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3), 61–74.

Platt, J. C. (1999). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *ADVANCES IN LARGE MARGIN CLASSIFIERS*, 61–74. Retrieved from: http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.1639.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. & Weinberger, K. Q. (2017). On fairness and calibration. *Advances in neural information processing systems*, 30.

Qin, T., Liu, T.-Y. & Li, H. (2010). A General Approximation Framework for Direct Optimization of Information Retrieval Measures. *Inf. Retr.*, 13(4), 375–397. doi: 10.1007/s10791-009-9124-x.

Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T. S. & Barzilay, R. (2023). Conformal Language Modeling. *arXiv preprint arXiv:2306.10193*.

Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T. S. & Barzilay, R. (2024). Conformal Language Modeling. *The Twelfth International Conference on Learning Representations*. Retrieved from: https://openreview.net/forum?id=pzUhfQ74c5.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al. (2021). Learning transferable visual models from natural language supervision. *International conference on machine learning*, pp. 8748–8763.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1), 5485–5551.

Rahaman, R. et al. (2021). Uncertainty quantification and deep ensembles. *Advances in neural information processing systems*, 34, 20063–20075.

Rahimi, A., Gupta, K., Ajanthan, T., Mensink, T., Sminchisescu, C. & Hartley, R. (2020). Post-hoc calibration of neural networks. *arXiv preprint arXiv:2006.12807*.

Rajchl, M., Lee, M. C., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W., Damodaram, M., Rutherford, M. A., Hajnal, J. V., Kainz, B. et al. (2016). Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE TMI*, 36(2), 674–683.

Rajpurkar, P., Zhang, J., Lopyrev, K. & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392.

Ranjbarzadeh, R., Bagherian Kasgari, A., Jafarzadeh Ghoushchi, S., Anari, S., Naseri, M. & Bendechache, M. (2021). Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images. *Scientific reports*, 11(1), 10930.

Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J. & Lu, J. (2022). Denseclip: Language-guided dense prediction with context-aware prompting. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18082–18091.

Ravi, S. N., Dinh, T., Lokhande, V. S. & Singh, V. (2019). Explicitly imposing constraints in deep networks via conditional gradients gives improved generalization and faster convergence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 4772–4779.

Real, E., Aggarwal, A., Huang, Y. & Le, Q. V. (2018). Regularized Evolution for Image Classifier Architecture Search. *arXiv preprint arXiv:1802.01548*.

Recht, B., Roelofs, R., Schmidt, L., & VaishaalShankar. (2019). Do imagenet classifiers generalize to imagenet? *International Conference on Machine Learning (ICML)*, pp. 5389–5400.

Ren, S., He, K., Girshick, R. & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Ricci Lara, M. A., Mosquera, C., Ferrante, E. & Echeveste, R. (2023). Towards unraveling calibration biases in medical image analysis. *Workshop on Clinical Image-Based Procedures*, pp. 132–141.

Roelofs, R., Cain, N., Shlens, J. & Mozer, M. C. (2022). Mitigating bias in calibration error estimation. *International Conference on Artificial Intelligence and Statistics*, pp. 4036–4054.

Romano, Y., Sesia, M. & Candes, E. (2020). Classification with valid and adaptive coverage. *Advances in neural information processing systems*, 33, 3581–3591.

Ronneberger, O., Fischer, P. & Brox, T. (2015a). U-net: Convolutional networks for biomedical image segmentation. *MICCAI*.

Ronneberger, O., Fischer, P. & Brox, T. (2015b). U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241.

Rony, J., Granger, E., Pedersoli, M. & Ben Ayed, I. (2021). Augmented lagrangian adversarial attacks. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7738–7747.

Rossetti, S., Zappia, D., Sanzari, M., Schaerf, M. & Pirri, F. (2022). Max Pooling with Vision Transformers reconciles class and shape in weakly supervised semantic segmentation. *Proceedings of the European Conference on Computer Vision*, pp. 446–463.

Rousseau, A.-J., Becker, T., Bertels, J., Blaschko, M. B. & Valkenborg, D. (2021). Post training uncertainty calibration of deep networks for medical image segmentation. *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1052–1056.

Ru, L., Du, B. & Wu, C. (2021). Learning Visual Words for Weakly-Supervised Semantic Segmentation. *IJCAI*, 5, 6.

Ru, L., Du, B., Zhan, Y. & Wu, C. (2022a). Weakly-Supervised Semantic Segmentation with Visual Words Learning and Hybrid Pooling. *International Journal of Computer Vision*, 130(4), 1127–1144.

Ru, L., Zhan, Y., Yu, B. & Du, B. (2022b). Learning Affinity from Attention: End-to-End Weakly-Supervised Semantic Segmentation with Transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16846–16855.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088), 533–536.

Sadinle, M., Lei, J. & Wasserman, L. (2019). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525), 223–234.

Salehi, S. S. M., Erdogmus, D. & Gholipour, A. (2017). Tversky loss function for image segmentation using 3D fully convolutional deep networks. *International workshop on machine learning in medical imaging*, pp. 379–387.

Sangalli, S., Erdil, E., Hötker, A., Donati, O. F. & Konukoglu, E. (2021). Constrained Optimization to Train Neural Networks on Critical and Under-Represented Classes. *Advances in Neural Information Processing Systems (NeurIPS)*.

Schick, T., Udupa, S. & Schütze, H. (2021). Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9, 1408–1424.

Sedai, S., Antony, B., Rai, R., Jones, K., Ishikawa, H., Schuman, J., Gadi, W. & Garnavi, R. (2019). Uncertainty Guided Semi-supervised Segmentation of Retinal Layers in OCT Images. *MICCAI*.

Shah, M. P., Merchant, S. & Awate, S. (2018). MS-Net:mixed-supervision fully-convolutional networks for full-resolution segmentation. *MICCAI*.

Sharifdeen, A., Munir, M. A., Baliah, S., Khan, S. & Khan, M. H. (2025). O-TPT: Orthogonality Constraints for Calibrating Test-time Prompt Tuning in Vision-Language Models. *arXiv preprint arXiv:2503.12096*.

Shattuck, D. W., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G., Narr, K. L., Poldrack, R. A., Bilder, R. M. & Toga, A. W. (2008). Construction of a 3D probabilistic atlas of human cortical structures. *Neuroimage*, 39(3), 1064–1080.

Shen, Z., Liu, Z., Xu, D., Chen, Z., Cheng, K.-T. & Savvides, M. (2021). Is Label Smoothing Truly Incompatible with Knowledge Distillation: An Empirical Study. *ICLR*.

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D. & Summers, R. M. (2016). Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging*, 35(5), 1285–1298. doi: 10.1109/TMI.2016.2528162.

Shu, M., Nie, W., Huang, D.-A., Yu, Z., Goldstein, T., Anandkumar, A. & Xiao, C. (2022). Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 14274–14289.

Silva Filho, T., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M. & Flach, P. (2023). Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, 1–50.

Silva-Rodriguez, J., Hajimiri, S., Ayed, I. B. & Dolz, J. (2024). A Closer Look at the Few-Shot Adaptation of Large Vision-Language Models. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Silva-Rodríguez, J., Ben Ayed, I. & Dolz, J. (2025). Conformal Prediction for Zero-Shot Models. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. et al. (2016). Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587), 484–489.

Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y. & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A. & Li, C.-L. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33, 596–608.

Song, C., Huang, Y., Ouyang, W. & Wang, L. (2019). Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3136–3145.

Soomro, K., Zamir, A. R. & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. *ArXiv Preprint*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), 1929–1958.

Su, Y., Sun, R., Lin, G. & Wu, Q. (2021). Context Decoupling Augmentation for Weakly Supervised Semantic Segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Sun, G., Wang, W., Dai, J. & Van Gool, L. (2020). Mining cross-image semantics for weakly supervised semantic segmentation. *Proceedings of the European Conference on Computer Vision*, pp. 347–365.

Sun, K., Shi, H., Zhang, Z. & Huang, Y. (2021). Ecs-net: Improving weakly supervised semantic segmentation by using connections between class activation maps. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7283–7292.

Sung, Y.-L., Cho, J. & Bansal, M. (2022). Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5227–5237.

Sutskever, I., Vinyals, O. & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2818–2826.

Tang, M., Djelouah, A., Perazzi, F., Boykov, Y. & Schroers, C. (2018a). Normalized cut loss for weakly-supervised CNN segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1818–1827.

Tang, M., Perazzi, F., Djelouah, A., Ben Ayed, I., Schroers, C. & Boykov, Y. (2018b). On regularized losses for weakly-supervised cnn segmentation. *Proceedings of the European Conference on Computer Vision*, pp. 507–522.

Tarvainen, A. & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30.

Thulasidasan et al. (2019a). On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks NeurIPS Reproducibility Challenge 2019. *NeurIPS*.

Thulasidasan, S., Chennupati, G., Bilmes, J. A., Bhattacharya, T. & Michalak, S. (2019b). On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *NeurIPS*, 32.

Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J. et al. (2021). Mlp-mixer: An all-mlp architecture for vision. *NeurIPS*, 34, 24261–24272.

Tomani, C., Gruber, S., Erdem, M. E., Cremers, D. & Buettner, F. (2021a). Post-Hoc Uncertainty Calibration for Domain Drift Scenarios. *CVPR*.

Tomani, C., Gruber, S., Erdem, M. E., Cremers, D. & Buettner, F. (2021b). Post-hoc uncertainty calibration for domain drift scenarios. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10124–10132.

Tomani, C., Cremers, D. & Buettner, F. (2022). Parameterized temperature scaling for boosting the expressive power in post-hoc uncertainty calibration. *European Conference on Computer Vision (ECCV)*, pp. 555–569.

Torralba, A. & Efros, A. A. (2011). Unbiased look at dataset bias. *CVPR 2011*, pp. 1521–1528.

Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of statistical physics*, 52(1), 479–487.

Tsimpoukelli, M., Menick, J. L., Cabi, S., Eslami, S., Vinyals, O. & Hill, F. (2021). Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34, 200–212.

Tu, W., Deng, W. & Gedeon, T. (2023). A Closer Look at the Robustness of Contrastive Language-Image Pre-Training (CLIP). *Thirty-seventh Conference on Neural Information Processing Systems*. Retrieved from: https://openreview.net/forum?id=wMNpMe0vp3.

Tu, W., Deng, W., Campbell, D., Gould, S. & Gedeon, T. (2024, 21–27 Jul). An Empirical Study Into What Matters for Calibrating Vision-Language Models. *Proceedings of the 41st International Conference on Machine Learning*, 235(Proceedings of Machine Learning Research), 48791–48808. Retrieved from: https://proceedings.mlr.press/v235/tu24a.html.

Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J. & Schön, T. (2019). Evaluating model calibration in classification. *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3459–3467.

Van Amersfoort, J., Smith, L., Teh, Y. W. & Gal, Y. (2020). Uncertainty estimation using a single deep deterministic neural network. *International conference on machine learning*, pp. 9690–9700.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *NeurIPS*, 30.

Vovk, V., Gammerman, A. & Saunders, C. (1999). Machine-learning applications of algorithmic randomness.

Vu, T.-H., Jain, H., Bucher, M., Cord, M. & Pérez, P. (2019). Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. *CVPR*.

Wah, C., Branson, S., Welinder, P., Perona, P. & Belongie, S. (2011a). *The Caltech-UCSD Birds-200-2011 Dataset* (Report n°CNS-TR-2011-001).

Wah, C., Branson, S., Welinder, P., Perona, P. & Belongie, S. (2011b). The caltech-ucsd birds-200-2011 dataset.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. & Bowman, S. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355. doi: 10.18653/v1/W18-5446.

Wang, C. & Golebiowski, J. (2023). Meta-calibration regularized neural networks. *arXiv preprint arXiv:2303.15057*.

Wang, C., Lawrence, C. & Niepert, M. (2021a). Uncertainty Estimation and Calibration with Finite-State Probabilistic RNNs. *ICLR'21*.

Wang, C., Balazs, J., Szarvas, G., Ernst, P., Poddar, L. & Danchenko, P. (2022a). Calibrating imbalanced classifiers with focal loss: An empirical study. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 145–153.

Wang, D.-B., Feng, L. & Zhang, M.-L. (2021b). Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. *NeurIPS*, 34, 11809–11820.

Wang, D., Gong, B. & Wang, L. (2023). On calibrating semantic segmentation models: Analyses and an algorithm. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23652–23662.

Wang, D., Li, M., Ben-Shlomo, N., Corrales, C. E., Cheng, Y., Zhang, T. & Jayender, J. (2019a). Mixed-Supervised Dual-Network for Medical Image Segmentation. *MICCAI*.

Wang, F., Li, M., Lin, X., Lv, H., Schwing, A. G. & Ji, H. (2022b). Learning to Decompose Visual Features with Latent Textual Prompts. *arXiv preprint arXiv:2210.04287*.

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

Wang, H., Ge, S., Lipton, Z. & Xing, E. P. (2019c). Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, J., Chen, J., Liu, J., Tang, D., Chen, D. Z. & Wu, J. (2025). A Survey on Ordinal Regression: Applications, Advances and Prospects. *arXiv preprint arXiv:2503.00952*.

Wang, P., Peng, J., Pedersoli, M., Zhou, Y., Zhang, C. & Desrosiers, C. (2021c). Self-paced and self-consistent co-training for semi-supervised image segmentation. *Medical Image Analysis*, 73, 102146.

Wang, S., Li, Y. & Wei, H. (2024a). Understanding and Mitigating Miscalibration in Prompt Tuning for Vision-Language Models. *arXiv preprint arXiv:2410.02681*.

Wang, S., Wang, J., Wang, G., Zhang, B., Zhou, K. & Wei, H. (2024b, 21–27 Jul). Open-Vocabulary Calibration for Fine-tuned CLIP. *Proceedings of the 41st International Conference on Machine Learning*, 235(Proceedings of Machine Learning Research), 51734–51754. Retrieved from: https://proceedings.mlr.press/v235/wang24bw.html.

Wang, X., Liu, S., Ma, H. & Yang, M.-H. (2020a). Weakly-supervised semantic segmentation by iterative affinity learning. *International Journal of Computer Vision*, 128(6), 1736–1749.

Wang, Y., Zhang, J., Kan, M., Shan, S. & Chen, X. (2020b). Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12275–12284.

Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M. & Liu, T. (2022c). Cris: Clip-driven referring image segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11686–11695.

Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y. & Cao, Y. (2022d). Simvlm: Simple visual language model pretraining with weak supervision. *International Conference on Learning Representations*.

Wang, Z., Zheng, J.-Q., Zhang, Y., Cui, G. & Li, L. (2024c). Mamba-unet: Unet-like pure visual mamba for medical image segmentation. *arXiv preprint arXiv:2402.05079*.

Wei, H., Xie, R., Cheng, H., Feng, L., An, B. & Li, Y. (2022). Mitigating neural network overconfidence with logit normalization. *International Conference on Machine Learning (ICML)*, pp. 23631–23644.

Wei, Y., Feng, J., Liang, X., Cheng, M.-M., Zhao, Y. & Yan, S. (2017). Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1568–1576.

Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J. & Huang, T. S. (2018). Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. *CVPR*, pp. 7268–7277.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S. & Perona, P. (2010). Caltech-UCSD birds 200.

Wenger, J., Kjellström, H. & Triebel, R. (2020). Non-parametric calibration for classification. *International Conference on Artificial Intelligence and Statistics*, pp. 178–190.

Wenzel, F., Snoek, J., Tran, D. & Jenatton, R. (2020). Hyperparameter ensembles for robustness and uncertainty quantification. *NeurIPS*.

Widmann, D., Lindsten, F. & Zachariah, D. (2019). Calibration tests in multi-class classification: A unifying framework. *Advances in neural information processing systems*, 32.

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Gontijo-Lopes, R., Hajishirzi, H., Farhadi, A., Namkoong, H. & Schmidt, L. (2022). Robust fine-tuning of zero-shot models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7959–7971.

Wu, J., Li, X., Xu, S., Yuan, H., Ding, H., Yang, Y., Li, X., Zhang, J., Tong, Y., Jiang, X. et al. (2024). Towards open vocabulary learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7), 5092–5113.

Wu, T., Huang, J., Gao, G., Wei, X., Wei, X., Luo, X. & Liu, C. H. (2021). Embedded discriminative attention mechanism for weakly supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16765–16774.

Wu, T., Gao, G., Huang, J., Wei, X., Wei, X. & Liu, C. H. (2022). Adaptive Spatial-BCE Loss for Weakly Supervised Semantic Segmentation. *Proceedings of the European Conference on Computer Vision*, pp. 199–216.

Wu, Z., Shen, C. & van den Hengel, A. (2019a). Wider or Deeper: Revisiting the ResNet Model for Visual Recognition. *Pattern Recognition*, 90, 119–133.

Wu, Z., Shen, C. & Van Den Hengel, A. (2019b). Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90, 119–133.

Xian, Y., Choudhury, S., He, Y., Schiele, B. & Akata, Z. (2019). Semantic projection network for zero-and few-label semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8256–8265.

Xiao, H., Rasul, K. & Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A. & Torralba, A. (2010). SUN database: Large-scale scene recognition from abbey to zoo. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3485–3492.

Xie, J., Shuai, B., Hu, J.-F., Lin, J. & Zheng, W.-S. (2018). Improving fast segmentation with teacher-student learning. *BMVC*.

Xie, J., Hou, X., Ye, K. & Shen, L. (2022). CLIMS: Cross Language Image Matching for Weakly Supervised Semantic Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4483–4492.

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

Xie, Q., Dai, Z., Hovy, E., Luong, T. & Le, Q. (2020). Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33, 6256–6268.

Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. (2017). Aggregated residual transformations for deep neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500.

Xing, Y., Wu, Q., Cheng, D., Zhang, S., Liang, G., Wang, P. & Zhang, Y. (2023). Dual modality prompt tuning for vision-language pre-trained model. *IEEE Transactions on Multimedia*.

Xu, C. & Xie, Y. (2023). Conformal prediction for time series. *IEEE transactions on pattern analysis and machine intelligence*, 45(10), 11575–11587.

Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., Sohel, F. & Xu, D. (2021a). Leveraging Auxiliary Tasks With Affinity Learning for Weakly Supervised Semantic Segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6984–6993.

Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., Sohel, F. & Xu, D. (2021b). Leveraging Auxiliary Tasks with Affinity Learning for Weakly Supervised Semantic Segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F. & Xu, D. (2022a). Multi-class Token Transformer for Weakly Supervised Semantic Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4310–4319.

Xu, Y., Du, B., Zhang, L., Zhang, Q., Wang, G. & Zhang, L. (2019). Self-ensembling attention networks:Addressing domain shift for semantic segmentation. *AAAI*.

Xu, Z., Lim, S., Shin, H.-K., Uhm, K.-H., Lu, Y., Jung, S.-W. & Ko, S.-J. (2022b). Risk-aware survival time prediction from whole slide pathological images. *Scientific reports*, 12(1), 21948.

Yang, J., Duan, J., Tran, S., Xu, Y., Chanda, S., Chen, L., Zeng, B., Chilimbi, T. & Huang, J. (2022). Vision-language pre-training with triple contrastive learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15671–15680.

Yang, S., Pappas, G. J., Mangharam, R. & Lindemann, L. (2023). Safe Perception-Based Control under Stochastic Sensor Uncertainty using Conformal Prediction. *arXiv preprint arXiv:2304.00194*.

Yang, Z., Dai, Z., Salakhutdinov, R. & Cohen, W. W. (2018). Breaking the Softmax Bottleneck: A High-Rank RNN Language Model. *International Conference on Learning Representations*.

Yao, H., Zhang, R. & Xu, C. (2023). Visual-language prompt tuning with knowledge-guided context optimization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6757–6767.

Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X. & Xu, C. (2021a). Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Yao, Y., Chen, T., Xie, G.-S., Zhang, C., Shen, F., Wu, Q., Tang, Z. & Zhang, J. (2021b). Non-salient region object mining for weakly supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2623–2632.

Ye, P., Huang, Y., Tu, C., Li, M., Chen, T., He, T. & Ouyang, W. (2023). Partial fine-tuning: A successor to full fine-tuning for vision transformers. *arXiv preprint arXiv:2312.15681*.

Yeung, M., Rundo, L., Nan, Y., Sala, E., Schönlieb, C.-B. & Yang, G. (2021). Calibrating the Dice loss to handle neural network overconfidence for biomedical image segmentation. *arXiv preprint arXiv:2111.00528*.

Yim, J., Joo, D., Bae, J. & Kim, J. (2017). A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. *CVPR*.

Yoon, H. S., Yoon, E., Tee, J. T. J., Hasegawa-Johnson, M. A., Li, Y. & Yoo, C. D. (2024). C-TPT: Calibrated Test-Time Prompt Tuning for Vision-Language Models via Text Feature Dispersion. *International Conference on Learning Representations (ICLR)*.

Yoon, S.-H., Kweon, H., Cho, J., Kim, S. & Yoon, K.-J. (2022). Adversarial Erasing Framework via Triplet with Gated Pyramid Pooling Layer for Weakly Supervised Semantic Segmentation. *Proceedings of the European Conference on Computer Vision*, pp. 326–344.

You, H., Zhou, L., Xiao, B., Codella, N., Cheng, Y., Xu, R., Chang, S.-F. & Yuan, L. (2022). Learning visual representation from modality-shared contrastive language-image pre-training. *European Conference on Computer Vision (ECCV)*, pp. 69–87.

Yu, L., Wang, S., Li, X., Fu, C.-W. & Heng, P.-A. (2019). Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 605–613.

Yu, L., Xiang, W., Fang, J., Chen, Y.-P. P. & Chi, L. (2023a). eX-ViT: A Novel eXplainable Vision Transformer for Weakly Supervised Semantic Segmentation. *Pattern Recognition*, 109666.

Yu, T., Lu, Z., Jin, X., Chen, Z. & Wang, X. (2023b). Task Residual for Tuning Vision-Language Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10899–10909.

Zadrozny, B. & Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. *International Conference on Machine Learning (ICML)*, 1, 609–616.

Zadrozny, B. & Elkan, C. (2002a). Transforming Classifier Scores into Accurate Multiclass Probability Estimates. *KDD*, pp. 694–699. doi: 10.1145/775047.775151.

Zadrozny, B. & Elkan, C. (2002b). Transforming Classifier Scores into Accurate Multiclass Probability Estimates. *KDD*.

Zadrozny, B. & Elkan, C. (2002c). Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pp. 694–699.

Zagoruyko, S. & Komodakis, N. (2016). Wide Residual Networks. *BMVC*.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T. & Dwork, C. (2013). Learning fair representations. *International conference on machine learning*, pp. 325–333.

Zhai, X., Kolesnikov, A., Houlsby, N. & Beyer, L. (2021). Scaling Vision Transformers. doi: 10.48550/arxiv.2106.04560.

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A. & Beyer, L. (2022). Lit: Zero-shot transfer with locked-image text tuning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18123–18133.

Zhang, B., Xiao, J., Jiao, J., Wei, Y. & Zhao, Y. (2021a). Affinity Attention Graph Neural Network for Weakly Supervised Semantic Segmentation. *Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, D., Zhang, H., Tang, J., Hua, X. & Sun, Q. (2020a). Causal Intervention for Weakly-Supervised Semantic Segmentation. *Advances in Neural Information Processing Systems*.

Zhang, D., Zhang, H., Tang, J., Hua, X. & Sun, Q. (2020b). Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*.

Zhang, F., Gu, C., Zhang, C. & Dai, Y. (2021b). Complementary patch for weakly supervised semantic segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7242–7251.

Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. (2018a). mixup: Beyond empirical risk minimization. *ICLR*.

Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. (2018b). mixup: Beyond Empirical Risk Minimization. *International Conference on Learning Representations*.

Zhang, J., Kailkhura, B. & Han, T. Y.-J. (2020c). Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. *ICML*, pp. 11117–11128.

Zhang, L., Deng, Z., Kawaguchi, K., Ghorbani, A. & Zou, J. (2021c). How Does Mixup Help With Robustness and Generalization? *ICLR*.

Zhang, L., Deng, Z., Kawaguchi, K. & Zou, J. (2022a). When and how mixup improves calibration. *International Conference on Machine Learning*, pp. 26135–26160.

Zhang, L., Li, X. & Chen, W. (2024). Camp-net: consistency-aware multi-prior network for accelerated MRI reconstruction. *IEEE journal of biomedical and health informatics*.

Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y. & Li, H. (2022b, 11). Tip-Adapter: Training-free CLIP-Adapter for Better Vision-Language Modeling. *European Conference on Computer Vision (ECCV)*, pp. 1–19.

Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y. & Li, H. (2022c). Tip-adapter: Training-free adaption of clip for few-shot classification. *Proceedings of the European Conference on Computer Vision*, pp. 493–510.

Zhang, R., Hu, X., Li, B., Huang, S., Deng, H., Qiao, Y., Gao, P. & Li, H. (2023). Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15211–15222.

Zhang, T., Lin, G., Liu, W., Cai, J. & Kot, A. (2020d). Splitting vs. merging: Mining object regions with discrepancy and intersection loss for weakly supervised semantic segmentation. *Proceedings of the European Conference on Computer Vision*, pp. 663–679.

Zhang, Y. & Zhang, J. (2021). Dual-Task Mutual Learning for Semi-Supervised Medical Image Segmentation. *arXiv preprint arXiv:2103.04708*.

Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D. P. & Chen, D. Z. (2017). Deep adversarial networks for biomedical image segmentation utilizing unannotated images. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 408–416.

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

Zhao, Y., Khalman, M., Joshi, R., Narayan, S., Saleh, M. & Liu, P. J. (2022). Calibrating sequence likelihood improves conditional language generation. *The Eleventh International Conference on Learning Representations*.

Zhao, Y., Joshi, R., Liu, T., Khalman, M., Saleh, M. & Liu, P. J. (2023). Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*.

Zhao, Z., Wallace, E., Feng, S., Klein, D. & Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. *ICML*, pp. 12697–12706.

Zheng, C., Zhou, H., Meng, F., Zhou, J. & Huang, M. (2023). On Large Language Models' Selection Bias in Multi-Choice Questions. *arXiv preprint arXiv:2309.03882*.

Zhong, Z., Friedman, D. & Chen, D. (2021a). Factual Probing Is [MASK]: Learning vs. Learning to Recall. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5017–5033.

Zhong, Z., Cui, J., Liu, S. & Jia, J. (2021b). Improving calibration for long-tailed recognition. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16489–16498.

Zhong, Z., Cui, J., Liu, S. & Jia, J. (2021c). Improving Calibration for Long-Tailed Recognition. *CVPR*.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. (2016). Learning deep features for discriminative localization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929.

Zhou, H., Wan, X., Proleev, L., Mincu, D., Chen, J., Heller, K. & Roy, S. (2023a). Batch calibration: Rethinking calibration for in-context learning and prompt engineering. *arXiv preprint arXiv:2309.17249*.

Zhou, K., Yang, J., Loy, C. C. & Liu, Z. (2022a). Conditional Prompt Learning for Vision-Language Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhou, K., Yang, J., Loy, C. C. & Liu, Z. (2022b). Conditional prompt learning for vision-language models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825.

Zhou, K., Yang, J., Loy, C. C. & Liu, Z. (2022c). Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision (IJCV)*.

Zhou, K., Yang, J., Loy, C. C. & Liu, Z. (2022d). Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9), 2337–2348.

Zhou, T., Zhang, M., Zhao, F. & Li, J. (2022e). Regional semantic contrast and aggregation for weakly supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4299–4309.

Zhou, Y., Wang, Y., Tang, P., Bai, S., Shen, W., Fishman, E. & Yuille, A. (2019). Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 121–140.

Zhou, Z., Lei, Y., Zhang, B., Liu, L. & Liu, Y. (2023b). Zegclip: Towards adapting clip for zero-shot semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11175–11185.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N. & Liang, J. (2020). UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Transactions on Medical Imaging*, 39(6), 1856–1867. doi: 10.1109/TMI.2019.2959609.

Zhu, B., Niu, Y., Han, Y., Wu, Y. & Zhang, H. (2023). Prompt-aligned gradient for prompt tuning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15659–15669.

Zoph, B., Vasudevan, V., Shlens, J. & Le, Q. V. (2018). Learning Transferable Architectures for Scalable Image Recognition. *CVPR*.

Zou, K., Chen, Z., Yuan, X., Shen, X., Wang, M. & Fu, H. (2023). A review of uncertainty estimation and its application in medical imaging. *Meta-Radiology*, 1(1), 100003.