Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Expectation maximisation pseudo labels

Moucheng Xu[a,b,f,*], Yukun Zhou[a,b], Chen Jin[a,c], Marius de Groot[e], Daniel C. Alexander[a,c], Neil P. Oxtoby[a,c,1], Yipeng Hu[a,b,g,1], Joseph Jacob[a,d,f,1]

[a] *UCL Centre for Medical Image Computing (CMIC), University College London, 90 High Holborn, London, WC1V 6LJ, UK*
[b] *UCL Department of Medical Physics and Biomedical Engineering, University College London, Gower Street, London, WC1E 6BT, UK*
[c] *UCL Department of Computer Science, University College London, Gower Street, London, WC1E 6BT, UK*
[d] *UCL Respiratory, University College London, 1st Floor, Rayne Institute, 5 University Street, London, WC1E 6JF, UK*
[e] *GSK, Gunnels Wood Road, Stevenage, SG1 2NY, UK*
[f] *Satsuma Lab, University College Londo, 90 High Holborn, WC1V 6LJ, UK*
[g] *Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS), University College London, Charles Bell House, 1a 43-45 Foley Street, London, W1 W 7TS, UK*

ARTICLE INFO

Keywords:
Pseudo labels
Bayesian deep learning
Expectation–maximisation
Semi-supervised learning
Segmentation
Generative models
Robustness

ABSTRACT

In this paper, we study pseudo-labelling. Pseudo-labelling employs raw inferences on unlabelled data as pseudo-labels for self-training. We elucidate the empirical successes of pseudo-labelling by establishing a link between this technique and the Expectation Maximisation algorithm. Through this, we realise that the original pseudo-labelling serves as an empirical estimation of its more comprehensive underlying formulation. Following this insight, we present a full generalisation of pseudo-labels under Bayes' theorem, termed Bayesian Pseudo Labels. Subsequently, we introduce a variational approach to generate these Bayesian Pseudo Labels, involving the learning of a threshold to automatically select high-quality pseudo labels. In the remainder of the paper, we showcase the applications of pseudo-labelling and its generalised form, Bayesian Pseudo-Labelling, in the semi-supervised segmentation of medical images. Specifically, we focus on: (1) 3D binary segmentation of lung vessels from CT volumes; (2) 2D multi-class segmentation of brain tumours from MRI volumes; (3) 3D binary segmentation of whole brain tumours from MRI volumes; and (4) 3D binary segmentation of prostate from MRI volumes. We further demonstrate that pseudo-labels can enhance the robustness of the learned representations. The code is released in the following GitHub repository: https://github.com/moucheng2017/EMSSL.

1. Introduction

Recent years have witnessed the rise of deep learning based AI technologies in a wide range of applications for the betterment of humanity. The training of a successful deep learning model demands a large volume of annotated data. Regrettably, the money and time costs associated with the annotation acquisition is very expensive, causing a common issue namely label scarcity. The issue of label scarcity is especially challenging in one of the key applications of AI, healthcare, where the annotation process requires the expertise of highly skilled medical professionals, adding extra costs. In the era of AI-enabled healthcare, medical image segmentation is one of the core tasks, aiming at accurately labelling all pixels within volumetric medical images. It serves as a foundational step for other downstream tasks of healthcare, including computer-aided diagnosis, surgical navigation, and endpoint decision-making in drug discovery, among others. In this paper, we focus on the task of medical image segmentation as an exemplar application.

To tackle the inevitable issue of label scarcity in deep learning, semi-supervised learning has emerged as a solution. This approach utilises both labelled and unlabelled data to enhance model performance. Typically, unlabelled data are more abundant than their labelled counterparts, yet they are often overlooked in supervised learning paradigms. Semi-supervised learning is advantageous because it leverages existing unlabelled data, thereby sidestepping the need for additional investment in label acquisition. While other strategies, such as outsourcing data labelling combined with federated learning (Li et al., 2021), have been developed to address label scarcity, they still do not entirely eliminate the associated costs. In contrast, semi-supervised learning offers an attractive trade-off between cost and performance improvement.

[*] Corresponding author at: UCL Centre for Medical Image Computing (CMIC), University College London, 90 High Holborn, London, WC1V 6LJ, UK.
E-mail address: moucheng.xu.18@alumni.ucl.ac.uk (M. Xu).

[1] This is author footnote for joint senior authors.

https://doi.org/10.1016/j.media.2024.103125
Received 4 March 2023; Received in revised form 18 February 2024; Accepted 26 February 2024
Available online 27 February 2024

1.1. Semi-supervised learning and entropy regularisation

Most semi-supervised learning methods focus on minimising the entropy of predictions for unlabelled data. Entropy minimisation has a long-standing history in the field of representation learning. One of the earliest forms of this approach originated from the mutual information between input and output in unsupervised learning (Bridle and Anthony Heading, 1991). The concept of entropy regularisation gained significant traction in the realm of semi-supervised image classification after it was proposed to minimise the entropy of unlabelled data as a strong form of regularisation (Grandvalet and Bengio, 2004). This technique aims to guide the model towards establishing a reliable decision boundary, focusing on maximising the firmness of the decisions. Since its introduction, entropy minimisation has evolved from its original explicit form into various implicit manifestations.

1.2. Consistency regularisation

Among the multifarious implicit forms of entropy minimisation, consistency regularisation stands as one of the most popular options, serving as the underpinning for a majority of cutting-edge methods in semi-supervised classification and segmentation (Sohn et al., 2020; Berthelot et al., 2020; Li et al., 2018; French et al., 2020; Xu et al., 2022a). Consistency regularisation utilises distance-based loss functions directly applied to the raw outputs or associated prediction probabilities. Consistency regularisation aims to engender predictive models that are resilient to perturbations at either the input or feature levels (Tarvainen and Valpola, 2017; Sohn et al., 2020; Berthelot et al., 2020, 2019; Xu et al., 2022a; Ouali et al., 2020; French et al., 2020; Chen et al., 2021).

For methods relying on input-level consistency, many are derived from a classic model called Mean-Teacher (Tarvainen and Valpola, 2017). In this model, the student's weight is an exponential moving average of the teacher model's weights. The teacher model processes a regular input, while the student model processes the same input with added Gaussian noise. In other words, the student model and the teacher model intake two different views of the same input. A mean square error is used for soft consistency regularisation between the outputs of the two models.

A more advanced teacher–student model, FixMatch, has achieved state-of-the-art performance in semi-supervised classification (Sohn et al., 2020). FixMatch employs two forward passes: one with weakly augmented input (e.g., flipping) and another with strongly augmented input (e.g., shearing, random intensity). The output of the weakly augmented input is then used to generate a pseudo-label as the ground truth for training the output of the strongly augmented input.

Although FixMatch and its variants have excelled in image classification, it has been observed that they are not directly applicable to image segmentation tasks, as the cluster assumption does not hold at the pixel level (French et al., 2020). To adapt consistency regularisation for segmentation, the authors in Ouali et al. (2020) found it feasible to apply perturbations at the feature level rather than the input level before implementing consistency regularisation. They directly apply augmentation techniques to the features of different decoders for semi-supervised image segmentation. Alternatively, perturbations can also be added through architectural modifications. For instance, one can train two identical models with different initialisations and apply consistency regularisation using pseudo-labels on both outputs (Chen et al., 2021). These methods, along with ours, are further tested and compared in a subsequent Section 5.

1.3. Pseudo labelling

Pseudo labelling is another form of entropy regularisation, requiring less computational resource. The concept of pseudo labelling was initially introduced in the context of semi-supervised multi-class image classification (Lee, 2013). In its prototypical form, pseudo labels are generated through the argmax operation applied to the output logits (essentially, the inferential outcomes) of the neural network for unlabelled data. Once generated, these pseudo labels are amalgamated with their corresponding unlabelled data and utilised to train the network in a manner akin to traditional supervised learning. One of the merits of this original approach lies in its computational efficiency; the generation of pseudo labels is performed "on-the-fly", in real-time. It is common practice to initially "warm-up" the network through purely supervised learning, followed by a gradual introduction—or "ramp-up"—of the weight attributed to the pseudo labels in the loss function.

Pseudo labelling has garnered significant attention within both semi-supervised and self-supervised learning paradigms, chiefly owing to its computational frugality coupled with its robust performance metrics. Notably, some empirical studies have posited that semi-supervised learning strategies, leveraging pseudo labelling with vast volumes of internet-sourced unlabelled data, can outperform their fully supervised counterparts in tasks such as ImageNet classification (Pham et al., 2021). Recent research endeavors have sought to curtail the increasing complexity inherent in consistency regularisation techniques. These efforts have yielded competitive performance metrics, achieved solely through the judicious use of pseudo labels (Rizve et al., 2021). Within the domain of image segmentation, novel methodologies underpinned by pseudo labelling have also demonstrated commendable results, especially when the pseudo labels are refined through self-attention mechanisms (Zou et al., 2021).

However, pseudo labelling is not devoid of limitations; a notable issue is the phenomenon of confirmation bias (Arazo et al., 2020). This occurs when erroneously generated pseudo labels are incorporated into the training process, thereby inducing a form of noisy training. The negative impact of these incorrect labels is not merely transient but tends to accumulate and amplify over the course of training. In the present manuscript, we propose a novel methodological framework aimed at mitigating this confirmation bias. Specifically, we introduce a stochastic training paradigm that is designed to learn the threshold of the pseudo labels, thereby generating high quality pseudo labels in an automatic manner (see Fig. 1).

1.4. Motivations and contributions

It has come to our attention that the majority of extant literature on pseudo-labelling primarily adopts an empirical methodology, conspicuously omitting an investigation into the foundational mechanisms underlying its empirically observed efficacy. Motivated by this lacuna, we embarked upon a more in-depth examination of pseudo-labelling and ascertained its significant theoretical relationship with the classical Expectation–Maximisation (EM) algorithm in machine learning. Furthermore, our study is catalysed by contemporary research in the domain of semi-supervised image classification, which posits that achieving competitive performance metrics is feasible through judicious selection of high-quality pseudo-labels (Rizve et al., 2021). In this manuscript, we offer a theoretical exegesis that elucidates the correlation between pseudo-labelling and the EM algorithm. Concurrently, we engage in empirical investigations to assess the applicability and robustness of pseudo-labelling in the context of semi-supervised medical image segmentation. We summarise our contributions in the following bullet points:

- We interpret pseudo labelling as Expectation Maximisation (EM) algorithm. As EM algorithm is guaranteed to converge to local minimum. We therefore partially explain the empirical success of pseudo labelling.

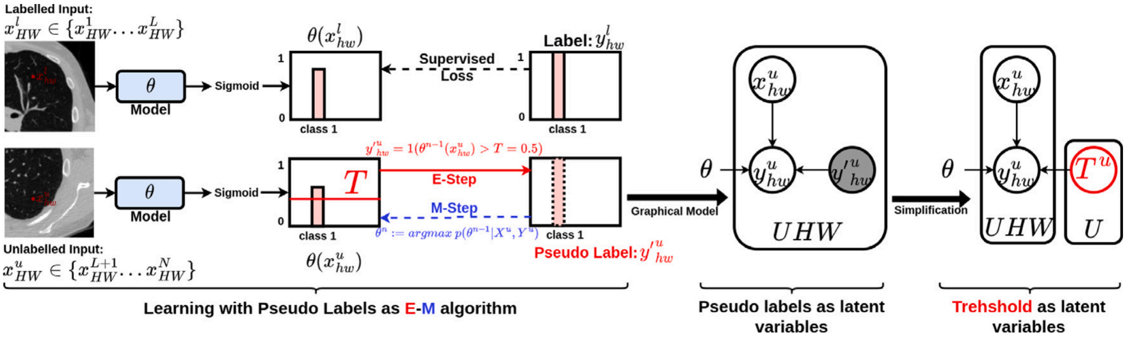

Fig. 1. Pseudo-labelling process for binary segmentation. Pseudo-label y'_n is generated using unlabelled data x_u and model with parameters from last iteration θ. Therefore, pseudo-labelling can be seen as the E-step in Expectation–Maximisation. The M-step updates θ using y'_n, y and data X. In our 1st implementation, namely SegPL, the threshold T is fixed for selecting the pseudo labels, which is the original pseudo labelling, as an empirical approximation of its true generalisation. In our 2nd implementation, namely SegPL-VI, the threshold T is dynamic and learnt via variational inference, which is an learnt approximation of its true generalisation.

- We demonstrate the generalised form of pseudo labels.
- We provide a learning method to learn the threshold for pseudo labelling in order to avoid confirmation bias and automatically pick up high quality pseudo labels.
- We investigate the use of pseudo labelling in semi-supervised medical image segmentation and its characteristics such as robustness.

Previously, a shorter version of this paper has been published at MICCAI 2022 (Xu et al., 2022b). This journal version includes a couple of extensions based on that MICCAI paper such as:

- We fulfilled the details of the proposed probabilistic model of pseudo labelling.
- We expanded the results section and included one more data set of prostate segmentation from MRI volumes.
- We included the results on the whole data set of the BraTS 2018 which was only partially used before in the previous MICCAI version.
- We extended the related work section by including more recently proposed works.

2. Related works

The landscape of semi-supervised segmentation is, to a significant extent, influenced by the advancements in semi-supervised classification techniques, as delineated in Section 1.1. Among the various frameworks adopted, the mean-teacher based consistency regularisation paradigm is particularly prevalent in the field of semi-supervised medical image segmentation (Chen et al., 2020; Li et al., 2020a; Xu et al., 2020; Hang et al., 2020; Xie et al., 2020; Ta et al., 2020; To et al., 2020; Unnikrishnan et al., 2020; Yang et al., 2020; Fotedar et al., 2020). One early contribution to this vein of research was made by Yu et al. who enriched the mean-teacher model by incorporating uncertainty measures to generate a mask. This mask then modulates the application of consistency regularisation to only low-uncertainty regions (Yu et al., 2019).

Beyond perturbations at the data level, feature-level perturbations for consistency regularisation have also garnered considerable attention. For example, Luo et al. employed distinct initialisations for different decoders to induce feature perturbations (Luo et al., 2022). Xu introduced the idea by applying consistency regularisation on features after different morphological perturbations (Xu et al., 2022a). Another intriguing work employed a multi-decoder architecture, utilising three decoders with divergent up-sampling layers, to enable mutual consistency regularisation across the decoder outputs (Wu et al., 2022).

Recent studies have also explored the application of consistency regularisation to align signed distance maps of object boundaries, derived from different views of a common unlabelled input (You et al., 2022). The method is intended to enforce the models to be more of awareness of the object boundaries. Similar initiatives have leveraged uncertainty estimates, acquired via MC dropout, to weight the consistency regularisation (Zhang et al., 2023). While these methods achieved good performances on some of the tasks, it is pertinent to note that the necessity for at least two forward passes substantially elevates the computational overhead. In contrast, our method does not bring extra computational burden and it requires minimalist changes of the original backbone segmentation model.

Alternatively, more computationally affordable strategies have been pursued, notably leveraging pseudo labelling in the realm of medical image segmentation. Our proposed method also belongs to this paradigm. For instance, Bai et al. utilised conditional random fields to filter out false positives in pseudo labels (Bai et al., 2017). Wang employed uncertainty measures to refine pseudo labels (Wang et al., 2022). Wu et al. amalgamated pseudo labels with a dual-headed neural network architecture to instantiate a cross pseudo-supervision framework (Wu et al., 2021). Moreover, a recent study employed a variational auto-encoder as a student model to learn from pseudo labels generated by a deterministic teacher model (Wang and Lukasiewicz, 2022).

3. Pseudo labelling as expectation-maximisation

In this section, we reinterpret pseudo labelling in semi-supervised learning through the lens of the Expectation–Maximisation (EM) algorithm. We specifically focus on binary segmentation, as it is commonly encountered in medical imaging tasks where the objective is to differentiate foreground from background. This framework can be easily extended to multi-class segmentation by employing a multi-channel Sigmoid function. Each channel is treated as a binary output and combined using the argmax operation for the final prediction.

3.1. Problem formulation

Given a set of N total available training images as $X = \{x_n \in R^{HW} : n \in (1, 2, \dots, L, L+1, \dots, N)\}$, where $X_L = \{x_l \in R^{HW} : l \in (1, \dots, L)\}$ are L labelled images; $Y_L = \{y_l \in R^{HW} : l \in (1, \dots, L)\}$ are L labels for X_L; $X_U = \{x_u \in R^{HW} : u \in (L+1, \dots, N)\}$ is the rest of the U or $(N-L)$ unlabelled images. We have a segmentation network with parameters as θ and our final goal is to predict the labels $p(Y|X,\theta)$ of the whole data X with respect to θ.

3.2. Pseudo labels as latent variables

In order to find the optimal parameters of θ, the common approach is maximum likelihood estimation for maximising the likelihood of $P(X|\theta)$ with respect to θ, which contains two parts, namely supervised learning part and unsupervised learning part. The supervised learning part is to find the following joint data density with known full information of the labels:

$$p(X_L, Y_L|\theta) \tag{1}$$

The unsupervised learning part is to find the beneath likelihood with the same parameters without full information of the data:

$$p(X_U|\theta) \tag{2}$$

Since labels are not observable for X_U, we can treat this as a missing data problem and introduce latent variables Y_U'. We therefore transform the above Eq. (2) to an estimation of the following marginal likelihood:

$$p(X_U|\theta) = \int p(X_U, Y_U'|\theta)dY_U' \tag{3}$$

The latent variable in the above Eq. (3) can be implemented as the pseudo labels. Eq. (3) also shows that it is not an easy task to train a model in semi-supervised fashion, because it is difficult to simultaneously estimate the optimal values of θ and Y_U'. To address this difficult learning problem, we can decompose this problem by iteratively estimating the latent variables Y_U' and the model θ. We now notice that this can be solved by a typical Expectation-Maximisation (EM) (Bishop, 2006) algorithm. By plugging the Jensen's inequality, one can iteratively refine the Evidence Lower Bound of the log likelihood of the data in Eq. (3) (see details in later Section 3.4).

3.3. E-M pseudo labelling

We now display each component of the pseudo labelling in the sense of EM algorithm in the following paragraphs.

E-step At the nth iteration, the E-step estimates the values of the latent variable with the model (θ^{n-1}) from the last iteration ($n-1$). According to the cluster assumption that similar data points are supposed to have similar labels (Cahpelle et al., 2006), the E-step runs the inference on unlabelled data and generate pseudo-labels according to its maximum predicted probability. In practice, in binary segmentation, the pseudo-labels for the foreground class 1 are picked using a fixed threshold value (T) between 0 and 1. Normally, this threshold is set up as 0.5. This binarization is actually equivalent to the plug-in principle (Grandvalet and Bengio, 2004), which is a common approach for estimating the posterior probability using an empirical estimation in statistics. Therefore, the pseudo-labelling itself is the E-step:

$$y_u^{hw'} = \mathbb{1}(\theta^{n-1}(x_u^{hw}) > T = 0.5) \tag{4}$$

The above Eq. (4) is pseudo-labelling at pixel-wise. Where h and w are the index for the height and the index for the width of the pixel location respectively, for each unlabelled image x_u. $y_u^{hw'}$ is the pixel-wise pseudo label. More details of the connection between E-step and pseudo labelling is in the later section Section 3.4 on the convergence of pseudo labelling.

M-step At the M-step of iteration n, we will update the model parameters θ^{n-1} using the estimated latent variables (pseudo-labels Y_U') from the E-step. The images X are ignored for simplicity in the following expression:

$$\theta^n := \underset{\theta}{\mathrm{argmax}}\, p(\theta^n|\theta^{n-1}, Y_n') \tag{5}$$

The above Eq. (5) is normally solved by setting the partial derivatives of the sum of the $p(Y_n')$ with respect to θ as zero, which can be calculated with modern automatic differentiation based deep learning toolbox such as Pytorch (Paszke et al., 2019). In practice, we optimise θ in Eq. (5) via stochastic gradient descent. To use the stochastic gradient descent, we need to define an objective function and we use the common Dice loss ($f_{dice}(.)$) (Milletari et al., 2016) as this is a segmentation task:

$$f_{dice}(a, b) = \frac{2 * a * b + \epsilon}{a + b + \epsilon} \tag{6}$$

where a is the prediction, b is the ground truth and ϵ is to prevent the division of zero.

Loss function of SegPL We weight Eq. (5) with a hyper-parameter α. For the whole data set including both unlabelled and labelled data, we can extend the Eqs. (5) and (4) to a combination between the supervised learning part L_L and the unsupervised learning part L_U:

$$\mathcal{L}_{SegPL} = \alpha \underbrace{\frac{1}{N-L}\sum_{u=L+1}^{N} f_{dice}(\theta^{n-1}(x_u), \mathbb{1}(\theta^{n-1}(x_u) > T = 0.5))}_{\mathcal{L}_U} + \underbrace{\frac{1}{L}\sum_{l=1}^{L} f_{dice}(\theta^{n-1}(x_l), y_l)}_{\mathcal{L}_L} \tag{7}$$

The above loss function (7) is the key component of our first proposed semi-supervised segmentation method, omitting pixels' locations for simplicity, which is referred as SegPL (Segmentation with Pseudo Labels) in the paper. $\mathcal{L}_L$ works to prevent the networks falling into trivial solutions, trivial solutions happen when networks constantly predict one single class for all of the pixels.

3.4. On the convergence of pseudo labelling from the perspective of EM

In this section, we explain how semi-supervised learning with pseudo-labelling will always converge. We first define an objective function, the value of which we aim to increase. In our case, it would be the log data likelihood $logp(X_U)$. We also need to introduce a surrogate function $q(Y_U')$ which is any arbitrary distribution over the latent variable Y_U'. We follow Bishop (2006) and display the lower bound of the log data likelihood in the form of the Free Energy:

$$\begin{aligned} logp(X_U) &:= log \int p(X_U, Y_U'|\theta)dY_U' \\ &\geq \int q(Y_U') log \frac{p(X_U, Y_U'|\theta)}{q(Y_U')} dY_U' := \mathcal{F}(q(Y_U'), \theta) \end{aligned} \tag{8}$$

The functional free energy can be transformed back to the log data likelihood (Bishop, 2006):

$$\mathcal{F}(q(Y_U'), \theta) = logp(X_U) - KL[q(Y_U') \parallel p(Y_U'|X, \theta)] \tag{9}$$

In the E-step of the iteration n, the free energy is:

$$\mathcal{F}(q(Y_U'), \theta^{n-1}) = logp(X_U) - KL[q(Y_U') \parallel p(Y_U'|X, \theta^{n-1})] \tag{10}$$

As KL can never be negative, the above Eq. (10) has an upper bound. In order to reach that upper bound of the free energy at the nth iteration, we need to minimise $KL[q(Y_U') \parallel p(Y_U'|X, \theta^{n-1})]$. The KL distance has its minimum value at zero only if $q(Y_U')$ is equal to $p(Y_U'|X, \theta^{n-1})$. Therefore, we can simply replace the arbitrary function of latent variable $q(Y_U')$ as the current estimated posterior of the latent variable:

$$q(Y_U') = p(Y_U'|X_U, \theta^{n-1}) \tag{11}$$

The above Eq. (11) can be implemented as pseudo-labelling in Eq. (4). In other words, pseudo-labelling essentially maximises the free energy of the log data likelihood in the E-step.

Intuitively, the subsequent M-step is applying supervised learning to optimise the model parameters with pseudo labels which are produced from the precursory E-step. As supervised learning can be seen as maximum likelihood estimation, thereby, M-step increases the log data likelihood. In more details, we know that the log data likelihood in the M-step after updating the model θ is:

$$logp(X_U) = \mathcal{F}(q(Y_U'), \theta^{n-1}) + KL[p(Y_U'|X_U, \theta^{n-1}) \parallel p(Y_U'|X_U, \theta^n)] \tag{12}$$

The KL term in the above Eq. (12) becomes positive as the posterior of the latent variable Y_U' is different from its previous value. Together, it is easy to tell that the M-step increases the data log likelihood by at least the increased amount of the lower bound.

Up to this point, it is clear to see that, pseudo labelling (E-step) combined with supervised optimisation of model parameters (M-step) can never decrease the log likelihood of the data, leading to guaranteed convergence towards local optima. Similar conclusion was reported in the original EM paper (Dempster et al., 1977).

4. Generalisation of pseudo labels via variational inference for segmentation

In the last Section 3, we use an empirical estimation of the posterior of the latent variables (pseudo labels) by setting the T as 0.5. The fixed empirical estimation of T could be sub-optimal especially in the early stage of training when the networks do not have good representations and the predictions are not very confident (Rizve et al., 2021). Potentially, noisy training with some "bad" pseudo labels could accumulate some errors into the learnt representations. To address this potential issue, we provide an alternative approach to learn to approximate the true posterior of the pseudo labels. This alternative approach can be seen as a generalisation of the empirical estimation approach in SegPL in Section 3.

4.1. Confidence threshold as latent variable

In the last Section 3, we directly treat pseudo labels as latent variables. However, in segmentation task, the pseudo labels are pixel-wise, making the generative task a difficult one. To address this, we now introduce a simplification of the graphical model of the pseudo-labelling in 3. The key of this simplification is to treat the threshold value T as the latent variable for instead:

$$p(X_U|\theta) = \int p(X_U, T|\theta)dT \tag{13}$$

This new latent variable T makes the computation of the posterior much easier. We know that T is a value between 0 and 1, so that we have a clear prior knowledge of the range of this single value T. That any distribution describing values between 0 and 1 can be used as a prior distribution to approximate the real distribution of T. The true posterior of the latent variable T is:

$$p(T|X_U, \theta) = \frac{p(X_U|T, \theta)p(T)}{p(X_U|\theta)} \tag{14}$$

The new E-step at iteration n with threshold as the latent variable now becomes:

$$p(T_n = i|X_U, \theta^{n-1}) = \frac{\prod_{u=L+1}^{N} p(x_u|\theta^{n-1}, T_n = i)p(T_n = i)}{\sum_{j\in[0,1]} \prod_{u=L+1}^{N} p(x_u|\theta^{n-1}, T_n = j)p(T_n = j)} \tag{15}$$

From the above Eq. (15), one can tell that the empirical estimation of the threshold T is actually necessary although not optimal. Because there are infinite possible values between 0 and 1 in the denominator in Eq. (15), the posterior of the pseudo-labels is still intractable.

4.2. Variational E-step

To address the aforementioned intractable issue in Eq. (15), we use variational inference for the approximation of $p(T)$. As mentioned before, the prior of T can be an arbitrary distribution describing values between 0 and 1. For the implementation simplicity, we adapt an univariate Normal distribution for the prior distribution and we denote the prior distribution of T as a surrogate distribution $q(\beta)$. We use extra model parameters ϕ to parameterise the log variance and the mean of the approximated posterior distribution of T, conditioning on the image features, see the beneath Eq. (17). ϕ is implemented as a average pooling layer followed by a single 3 × 3 convolutional block including ReLU and normalisation layer, then two 1 × 1 convolutional layers for μ and $Log(\sigma^2)$ respectively. Alternatively, a simple fully connected layer can also be used as ϕ, we found no performance differences among different choices of architectures for ϕ.

$$(\mu, Log(\sigma^2)) = \phi(\theta(X_U)) \tag{16}$$

$$p(T|X_U, \theta, \phi) \approx \mathcal{N}(\mu, \sigma) \tag{17}$$

Differing from the fixed T in E-step in Eq. (4), the T in variational E-step is dynamic, we denote the stochastic threshold as $\mathbf{T}$ for clarity. We use the standard reparameterisation trick (Kingma and Welling, 2014) to generate the threshold in each iteration:

$$\begin{aligned} &\mathbf{T} = \mu + rand * e^{0.5*log(\sigma^2)} \\ &rand \sim \mathcal{N}(0, 1) \end{aligned} \tag{18}$$

As demonstrated in previous Eq. (8) that the log data likelihood term has an Evidence Lower Bound (ELBO) which contains a conditional probability of the data given latent variable and a KL distance between the posterior and the prior of the latent variable. We therefore write down variational unsupervised learning objective as:

$$\begin{aligned} &Log(P(X_U) \geq \\ &\sum_{u=L+1}^{N} \mathbb{E}_{T\sim P(\mathbf{T})}[Log(P(x_u|\mathbf{T}))] - KL(p(\mathbf{T})||q(\beta)) \end{aligned} \tag{19}$$

Loss function of BPL The new learning objective $P(X, \mathbf{T}, \theta)$ over the whole data set has a supervised learning $P(X_L, \mathbf{T}, \theta)$ which has not changed from Eq. (7), and an unsupervised learning part $P(X_U, \mathbf{T}, \theta)$ from the above Eq. (19). The final loss function is an ELBO over the whole data set:

$$\begin{aligned} \mathcal{L}_{SegPL}^{VI} = &\underbrace{\frac{1}{L}\sum_{l=1}^{L} f_{dice}(\theta^{n-1}(x_l), y_l)}_{\mathcal{L}_L} + \\ &\underbrace{\alpha\frac{1}{N-L}\sum_{u=L+1}^{N} f_{dice}(\theta^{n-1}(x_u), \mathbb{1}(\theta^{n-1}(x_u) > \mathbf{T}))}_{\mathcal{L}_U} + \\ &\underbrace{Log(\sigma_\beta) - Log(\sigma) + \frac{\sigma^2 + (\mu - \mu_\beta)^2}{2 * (\sigma_\beta)^2} - 0.5}_{\mathcal{L}_{KL}:\ KL(p(T)||q(\beta)),\ \beta\sim\mathcal{N}(\mu_\beta, \sigma_\beta)} \end{aligned} \tag{20}$$

where $\mathbf{T}$ can be found in Eq. (18). Different data sets might need different priors for the best empirical performances. Although we suggest to use higher mean such as 0.9 as a starting point. A schematic illustration of the implementation of Bayesian Pseudo Labels is shown in Fig. 2.

5. Experimental results

5.1. Data sets

The classification of pulmonary arteries and veins (CARVE) We use CARVE for demonstration of 3D binary segmentation of lung vessel of CT images. The CARVE data set (Charbonnier et al., 2016) comprises 10 fully annotated non-contrast low-dose thoracic CT scans. Each case has between 399 and 498 images, acquired at various spatial resolutions ranging from (282 × 426) to (302 × 474). We randomly select 1 case for labelled training, 2 cases for unlabelled training, 1 case for validation and the remaining 5 cases for testing. All image and label volumes were cropped to 176 × 176 × 3. To test the influence of the number of labelled training data, we prepared four sets of labelled training volumes with differing numbers of labelled volumes at: 2, 5, 10, 20. Normalisation was performed at case wise. Data curation resulted in 479 volumes for testing, which is equivalent to 1437 images. No data augmentation is used.

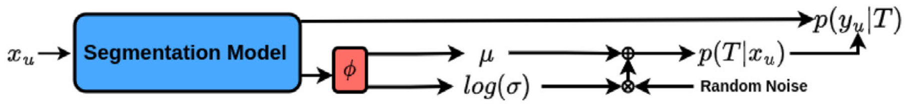

Fig. 2. The implementation of the proposed Bayesian Pseudo Labels. Only unsupervised learning part is illustrated.

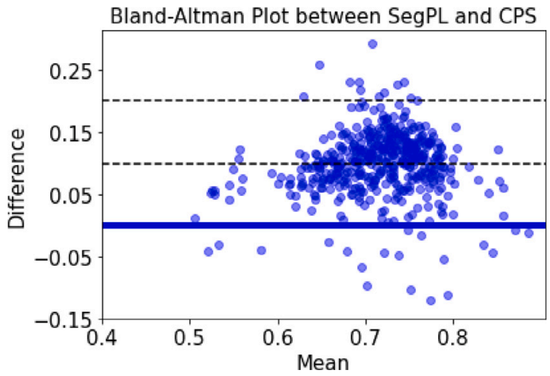

Fig. 3. SegPL statistically outperforms the best performing baseline CPS when trained on 2 labelled volumes from the CARVE data set. Each data point represents a single testing image.

BRATS 2018 We use BRATS 2018 (Menze et al., 2015; Bakas et al., 2018) for demonstration of 2D multi-class segmentation of brain tumour of MRI images. The BRATS 2018 comprises 210 high-grade glioma and 76 low-grade glioma MRI cases. Each case contains 155 slices. We focus on multi-class segmentation of sub-regions of tumours in high grade gliomas (HGG). All slices were centre-cropped to 176 × 176. We prepared three different sets of 2D slices for labelled training data: 50 slices from one case, 150 slices from one case and 300 slices from two cases. We use another 2 cases for unlabelled training data and 1 case for validation. 50 HGG cases were randomly sampled for testing. Case-wise normalisation was performed and all modalities were concatenated. A total of 3433 images were included for testing. No data augmentation is used.

Task01 Brain Tumour We use Task01 Brain Tumour from Medical Segmentation Decathlon consortium (Antonelli et al., 2022) as a demonstration of 3D binary segmentation of brain tumour of MRI images. The Task01 Brain Tumour is based on BRATS 2017 with different naming format from BRATS 2018. This data set was not in our previous MICCAI version but we included this data set here because it is easy to download and use for the readers for the future follow-up works. Each case in The Task01 Brain Tumour has 155 slices with 240 × 240 spatial dimension. We merge all of the tumour classes into one tumour class for simplicity. We do not apply centre cropping in the pre-processing here. In the training, we randomly crop volumes on the fly with size of 64 × 64 × 64. We separate the original training cases as labelled training data and testing data. We use the original testing cases as unlabelled data. For the labelled training data, we use 8 cases with index number from 1 to 8. We have 476 cases for testing and 266 cases for unlabelled training data. We apply normalisation with statistics of intensities across the whole training data set. We keep all of the MRI modalities as 4 channel input.

Task05 Prostate We also use Task05 Prostate from Medical Segmentation Decathlon consortium (Antonelli et al., 2022) as another demonstration of 3D binary segmentation of prostate of MRI images. Each case in the Task05 Prostate is a 4D volume: 2 modalities, 15 slices with 320 × 320 spatial dimension. We divide the original training cases into three parts, 1 case as labelled training data, 16 cases as unlabelled training data and the rest 14 cases as unseen testing data. During training, we randomly crop volumes on the fly with a target size of 192 × 192 × 8.

Table 1
Hyper-parameters used across experiments. Different data might need different α. LR: learning rate. Ratio: Unlabelled/labelled in each batch.

Data	Batch size	L.R.	Steps	α	Ratio.
BRATS(2D)	2	0.03	200	0.05	5
CARVE(3D)	2	0.01	800	1.0	4
Task01(3D)	1	0.0004	25 000	0.1	2
Task05(3D)	1	0.001	2000	0.002	4

Table 2
Different prior values of T used across experiments.

Data	Mean	Std.
BRATS(2D)	0.5	0.1
CARVE(3D)	0.4	0.1
Task01(3D)	0.9	0.1
Task05(3D)	0.9	0.1

5.2. Baselines

Our baselines include both supervised and semi-supervised learning methods. We use U-net (Ronneberger et al., 2015) in SegPL as an example of segmentation network. Partly due to computational constraints, for 3D experiments we used a 3D U-net with 8 channels in the first encoder such that unlabelled data can be included in the same batch. For 2D experiments, we used a 2D U-net with 16 channels in the first encoder. The first baseline utilises supervised training on the backbone and is trained with labelled data denoted as "Sup". We compared SegPL with state-of-the-art consistency based methods: (1) "cross pseudo supervision" or CPS (Chen et al., 2021), which is considered the current state-of-the-art for semi-supervised segmentation; (2) another recent state-of-the-art model "cross consistency training" (Ouali et al., 2020), denoted as "CCT", due to hardware restriction, our implementation shares most of the decoders apart from the last convolutional block; (3) a classic model called "FixMatch" (FM) (Sohn et al., 2020). To adapt FixMatch for a segmentation task, we added Gaussian noise as weak augmentation and "RandomAug" (Cubuk et al., 2020) for strong augmentation; (4) "self-loop (Li et al., 2020b)", which solves a self-supervised jigsaw problem as pre-training and combines with pseudo-labelling.

5.3. Training

We use Adam optimiser (Kingma and Ba, 2015) with default settings. Our code is implemented using Pytorch 1.0 (Paszke et al., 2019) and released in https://github.com/moucheng2017/EMSSL. We trained all of the experiments with a TITAN V GPU with 12 GB memory. The training hyperparameters are included in Table 1. The prior values used are presented in Table 2.

5.4. Segmentation performances

The segmentation performances of CARVE 2014, BRATS 2018, Task 01 can be found in Tables 3–5, respectively. As reflected in the quantitative results in tables, pseudo labelling based SegPL consistently achieves better results than the baselines of semi-supervised and supervised methods. Especially, as shown in Fig. 3 of the Bland–Altman plot between the best performing baseline CPS and our SegPL on CARVE when only 2 labelled volumes are used for training, SegPL statistically

Table 3
Our model vs. Baselines on a binary vessel segmentation task on 3D CT images of the CARVE data set. Metric is Intersection over Union (IoU (↑) in %). Avg performance of 5 training. blue: 2nd best. red: best.

Data	Supervised	Semi-supervised				
Labelled Volumes	3D U-net Ronneberger et al. (2015)	FixMatch Sohn et al. (2020)	CCT Ouali et al. (2020)	CPS Chen et al. (2021)	SegPL (Ours, 2022)	SegPL+VI (Ours, 2022)
2	56.79 ± 6.44	62.35 ± 7.87	51.71 ± 7.31	66.67 ± 8.16	**69.44 ± 6.38**	**70.65 ± 6.33**
5	58.28 ± 8.85	60.80 ± 5.74	55.32 ± 9.05	70.61 ± 7.09	**76.52 ± 9.20**	**73.33 ± 8.61**
10	67.93 ± 6.19	72.10 ± 8.45	66.94 ± 12.22	75.19 ± 7.72	**79.51 ± 8.14**	**79.73 ± 7.24**
20	81.40 ± 7.45	80.68 ± 7.36	80.58 ± 7.31	81.65 ± 7.51	**83.08 ± 7.57**	**83.41 ± 7.14**
Computational need						
Train (s)	1014	2674	4129	2730	1601	1715
Flops	6.22	12.44	8.3	12.44	6.22	6.23
Para (K)	626.74	626.74	646.74	1253.48	626.74	630.0

Table 4
Our model vs. Baselines on multi-class tumour segmentation on 2D MRI images of BRATS 2018. Metric is Intersection over Union (IoU (↑) in %). Avg performance of 5 runs. blue: 2nd best. red: best.

Data	Supervised	Semi-supervised				
Labelled Slices	2D U-net Ronneberger et al. (2015)	Self-Loop Li et al. (2020b)	FixMatch Sohn et al. (2020)	CPS Chen et al. (2021)	SegPL (Ours, 2022)	SegPL+VI (Ours, 2022)
50	54.08 ± 10.65	65.91 ± 10.17	67.35 ± 9.68	63.89 ± 11.54	**70.60 ± 12.57**	**71.20 ± 12.77**
150	64.24 ± 8.31	68.45 ± 11.82	69.54 ± 12.89	69.69 ± 6.22	**71.35 ± 9.38**	**72.93 ± 12.97**
300	67.49 ± 11.40	70.80 ± 11.97	70.84 ± 9.37	71.24 ± 10.80	**72.60 ± 10.78**	**75.12 ± 13.31**

Table 5
Our model vs. Supervised baseline on 3D binary tumour segmentation of Task 01 Brain Tumour (BRATS 2017). Metric is Intersection over Union (IoU (↑) in %). Avg performance of models between iteration 20 000 and 25 000 with 1000 as the interval. The shape of each training sample is 96^3. red: best.

Testing size	32 × 32 × 32	64 × 64 × 64	96 × 96 × 96
Supervised	61.07 ± 7.93	66.94 ± 12.4	70.13 ± 13.22
SegPL-VI	64.44 ± 8.3	71.43 ± 11.91	73.07 ± 11.71

Table 6
Our model vs. Supervised baseline on 3D binary segmentation of prostate Task 05 from Medical Decathlon. Metric is Intersection over Union (IoU (↑) in %). Performance of the models which achieved the highest training accuracy. The shape of each training sample is 192 × 192 × 8. red: best.

Testing size	192 × 192 × 8	160 × 160 × 8	128 × 128 × 8
Supervised	67.68 ± 10.06	61.39 ± 11.86	60.53 ± 9.94
SegPL-VI	70.15 ± 10.59	63.15 ± 11.62	61.06 ± 10.39

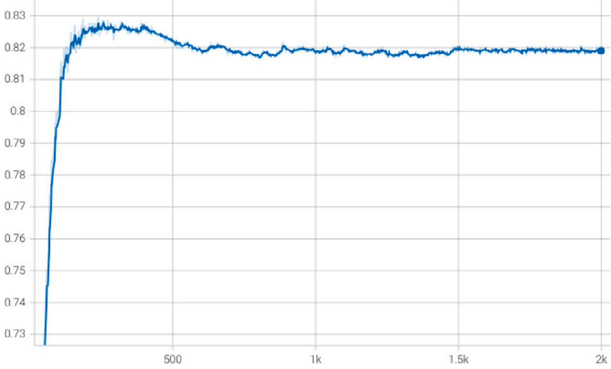

Fig. 4. Y-axis: Learnt threshold in the experiment of Task01 Brain Tumour. X-axis: training iterations. The mean of the prior is 0.9 and the std of the prior is 0.1. The learnt threshold converged around 0.82 after 2000 iterations.

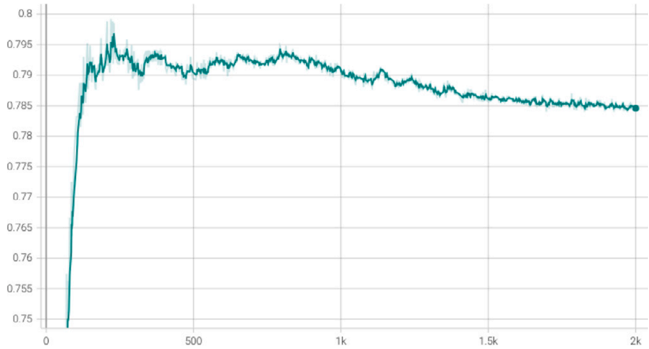

Fig. 5. Y-axis: Learnt threshold in the experiment of Task05 Prostate. X-axis: training iterations. The mean of the prior is 0.9 and the std of the prior is 0.1. The learnt threshold converged around 0.785 after 2000 iterations.

outperforms the best baseline. We further confirm the statistical difference by performing Mann Whitney test on the same results on 2 labelled volumes and we found the p-vale less than 1e–4. By extending the SegPL with variational inference to SegPL-VI, we found further improvements on segmentation on most of the experiments. Interestingly, the improvements brought by SegPL-VI is more obvious on multi-class experiments on BRATS 2018. As the outputs on BRATS are multi-channel but SegPL-VI learns one threshold across all of the channel, we suspect that might bring in strong regularisation effect which results in noticeable improvements. We also noticed that SegPL-VI could fail to learn optimal threshold sometimes as the result of SegPL-VI on CARVE with 5 labelled volumes are inferior to the corresponding result of SegPL. We expect that more hyper-parameter searching could improve the performance of SegPL.

As shown in the qualitative results in Fig. 6 of CARVE, SegPL successfully learnt better decision boundary than other baselines that SegPL can partially separate the foreground lung vessels from the background whereas most of the other methods classifies everything as background. However, SegPL seemed to have overconfident predictions on the edges of the foreground that it has a lot of false positive results. Similarly in BRATS, SegPL detected one more class of brain tumour (blue) than the other baselines in Fig. 7. However, none of the methods including SegPL can detect the most rare green class of tumour.

One phenomenon worthy mentioning is shown in Table 5 on 3D binary segmentation of whole tumour and Table 6 on 3D binary segmentation of prostate. During training on whole brain tumour segmentation, we use random cropping with fixed size at 64 × 64 × 64 to compensate with the memory of GPU. On testing data, we examined the models with different sizes of cropped volumes at 32^3, 64^3, 96^3 and

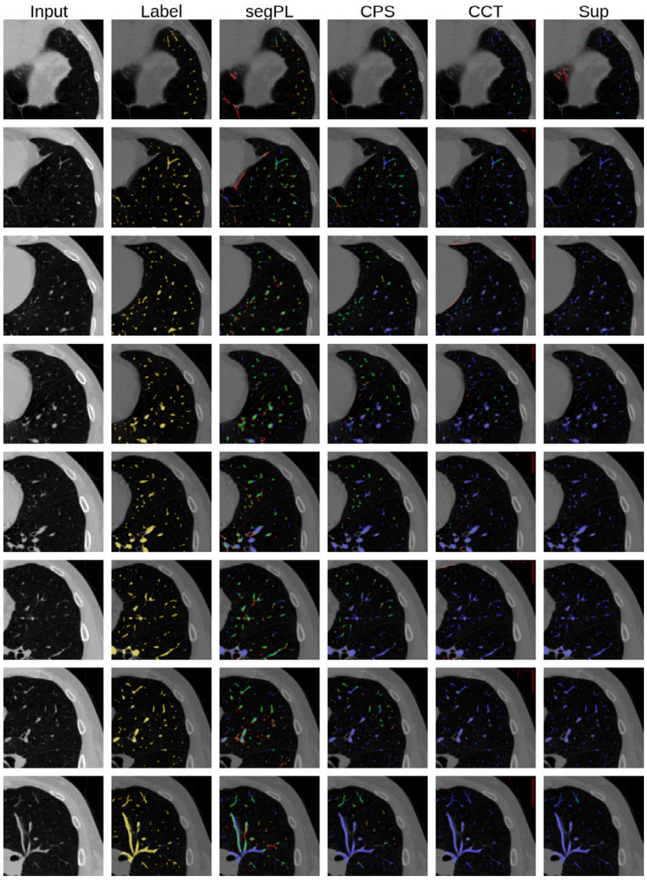

Fig. 6. Visual results. CARVE trained with 5 labelled volumes. Red: false positive. Green: true positive. Blue: false negative. Yellow: ground truth. GT: Ground truth. CPS: cross pseudo labels (CVPR 2021). CCT: cross consistency training (CVPR 2020). Sup: supervised training.

128^3. The models actually generalise well on the scales that they have not seen during the training. In fact, larger cropped volumes result in better results. The results in Table 6 on segmentation of prostate also confirms this phenomenon.

Although SegPL achieves higher segmentation accuracy, SegPL enjoys a low computational burden. As illustrated in the computational need section in Table 3, SegPL has the least computational burden among all of the tested semi-supervised learning baselines. Especially in terms of FLOPs, SegPL is very close to supervised learning methods. This shows that our model has the scaling potential for large models and large data sets.

5.5. Sensitivity studies of hyper-parameters

We performed brief sensitivity studies on hyper-parameters on BRATS with 150 labelled slices. As shown in Fig. 8,(a) shows that SegPL is very sensitive to learning rate that it should be at least 0.01. We found that other baselines also needed large learning rate. Fig. 8.(b) shows the impact of warm-up schedule of α from 0 to final α value. x axis is the length of linear warming-up of α in terms of whole steps. It appears that SegPL is not sensitive to warm-up schedule of α. Fig. 8.(c) illustrates the effect of the ratio between unlabelled images to labelled images in each batch. The suitable range of unlabelled/labelled ratio is quite wide and between 1 to 10. Fig. 8.(d) shows that the pseudo supervision cannot be too strong. This confirms the suggestions from the original pseudo labelling paper that pseudo supervision should not dominate the training.

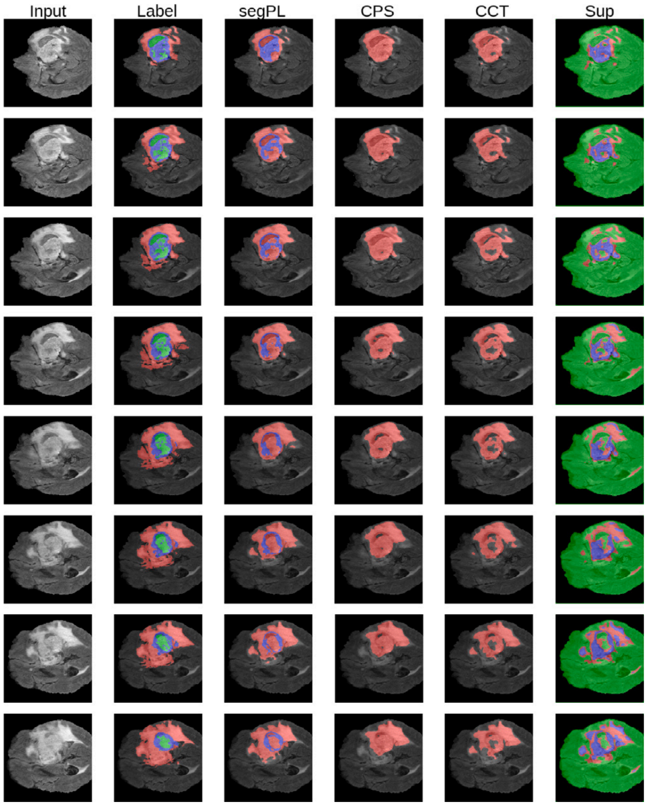

Fig. 7. Visual results. BRATS 2018 trained with 300 labelled slices. Red: whole tumour. Green: tumour core. Blue: enhancing tumour core. GT: Ground truth. CPS: cross pseudo labels (CVPR 2021). CCT: cross consistency training (CVPR 2020). Sup: supervised training.

5.6. Robustness

In medical imaging, models often face challenges due to out-of-distribution (OOD) noise such as variations in scan acquisition parameters or differing patient populations. These factors can significantly degrade model performance in real-world applications. To evaluate the robustness of our proposed SegPL model against OOD noise, we conduct experiments using models trained on the CARVE data set.

We simulate OOD noises with unseen random contrast and Gaussian noise, we then apply mix-up (Zhang et al., 2018) to create new testing samples by adding the OOD noises on original images. Specifically, for a given original testing image x_t, we applied random contrast and noise augmentation on x_t to derive OOD samples x_t'. We arrived at the testing sample $(\hat{x}_t)$ via $\gamma x_t' + (1-\gamma)x_t$. As shown in Fig. 9, as testing difficulty increases, the performances across all baselines drop exponentially. SegPL outperformed all of the baselines across all of the tested experimental settings. The findings suggest that SegPL is more robust when testing on OOD samples and achieves better generalisation performance against that from the baselines.

In the context of privacy and security, especially as federated learning across hospitals gains popularity, the robustness against adversarial attacks becomes crucial. We assess SegPL's resilience to such attacks using the fast gradient sign method (FGSM) (Kurakin et al., 2017). FGSM perturbs an image by computing the gradient of the loss function with respect to the input image and adding a noise term proportional to the sign of the gradient.

Our experiments show that the performance of all models, including SegPL, declines as the strength of the adversarial attack (measured by Epsilon) increases. However, SegPL exhibits a smaller drop in performance compared to baseline models, as illustrated in Fig. 10. These results further substantiate the robustness of SegPL under various challenging conditions.

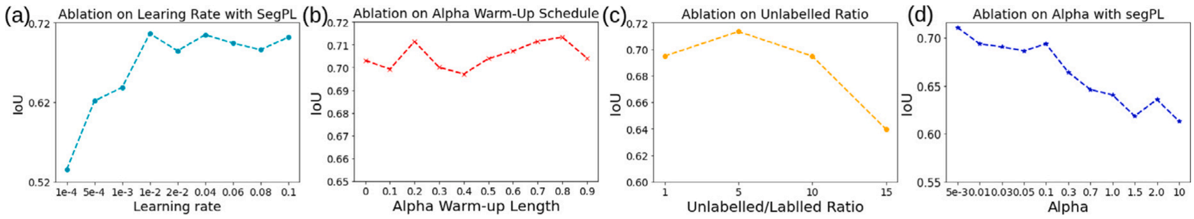

Fig. 8. Sensitivity studies of hyper-parameters on BRATS with 150 labelled slices.

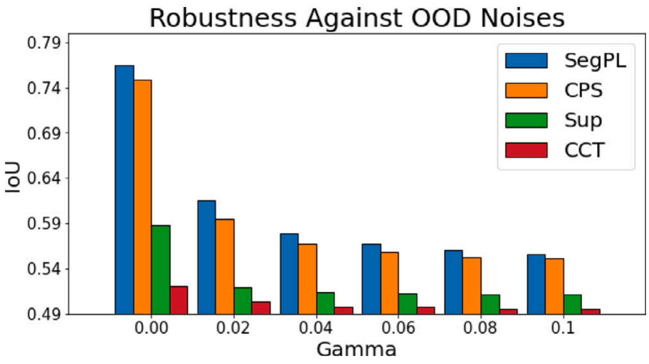

Fig. 9. Robustness against out-of-distribution noise. Gamma is the strength of the out-of-distribution noises. Using 2 labelled volumes from CARVE.

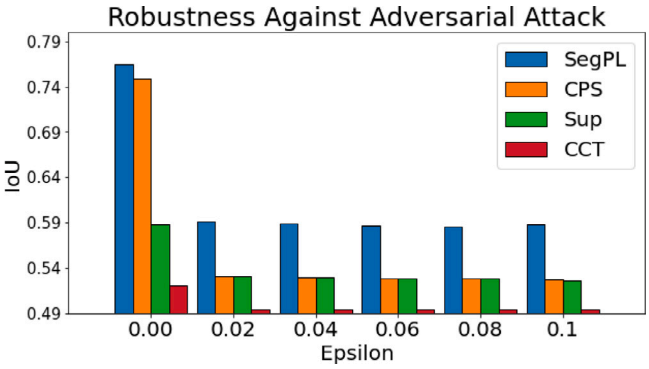

Fig. 10. Robustness against adversarial attack. Epsilon is the strength of the FGSM (Kurakin et al., 2017) attack. Using 2 labelled volumes from CARVE.

5.7. Uncertainty

Since SegPL-VI is trained with stochastic threshold for unlabelled data therefore not suffering from posterior collapse. Consequently, SegPL can generate plausible segmentation during inference using stochastic thresholds. To test the performance of SegPL-VI on uncertainty quantification, we use random latent variable values (threshold) with 5 Monte Carlo samples. We focus experimenting on models trained with 5 labelled volumes of CARVE data set. For comparison, we adopt Deep Ensemble, as it is the gold-standard baseline for uncertainty estimation (Lakshminarayanan et al., 2017; Snoek et al., 2019). Both the tested methods Deep Ensemble and SegPL-VI achieved the same Brier score at 0.97. This result shows that SegPL-VI has the potential to become a benchmark method for uncertainty quantification. The Brier score is calculated using beneath equation, where, y_{ij} is the ground truth label at pixel at location i, j, y_{ij} is 1 for foreground pixel and y_{ij} is 0 for background pixel. p_{ij} is the predicted probability of the pixel being the foreground pixel.

$$Brier = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}(p_{ij} - y_{ij})^2 \tag{21}$$

6. Limitations and future works

There are two main limitations of the proposed Bayesian pseudo labels. The first limitation is that once the model starts to over-fit, the model becomes overconfident that it predicts with very high confidence, while the learnt threshold also converges to a value such as shown in Figs. 4 and 5. In this situation, if the prior of the mean is too low, then the learnt threshold will not be able to mask out the bad over confident pseudo labels. Thus calibration becomes very important here. In future work, one could extend the formulation of pseudo labels to take into account of calibration.

The second limitation is the use of the prior in the current paper. We use Gaussian due to its simplicity and easy to implement. However, Gaussian prior might not be the most optimal one here. Future work can explore the impact of other priors of learnt threshold. Candidate prior distributions include categorical and Beta distributions.

In terms of implementation, the current Bayesian Pseudo Labels only learns a single threshold for all of the images in the same batch size. However, more adaptive implementations could potentially boost the performance of Bayesian Pseudo Labels, such as learning one threshold for each image or even each pixel. If the thresholds are learnt per pixel of an image, one might also need to consider the spatial correlations among the thresholds.

Theoretically, another interesting future work can be studying the impact of labelled data in terms of preventing collapsed representations. Other future work can also look into the convergence property of SegPL-VI.

The feasibility of the applications of the proposed methods on other tasks such as uncertainty quantification, classification and registration also remain unexplored.

In the future pipeline for learning with limited annotations, we also expect to exploit SegPL-VI's full potential by combining with large-scale pre-training techniques.

7. Conclusions

In this paper, we revisit pseudo-labelling and provide an interpretation of its empirical success by formulating the pseudo-labelling process as the EM algorithm. We as well unravel its full formulation along with a learning based approach to approximate it. Empirically, we examined that the original pseudo-labelling (Lee, 2013) and its Bayesian generalisation on semi-supervised medical image segmentation and we report that pseudo-labelling as a competitive and robust baseline.

CRediT authorship contribution statement

Moucheng Xu: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Yukun Zhou:** Investigation, Writing – original draft, Writing – review & editing, Conceptualization. **Chen Jin:** Conceptualization, Writing – original draft, Writing – review & editing. **Marius de Groot:** Funding acquisition, Supervision, Writing – original draft, Writing – review & editing. **Daniel C. Alexander:** Conceptualization, Supervision, Writing – original draft, Writing – review & editing, Funding acquisition. **Neil P. Oxtoby:** Writing – original draft, Writing – review & editing, Supervision. **Yipeng Hu:** Conceptualization, Writing – original draft, Writing – review & editing, Supervision. **Joseph Jacob:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Mou-Cheng Xu reports financial support was provided by University College London. Mou-Cheng Xu reports financial support was provided by GSK. Neil P. Oxtoby reports financial support was provided by UK Research and Innovation. Daniel C. Alexander reports financial support was provided by Engineering and Physical Sciences Research Council. Joseph Jacob reports financial support was provided by Wellcome Trust. Joseph Jacob reports financial support was provided by NIHR University College London Hospitals Biomedical Research Centre. Daniel C. Alexander reports financial support was provided by Wellcome Trust. Daniel C. Alexander reports was provided by NIHR University College London Hospitals Biomedical Research Centre. Neil P. Oxtoby reports financial support was provided by NIHR University College London Hospitals Biomedical Research Centre. Marius de Groot reports financial support was provided by GSK.

Data availability

Data will be made available on request.

Acknowledgements

MCX was supported by GSK, United Kingdom (BIDS3000034123) and UCL Engineering Dean's Prize. NPO is supported by a UKRI Future Leaders Fellowship (MR/S03546X/1). DCA is supported by UK EPSRC grants M020533, R006032, R014019, V034537, Wellcome Trust, United Kingdom UNS113739, Wellcome Trust 221915/Z/20/Z. JJ is supported by Wellcome Trust Clinical Research Career Development Fellowship 209,553/Z/17/Z. NPO, DCA, and JJ are supported by the NIHR UCLH Biomedical Research Centre, UK. This research was funded in whole or in part by the Wellcome Trust [209553/Z/17/Z]. For the purpose of open access, the author has applied a CC-BY public copyright licence to any author accepted manuscript version arising from this submission.

References

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Schneider, A., Landman, B., Ligjens, G., Menze, B., Ronneberger, O., SUmmers, R., Ginneken, B., Bilello, M., Bilic, P., Christ, P., Do, R., Gollub, M., Heckers, S.H., Huisman, H., Jarnagin, W.R., McHugo, M.K., Napel, S., Pernicka, J.S.G., Rhode, K., Tobon-Gomez, C., Vorontsov, E., Meakin, J.A., Ourselin, S., Wiesenfarth, M., Arbeláez, P., Bae, B., Chen, S., Daza, L., Feng, J., He, B., Isensee, F., Ji, Y., Jia, F., Kim, I., Maier-Hein, K., Merhof, D., Pai, A., Park, B., Perslev, M., Rezaiifar, R., Rippel, O., Sarasua, I., Shen, W., Son, J., Wachinger, C., Wang, L., Wang, Y., Xia, Y., Xu, D., Xu, Z., Zheng, Y., Simpson, A.L., Maier-Hein, L., Cardoso, M.J., 2022. The medical segmentation decathlon. Nature Commun. 13, 4128. http://dx.doi.org/10.1038/s41467-022-30695-9.

Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K., 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: 2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020. IEEE, pp. 1–8. http://dx.doi.org/10.1109/IJCNN48605.2020.9207304.

Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A.P., Matthews, P.M., Rueckert, D., 2017. Semi-supervised learning for network-based cardiac MR image segmentation. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2017 - 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II. In: Lecture Notes in Computer Science, vol. 10434, Springer, pp. 253–260. http://dx.doi.org/10.1007/978-3-319-66185-8_29.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., Prastawa, M., Alberts, E., Lipková, J., Freymann, J.B., Kirby, J.S., Bilello, M., Fathallah-Shaykh, H.M., Wiest, R., Kirschke, J., Wiestler, B., Colen, R.R., Kotrotsou, A., LaMontagne, P., Marcus, D.S., Milchenko, M., Nazeri, A., Weber, M., Mahajan, A., Baid, U., Kwon, D., Agarwal, M., Alam, M., Albiol, A., Albiol, A., Varghese, A., Tuan, T.A., Arbel, T., Avery, A., B., P., Banerjee, S., Batchelder, T., Batmanghelich, K.N., Battistella, E., Bendszus, M., Benson, E., Bernal, J., Biros, G., Cabezas, M., Chandra, S., Chang, Y., et al., 2018. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. CoRR abs/1811.02629. URL: http://arxiv.org/abs/1811.02629. arXiv:1811.02629.

Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C., Sohn, K., 2020. Remixmatch: semi-supervised learning with distribution alignment and augmentation anchoring. In: International Conference on Learning Representation. ICLR, URL: https://openreview.net/forum?id=HklkeR4KPB.

Berthelot, D., Carlini, N., Goodfellow, I., Oliver, A., Papernot, N., Raffel, C., 2019. MixMatch: A holistic approach to semi-supervised learning. In: Neural Information Processing Systems. NeurIPS, Vol. 32, URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/1cd138d0499a68f4bb72bee04bbec2d7-Paper.pdf.

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

Bridle, J.S., Anthony Heading, D.J.M., 1991. Unsupervised classifiers, mutual information and 'phantom targets'. In: Advances in Neural Information Processing Systems. NeurIPS, Vol. 4, URL: https://proceedings.neurips.cc/paper_files/paper/1991/file/a8abb4bb284b5b27aa7cb790dc20f80b-Paper.pdf.

Cahpelle, O., Scholkopf, B., Zien, A., 2006. Semi-supervised learning. The MIT Press, URL: http://dblp.uni-trier.de/db/books/collections/CSZ2006.html.

Charbonnier, J., Brink, M., Ciompi, F., Scholten, E.T., Schaefer-Prokop, C., van Rikxoort, E.M., 2016. Automatic pulmonary artery-vein separation and classification in computed tomography using tree partitioning and peripheral vessel matching. IEEE Trans. Med. Imaging 35 (3), 882–892. http://dx.doi.org/10.1109/TMI.2015.2500279.

Chen, C., Chen, Q., Huaqi, Q., Cheng, O., Shuo, W., Liang, C., Giacomo, T., Wenjia, B., Daniel, R., 2020. Realistic adversarial data augmentation for MR image segmentation. In: Medical Image Computing and Computer Assisted Intervention. MICCAI 2020, Springer, pp. 667–677.

Chen, X., Yuan, Y., Zeng, G., Wang, J., 2021. Semi-supervised semantic segmentation with cross pseudo supervision. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, June 19-25, 2021. Computer Vision Foundation / IEEE, pp. 2613–2622. http://dx.doi.org/10.1109/CVPR46437.2021.00264, URL: https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Semi-Supervised_Semantic_Segmentation_With_Cross_Pseudo_Supervision_CVPR_2021_paper.html.

Cubuk, E.D., Zoph, B., Shlens, J., Le, Q., 2020. RandAugment: Practical automated data augmentation with a reduced search space. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, Virtual. URL: https://proceedings.neurips.cc/paper/2020/hash/d85b63ef0ccb114d0a3bb7b7d808028f-Abstract.html.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Stat. Methodol. 39, 1–38, URL: http://web.mit.edu/6.435/www/Dempster77.pdf.

Fotedar, G., Tajbakhsh, N., Ananth, S.P., Ding, X., 2020. Extreme consistency: Overcoming annotation scarcity and domain shifts. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part I. In: Lecture Notes in Computer Science, vol. 12261, Springer, pp. 699–709. http://dx.doi.org/10.1007/978-3-030-59710-8_68.

French, G., Laine, S., Aila, T., Mackiewicz, M., Finlayson, G.D., 2020. Semi-supervised semantic segmentation needs strong, varied perturbations. In: 31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020. BMVA Press, URL: https://www.bmvc2020-conference.com/assets/papers/0680.pdf.

Grandvalet, Y., Bengio, Y., 2004. Semi-supervised learning by entropy minimization. In: Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]. pp. 529–536, URL: https://proceedings.neurips.cc/paper/2004/hash/96f2b50b5d3613adf9c27049b2a888c7-Abstract.html.

Hang, W., Feng, W., Liang, S., Yu, L., Wang, Q., Choi, K., Qin, J., 2020. Local and global structure-aware entropy regularized mean teacher model for 3D left atrium segmentation. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part I. In: Lecture Notes in Computer Science, vol. 12261, Springer, pp. 562–571. http://dx.doi.org/10.1007/978-3-030-59710-8_55.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. URL: http://arxiv.org/abs/1412.6980.

Kingma, D.P., Welling, M., 2014. Auto-encoding variational Bayes. In: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings. URL: http://arxiv.org/abs/1312.6114.

Kurakin, A., Goodfellow, I.J., Bengio, S., 2017. Adversarial examples in the physical world. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings. OpenReview.net, URL: https://openreview.net/forum?id=HJGU3Rodl.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 6402–6413, URL: https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Lee, D.-H., 2013. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. In: ICML Workshop on Challenges in Representation Learning. URL: https://www.researchgate.net/publication/280581078_Pseudo-Label_The_Simple_and_Efficient_Semi-Supervised_Learning_Method_for_Deep_Neural_Networks.

Li, Y., Chen, J., Xie, X., Ma, K., Zheng, Y., 2020b. Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part I. In: Lecture Notes in Computer Science, vol. 12261, Springer, pp. 614–623. http://dx.doi.org/10.1007/978-3-030-59710-8_60.

Li, X., Jiang, M., Zhang, X., Kamp, M., Dou, Q., 2021. FedBN: Federated learning on non-IID features via local batch normalization. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, URL: https://openreview.net/forum?id=6YEQUn0QICG.

Li, K., Wang, S., Yu, L., Heng, P., 2020a. Dual-teacher: Integrating intra-domain and inter-domain teachers for annotation-efficient cardiac segmentation. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part I. In: Lecture Notes in Computer Science, vol. 12261, Springer, pp. 418–427. http://dx.doi.org/10.1007/978-3-030-59710-8_41.

Li, X., Yu, L., Chen, H., Fu, C., Heng, P., 2018. Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model. In: British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018. BMVA Press, p. 63, URL: http://bmvc2018.org/contents/papers/0162.pdf.

Luo, X., Hu, M., Song, T., Wang, G., Zhang, S., 2022. Semi-supervised medical image segmentation via cross teaching between CNN and transformer. In: International Conference on Medical Imaging with Deep Learning, MIDL 2022, 6-8 July 2022, Zurich, Switzerland. In: Proceedings of Machine Learning Research, vol. 172, PMLR, pp. 820–833, URL: https://proceedings.mlr.press/v172/luo22b.html.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J.S., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E.R., Weber, M., Arbel, T., Avants, B.B., Ayache, N., Buendia, P., Collins, D.L., Cordier, N., Corso, J.J., Criminisi, A., Das, T., Delingette, H., Demiralp, Ç., 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.T., Sarikaya, D., Schwartz, L.H., Shin, H., Shotton, J., Silva, C.A., Sousa, N.J., Subbanna, N.K., Székely, G., Taylor, T.J., Thomas, O.M., Tustison, N.J., Ünal, G.B., Vasseur, F., Wintermark, M., Ye, D.H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., Leemput, K.V., 2015. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging 34 (10), 1993–2024. http://dx.doi.org/10.1109/TMI.2014.2377694.

Milletari, F., Navab, N., Ahmadi, S., 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, October 25-28, 2016. IEEE Computer Society, pp. 565–571. http://dx.doi.org/10.1109/3DV.2016.79.

Ouali, Y., Hudelot, C., Tami, M., 2020. Semi-supervised semantic segmentation with cross-consistency training. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. Computer Vision Foundation / IEEE, pp. 12671–12681. http://dx.doi.org/10.1109/CVPR42600.2020.01269, URL: https://openaccess.thecvf.com/content_CVPR_2020/html/Ouali_Semi-Supervised_Semantic_Segmentation_With_Cross-Consistency_Training_CVPR_2020_paper.html.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E.Z., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. pp. 8024–8035, URL: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Pham, H., Dai, Z., Xie, Q., Le, Q.V., 2021. Meta pseudo labels. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, June 19-25, 2021. Computer Vision Foundation / IEEE, pp. 11557–11568. http://dx.doi.org/10.1109/CVPR46437.2021.01139, URL: https://openaccess.thecvf.com/content/CVPR2021/html/Pham_Meta_Pseudo_Labels_CVPR_2021_paper.html.

Rizve, M.N., Duarte, K., Rawat, Y.S., Shah, M., 2021. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=-ODN6SbiUU.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III. In: Lecture Notes in Computer Science, vol. 9351, Springer, pp. 234–241. http://dx.doi.org/10.1007/978-3-319-24574-4_28.

Snoek, J., Ovadia, Y., Fertig, E., Lakshminarayanan, B., Nowozin, S., Sculley, D., Dillon, J.V., Ren, J., Nado, Z., 2019. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. pp. 13969–13980, URL: https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C., Cubuk, E.D., Kurakin, A., Li, C., 2020. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, Virtual. URL: https://proceedings.neurips.cc/paper/2020/hash/06964dce9addb1c5cb5d6e3d9838f733-Abstract.html.

Ta, K., Ahn, S.S., Stendahl, J.C., Sinusas, A.J., Duncan, J.S., 2020. A semi-supervised joint network for simultaneous left ventricular motion tracking and segmentation in 4D echocardiography. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part VI. In: Lecture Notes in Computer Science, vol. 12266, Springer, pp. 468–477. http://dx.doi.org/10.1007/978-3-030-59725-2_45.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 1195–1204, URL: https://proceedings.neurips.cc/paper/2017/hash/68053af2923e00204c3ca7c6a3150cf7-Abstract.html.

To, M.N.N., Sankineni, S., Xu, S., Turkbey, B., Pinto, P.A., Moreno, V., Merino, M., Wood, B.J., Kwak, J.T., 2020. Improving dense pixelwise prediction of epithelial density using unsupervised data augmentation for consistency regularization. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part I. In: Lecture Notes in Computer Science, vol. 12261, Springer, pp. 572–581. http://dx.doi.org/10.1007/978-3-030-59710-8_56.

Unnikrishnan, B., Nguyen, C.M., Balaram, S., Foo, C.S., Krishnaswamy, P., 2020. Semi-supervised classification of diagnostic radiographs with noteacher: A teacher that is not mean. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part I. In: Lecture Notes in Computer Science, vol. 12261, Springer, pp. 624–634. http://dx.doi.org/10.1007/978-3-030-59710-8_61.

Wang, J., Lukasiewicz, T., 2022. Rethinking Bayesian deep learning methods for semi-supervised volumetric medical image segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, la, USA, June 18-24, 2022. IEEE, pp. 182–190. http://dx.doi.org/10.1109/CVPR52688.2022.00028.

Wang, G., Zhai, S., Lasio, G., Zhang, B., Yi, B., Chen, S., Macvittie, T.J., Metaxas, D.N., Zhou, J., Zhang, S., 2022. Semi-supervised segmentation of radiation-induced pulmonary fibrosis from lung CT scans with multi-scale guided dense attention. IEEE Trans. Medical Imaging 41 (3), 531–542. http://dx.doi.org/10.1109/TMI.2021.3117564.

Wu, Y., Ge, Z., Zhang, D., Xu, M., Zhang, L., Xia, Y., Cai, J., 2022. Mutual consistency learning for semi-supervised medical image segmentation. Med. Image Anal. 81, 102530. http://dx.doi.org/10.1016/j.media.2022.102530, URL: https://www.sciencedirect.com/science/article/pii/S1361841522001773.

Wu, Y., Xu, M., Ge, Z., Cai, J., Zhang, L., 2021. Semi-supervised left atrium segmentation with mutual consistency training. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part II. In: Lecture Notes in Computer Science, vol. 12902, Springer, pp. 297–306. http://dx.doi.org/10.1007/978-3-030-87196-3_28.

Xie, Y., Zhang, J., Liao, Z., Verjans, J., Shen, C., Xia, Y., 2020. Pairwise relation learning for semi-supervised gland segmentation. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part V. In: Lecture Notes in Computer Science, vol. 12265, Springer, pp. 417–427. http://dx.doi.org/10.1007/978-3-030-59722-1_40.

Xu, J., Lala, S., Gagoski, B.A., Turk, E.A., Grant, P.E., Golland, P., Adalsteinsson, E., 2020. Semi-supervised learning for fetal brain MRI quality assessment with ROI consistency. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part VI. In: Lecture Notes in Computer Science, vol. 12266, Springer, pp. 386–395. http://dx.doi.org/10.1007/978-3-030-59725-2_37.

Xu, M., Zhou, Y., Jin, C., Blumberg, S.B., Wilson, F.J., de Groot, M., Alexander, D.C., Oxtoby, N.P., Jacob, J., 2022a. Learning morphological feature perturbations for calibrated semi-supervised segmentation. In: International Conference on Medical Imaging with Deep Learning, MIDL 2022, 6-8 July 2022, Zurich, Switzerland. In:

Proceedings of Machine Learning Research, vol. 172, PMLR, pp. 1413–1429, URL: https://proceedings.mlr.press/v172/xu22a.html.

Xu, M., Zhou, Y., Jin, C., de Groot, M., Alexander, D.C., Oxtoby, N.P., Hu, Y., Jacob, J., 2022b. Bayesian pseudo labels: Expectation maximization for robust and efficient semi-supervised segmentation. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2022 - 25th International Conference, Singapore, September 18-22, 2022, Proceedings, Part V. In: Lecture Notes in Computer Science, vol. 13435, Springer, pp. 580–590. http://dx.doi.org/10.1007/978-3-031-16443-9_56.

Yang, H., Shan, C., Kolen, A.F., de With, P.H.N., 2020. Deep Q-network-driven catheter segmentation in 3D US by hybrid constrained semi-supervised learning and dual-unet. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part I. In: Lecture Notes in Computer Science, vol. 12261, Springer, pp. 646–655. http://dx.doi.org/10.1007/978-3-030-59710-8_63.

You, C., Zhou, Y., Zhao, R., Staib, L.H., Duncan, J.S., 2022. SimCVD: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. IEEE Trans. Med. Imaging 41 (9), 2228–2237. http://dx.doi.org/10.1109/TMI.2022.3161829.

Yu, L., Wang, S., Li, X., Fu, C., Heng, P., 2019. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part II. In: Lecture Notes in Computer Science, vol. 11765, Springer, pp. 605–613. http://dx.doi.org/10.1007/978-3-030-32245-8_67.

Zhang, H., Cissé, M., Dauphin, Y.N., Lopez-Paz, D., 2018. mixup: Beyond empirical risk minimization. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, URL: https://openreview.net/forum?id=r1Ddp1-Rb.

Zhang, Y., Jiao, R., Liao, Q., Li, D., Zhang, J., 2023. Uncertainty-guided mutual consistency learning for semi-supervised medical image segmentation. Artif. Intell. Med. 138, 102476. http://dx.doi.org/10.1016/J.ARTMED.2022.102476.

Zou, Y., Zhang, Z., Zhang, H., Li, C., Bian, X., Huang, J., Pfister, T., 2021. PseudoSeg: Designing pseudo labels for semantic segmentation. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, URL: https://openreview.net/forum?id=-TwO99rbVRu.