

Bayesian Pseudo Labels: Expectation Maximization for Robust and Efficient Semi-Supervised Segmentation

Mou-Cheng Xu¹, Yukun Zhou¹, Chen Jin¹, Marius de Groot², Daniel C. Alexander¹, Neil P. Oxtoby^{*1}, Yipeng Hu^{*1}, and Joseph Jacob^{*1}

¹ Centre for Medical Image Computing, University College London, UK

² GSK R & D, Stevenage, UK

³ xumoucheng28@gmail.com, code: <https://github.com/moucheng2017/EMSSL/>

Abstract. This paper concerns pseudo labelling in segmentation. Our contribution is fourfold. Firstly, we present a new formulation of pseudo-labelling as an Expectation-Maximization (EM) algorithm for clear statistical interpretation. Secondly, we propose a semi-supervised medical image segmentation method purely based on the original pseudo labelling, namely SegPL. We demonstrate SegPL is a competitive approach against state-of-the-art consistency regularisation based methods on semi-supervised segmentation on a 2D multi-class MRI brain tumour segmentation task and a 3D binary CT lung vessel segmentation task. The simplicity of SegPL allows less computational cost comparing to prior methods. Thirdly, we demonstrate that the effectiveness of SegPL may originate from its robustness against out-of-distribution noises and adversarial attacks. Lastly, under the EM framework, we introduce a probabilistic generalisation of SegPL via variational inference, which learns a dynamic threshold for pseudo labelling during the training. We show that SegPL with variational inference can perform uncertainty estimation on par with the gold-standard method Deep Ensemble.

Keywords: Semi-Supervised Segmentation · Pseudo Labels · Expectation-Maximization · Variational Inference · Uncertainty · Probabilistic Modelling · Out-Of-Distribution · Adversarial Robustness

1 Introduction

Image segmentation is a fundamental component of medical image analysis, essential for subsequent clinical tasks such as computer-aided-diagnosis and disease progression modelling. In contrast to natural images, the acquisition of pixel-wise labels for medical images requires input from clinical specialists making such labels expensive to obtain. Semi-supervised learning is a promising approach to address label scarcity by leveraging information contained within the additional unlabelled datasets. Consistency regularisation methods are the current state-of-the-art strategy for semi-supervised learning.

* Joint Senior Authorships.

Consistency regularisation methods [24,23,2,3,27] make the networks invariant to perturbations at the input level, the feature level [18,8] or the network architectural level [6]. These methods heavily rely on specifically designed perturbations at the input, feature or the network level, which lacks transferability across different tasks [21]. In contrast, pseudo-labelling is a simple and general approach which was proposed for semi-supervised image classification. Pseudo-labelling [14] generates pseudo-labels using a fixed threshold and uses pseudo-labels to supervise the training of unlabelled images. A recent work [21] argued that pseudo labelling on its own can achieve results comparable with consistency regularisation methods in image classification.

Recent works in semi-supervised segmentation with consistency regularisation are complex in their model design and training. For example, a recent work [18] proposed a model with multiple decoders, where the features in each decoder are also augmented differently, before the consistency regularisation is applied on predictions of different decoders. Another recent method [6] proposed an ensemble approach and applied the consistency regularisation between the predictions of different models. Other approaches include applying different data augmentation on the input in order to create two different predictions. One of the predictions is then used as a pseudo-label to supervise the other prediction [23]. The aforementioned related works are all tested as baselines in our experiments presented in Sect. 4.

In this paper, we propose a straightforward pseudo label-based algorithm called SegPL, and report its improved or comparable efficacy on two medical image segmentation applications following comparison with all tested alternatives. We highlight the simplicity of the proposed method, compared to existing approaches, which allows efficient model development and training. We also explore the benefits of pseudo labels at improving model generalisation with respect to out-of-distribution noise and model robustness against adversarial attacks in segmentation. Theoretically, we as well provide a new perspective of pseudo-labelling as an EM algorithm. Lastly, under the EM framework, we show that SegPL with variational inference can learn the threshold for pseudo-labelling leading to uncertainty quantification as well as Deep Ensemble [13]. A high level introduction of SegPL can be found in Fig.1.

2 Pseudo Labelling As Expectation-Maximization

We show that training with pseudo labels is an Expectation-Maximization (EM) algorithm using a binary segmentation case. Multi-class segmentation can be turned into a combination of multiple binary segmentation via one-hot encoding and applying Sigmoid on each class individually (multi-channel Sigmoid).

Let $X = \{(x_1, y_1), \dots, (x_L, y_L), x_{L+1}, \dots, x_N\}$ be the available training images, where $X_L = \{(x_l, y_l) : l \in (1, \dots, L)\}$ is a batch of L labelled images; $X_U = \{x_u : u \in (L+1, \dots, N)\}$ is an additional batch of unlabelled images. Our goal is to optimise a segmentation network (θ) so that we maximise the marginal likelihood of $P(X, \theta)$ w.r.t θ . Given that pseudo-labelling is a generative task,

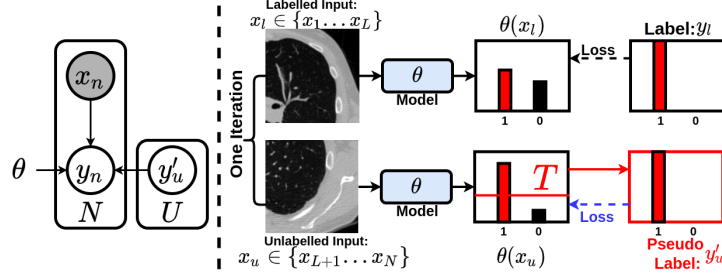


Fig. 1. Left: Graphical model of pseudo-labelling. Pseudo-label y'_n is generated using data x_u and model θ , therefore, pseudo-labelling can be seen as the E-step in Expectation-Maximization. The M-step updates θ using y'_n and data X . Right: Pseudo-labelling for binary segmentation. In our 1st implementation, namely SegPL, the threshold T is fixed for selecting the pseudo labels. In our 2nd implementation, namely SegPL-VI, the threshold T is dynamic and learnt via variational inference.

we can cast pseudo-labels ($Y'_U = \{y'_u \in (L + 1, ..N)\}$) as latent variables. Now our optimisation goal is to maximise the likelihood of $P(X, Y'_U, \theta)$. We can solve θ and Y'_U iteratively like the Expectation-Maximization (EM)[4] algorithm.

At the n^{th} iteration, the E-step generates the latent variable by estimating its posterior with the old model (θ^{n-1}). In the original pseudo-labelling [14], pseudo-labels are generated according to its maximum predicted probability. For example in the binary setting, the pseudo-labels are generated using a fixed threshold value (T) which is normally set up as 0.5. Pseudo-labelling is equivalent to applying the plug-in principle [9] to estimate the posterior of Y'_U using its empirical estimation. Hence, the pseudo-labelling is the E-step:

$$y'_u = \mathbb{1}(\theta^{n-1}(x_u) > T = 0.5) \quad (1)$$

At the M-step of iteration n , Y'_U is fixed and we update the old model θ^{n-1} using the estimated latent variables (pseudo-labels Y'_U) and the data X including both the labelled and unlabelled data:

$$\theta^n := \operatorname{argmax}_{\theta} p(\theta^n | X, \theta^{n-1}, Y'_n) \quad (2)$$

Eq.2 can be solved by stochastic gradient descent given an objective function. We use Dice loss function ($f_{dice}(\cdot)$) [17] as the objective function and we weight the unsupervised learning with a hyper-parameter α , the Eq.2 and Eq.1 can thereby be extended to a loss function over the whole dataset (supervised learning part L_L and unsupervised learning part L_U):

$$\mathcal{L}_{SegPL} = \underbrace{\alpha \frac{1}{N-L} \sum_{u=L+1}^N f_{dice}(\theta^{n-1}(x_u), y'_u)}_{\mathcal{L}_U} + \underbrace{\frac{1}{L} \sum_{l=1}^L f_{dice}(\theta^{n-1}(x_l), y_l)}_{\mathcal{L}_L} \quad (3)$$

3 Generalisation of Pseudo Labels via Variational Inference for Segmentation

We now derive a generalisation of SegPL under the same EM framework. As shown in the E-step in Eq.1, the pseudo-labelling depends on the threshold value T , so finding the pseudo-label y'_u is the same with finding the threshold value T . We therefore can recast the learning objective of EM as estimating the likelihood $P(X, T, \theta)$ w.r.t θ . The benefit of flexible T is that different stages of training might need different optimal T to avoid wrong pseudo labels which cause noisy training[21]. The E-step is now estimating the posterior of T . According to the Bayes' rule, the posterior of T is $p(T|X, \theta) \propto p(X|\theta, T)p(T)$. The E-step at iteration n is:

$$p(T_n = i|X, \theta^{n-1}) = \frac{\prod_{u=L+1}^N p(x_u|\theta^{n-1}, T_n = i)p(T_n = i)}{\sum_{j \in [0,1]} \prod_{u=L+1}^N p(x_u|\theta^{n-1}, T_n = j)p(T_n = j)} \quad (4)$$

As shown in Eq.4, the denominator is computationally intractable as there are infinite possible values between 0 and 1. This also confirms that the empirical estimation in Eq.1 is actually necessary although not optimal. We use variational inference to approximate the true posterior of $p(T|X)$. We use new parameters ϕ ⁴ to parametrize $p(T|X)$ given the image features, assuming that T is drawn from a Normal distribution between 0 and 1:

$$(\mu, \text{Log}(\sigma)) = \phi(\theta(X)) \quad (5)$$

$$p(T|X, \theta, \phi) \approx \mathcal{N}(\mu, \sigma) \quad (6)$$

Following [11], we introduce another surrogate prior distribution over T and denote this as $q(\beta)$, $\beta \sim \mathcal{N}(\mu_\beta, \sigma_\beta)$. The E-step now minimises $KL(p(T)||q(\beta))$. The M-step is the same with Eq.3 but with a learnt threshold T . The learning objective $P(X, T, \theta)$ has supervised learning $P(X_L, T, \theta)$ and unsupervised learning part $P(X_U, T, \theta)$. Although $P(X_L, T, \theta)$ is not changed, $P(X_U, T, \theta)$ now estimates the Evidence Lower Bound (ELBO):

$$\text{Log}(P(X_U, T, \theta, \phi)) \geq \sum_{u=L+1}^N \mathbb{E}_{T \sim P(T)} [\text{Log}(P(y_u|x_n, T, \theta, \phi))] - KL(p(T)||q(\beta)) \quad (7)$$

Where Log is taken for expression simplicity in Eq.7, y_n is the prediction of unlabelled image x_u . The final loss function is:

$$\mathcal{L}_{SegPL}^{VI} = \mathcal{L}_{SegPL} + \underbrace{\text{Log}(\sigma_\beta) - \text{Log}(\sigma) + \frac{\sigma^2 + (\mu - \mu_\beta)^2}{2 * (\sigma_\beta)^2} - 0.5}_{KL(p(T)||q(\beta)), \beta \sim \mathcal{N}(\mu_\beta, \sigma_\beta)} \quad (8)$$

Where $\mathcal{L}_{SegPL}$ can be found in Eq.3 but with a learnt threshold T . Different data sets need different priors. Priors are chosen empirically for different data sets. $\beta \sim \mathcal{N}(0.5, 0.1)$ for BRATS and $\beta \sim \mathcal{N}(0.4, 0.1)$ for CARVE.

⁴ ϕ : Average pooling layer followed by one 3 x 3 convolutional block including ReLU and normalisation layer, then two 1 x 1 convolutional layers for μ and $\text{Log}(\sigma)$

4 Experiments & Results

Segmentation Baselines ⁵ We compare SegPL to both supervised and semi-supervised learning methods. We use U-net [22] 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 [6], which is considered the current state-of-the-art for semi-supervised segmentation; 2) another recent state-of-the-art model “cross consistency training” [18], 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) [23]. To adapt FixMatch for a segmentation task, we added Gaussian noise as weak augmentation and “RandomAug” [7] for strong augmentation; 4) “self-loop [15]”, which solves a self-supervised jigsaw problem as pre-training and combines with pseudo-labelling.

CARVE 2014 The classification of pulmonary arteries and veins (CARVE) dataset [5] 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 x 426) to (302 x 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 $176 \times 176 \times 3$. 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.

BRATS 2018 BRATS 2018 [16] 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.

4.1 SegPL Outperforms Baselines with Less Computational Resource and Less Training Time

As shown in Tab.2, Tab.1 and Fig.3, SegPL consistently outperformed the state-of-the-art semi-supervised segmentation methods and the supervised counterpart

⁵ Training: Adam optimiser[10]. Our code is implemented using Pytorch 1.0[20], trained with a TITAN V GPU. See Appendix for details of hyper-parameters.

Table 1. Our model vs Baselines on a binary vessel segmentation task on 3D CT images of the CARVE dataset. Metric is Intersection over Union (IoU ($\uparrow$) in %) . All 3D networks have 8 channels in the first encoder. Avg performance of 5 training.

Data	Supervised	Semi-Supervised				
Labelled Volumes	3D U-net [22](2015)	FixMatch [23](2020)	CCT [18](2020)	CPS [6](2021)	SegPL (Ours, 2022)	SegPL+VI (Ours, 2022)
2	56.79 $\pm$ 6.44	62.35 $\pm$ 7.87	51.71 $\pm$ 7.31	66.67 $\pm$ 8.16	69.44$\pm$6.38	70.65$\pm$6.33
5	58.28 $\pm$ 8.85	60.80 $\pm$ 5.74	55.32 $\pm$ 9.05	70.61 $\pm$ 7.09	76.52$\pm$9.20	73.33$\pm$8.61
10	67.93 $\pm$ 6.19	72.10 $\pm$ 8.45	66.94 $\pm$ 12.22	75.19 $\pm$ 7.72	79.51$\pm$8.14	79.73$\pm$7.24
20	81.40 $\pm$ 7.45	80.68 $\pm$ 7.36	80.58 $\pm$ 7.31	81.65 $\pm$ 7.51	83.08$\pm$7.57	83.41$\pm$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

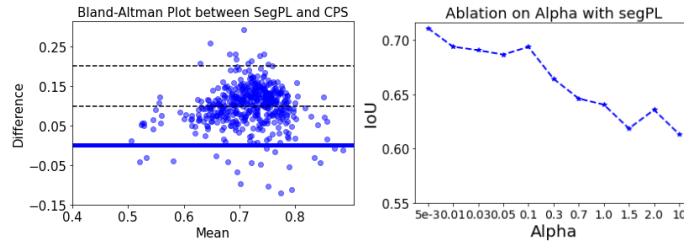


Fig. 2. Left: SegPL statistically outperforms the best performing baseline CPS when trained on 2 labelled volumes from the CARVE dataset. Each data point represents a single testing image. Right: Ablation study on the α which is the weight on the unsupervised learning part. Different data might need different α .

for different labelled data regimes, on both 2D multi-class and 3D binary segmentation tasks. The Bland-Altman plot in Fig.2.Left shows SegPL performs significantly better than the best performed baseline CPS in an extremely low data regime (2 labelled volume on CARVE). We also applied Mann Whitney test on results when trained with 2 labelled volume and CARVE and the p-value was less than $1e-4$. As shown in bottom three rows in Tab.1, SegPL also is the most close one to the supervised learning counterpart in terms of computational cost, among all of the tested methods. As shown in Tab.2⁶, SegPL-VI further improves segmentation performance on the MRI based brain tumour segmentation task. It also appears that the variational inference is sensitive to the data because we didn’t observe obvious performance gain with SegPL-VI on CARVE. The performance of SegPL-VI on CARVE could be further improved with more hyper-parameter searching.

⁶ blue: 2nd best. red: best

Table 2. Our model vs Baselines on multi-class tumour segmentation on 2D MRI images of BRATS. Metric is Intersection over Union (IoU ($\uparrow$) in %). All 2D networks have 16 channels in the first encoder. Avg performance of 5 runs.

Data	Supervised	Semi-Supervised				
Labelled Slices	2D U-net [22](2015)	Self-Loop [15](2020)	FixMatch [23](2020)	CPS [6](2021)	SegPL (Ours, 2022)	SegPL+VI (Ours, 2022)
50	54.08 $\pm$ 10.65	65.91 $\pm$ 10.17	67.35 $\pm$ 9.68	63.89 $\pm$ 11.54	70.60$\pm$12.57	71.20$\pm$12.77
150	64.24 $\pm$ 8.31	68.45 $\pm$ 11.82	69.54 $\pm$ 12.89	69.69 $\pm$ 6.22	71.35$\pm$9.38	72.93$\pm$12.97
300	67.49 $\pm$ 11.40	70.80 $\pm$ 11.97	70.84 $\pm$ 9.37	71.24 $\pm$ 10.80	72.60$\pm$10.78	75.12$\pm$13.31

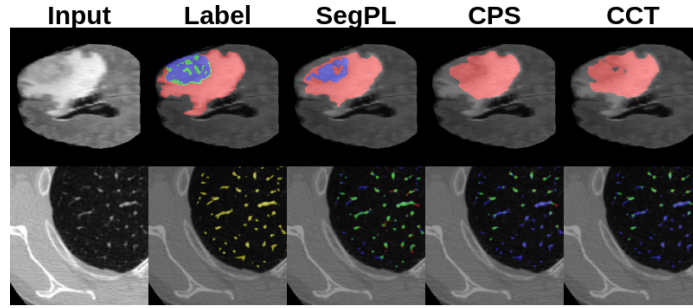


Fig. 3. Row1: BRATS with 300 labelled slices. Red: whole tumour. Green: tumour core. Blue: enhancing tumour core. Row2: CARVE with 5 labelled volumes. Red: false positive. Green: true positive. Blue: false negative. Yellow: ground truth.

4.2 Ablation Studies

Ablation studies are performed using BRATS with 150 labelled slices. The suitable range for α is $5e - 3$ to 0.1 as shown in Fig.2.Right. The other ablation studies can be found in the Appendix. We found that SegPL needs large learning rate from at least 0.01. We noticed that SegPL is not sensitive to the warm-up schedule of α . The appropriate range of the ratio between unlabelled images to labelled images in each batch is between 2 to 10. Different data sets might need different ranges of α for optimal performances.

4.3 Better Generalisation on Out-Of-Distribution (OOD) Samples and Better Robustness against Adversarial Attack

We mimic OOD image noise resulting from variations in scan acquisition parameters and patient characteristics that are characteristic of medical imaging. We simulated OOD samples with unseen random contrast and Gaussian noise on CARVE and mix-up [28] to create new testing samples. 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.4, as testing difficulty increases, the performances across all baselines drop exponentially. SegPL outperformed all of

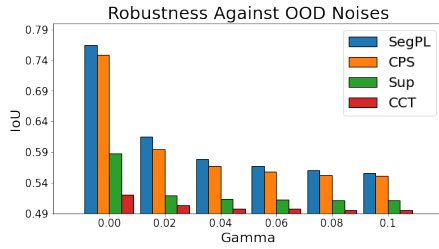


Fig. 4. Robustness against out-of-distribution noise. Gamma is the strength of the out-of-distribution noises. Using 2 labelled volumes from CARVE.

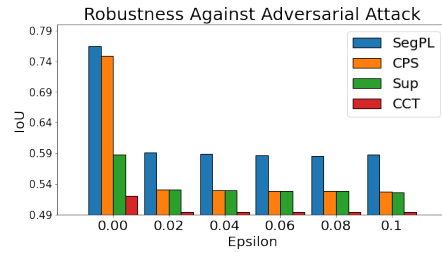


Fig. 5. Robustness against adversarial attack. Epsilon is the strength of the FGSM[12] attack. Using 2 labelled volumes from CARVE.

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. As robustness has become important in privacy-preserving collaborative learning among hospitals, we also examined SegPL’s robustness against adversarial attack using fast gradient sign method (FGSM)[12]. With increasing strength of adversarial attack (Epsilon), all the networks suffered performance drop. Yet, as shown in Fig.5, SegPL suffered much less than the baselines.

4.4 Uncertainty Estimation with SegPL-VI

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. We tested the uncertainty estimation of SegPL-VI with the common metric Brier score on models trained with 5 labelled volumes of CARVE dataset. We compared SegPL-VI with Deep Ensemble, which is regarded as a gold-standard baseline for uncertainty estimation [13][19]. Both SegPL-VI and Deep Ensemble achieved the same Brier score at 0.97, using 5 Monte Carlo samples.

5 Related Work

Although we are the first to solely use the original pseudo labelling in semi-supervised segmentation of medical images, a few prior works explored the use of pseudo labels in other forms. Bai [1] post-processed pseudo-labels via conditional random field. Wu [26] proposed sharpened cycled pseudo labels along with a two-headed network. Wang [25] post-processed pseudo-labels with uncertainty-aware refinements for their attention based models.

6 Conclusion

We verified that the original pseudo-labelling [14] is a competitive and robust baseline in semi-supervised medical image segmentation. We further interpret pseudo-labelling with EM and explore the potential improvement with variational inference following generalised EM. The most prominent future work is to improve the quality of the pseudo labels by exploring different priors (e.g. Beta, Categorical) or auto-search of hyper parameters. Other future works include investigating the convergence property of SegPL-VI and improving SegPL-VI at uncertainty quantification. Future studies can also investigate from the perspective of model calibration [27] to understand why pseudo labels work. SegPL and SegPL-VI can also be used in other applications such as classification or registration.

7 Acknowledgements

We thank the anonymous reviewer 3 for the suggestion of using Beta prior and an interesting discussion about posterior collapse issue in uncertainty quantification. MCX is supported by GlaxoSmithKline (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 UNS113739. 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.

References

1. Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P.M., Rueckert, D.: Semi-supervised learning for network-based cardiac mr image segmentation. MICCAI (2017)
2. Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C., Sohn, K.: Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. International Conference On Learning Representation (ICLR) (2020)
3. Berthelot, D., Carlini, N., Goodfellow, I., Oliver, A., Papernot, N., Raffel, C.: Mixmatch: A holistic approach to semi-supervised learning. Neural Information Processing Systems (NeurIPS) (2019)
4. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg (2006)
5. Charbonnier, J.P., Brink, M., Ciompi, F., Scholten, E.T., Schaefer-Prokop, C.M., van Rikxoort, E.M.: Automatic pulmonary artery-vein separation and classification in computed tomography using tree partitioning and peripheral vessel matching. IEEE Transaction on Medical Imaging (2015)
6. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. Computer Vision and Pattern Recognition (CVPR) (2021)

7. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. *Neural Information Processing Systems (NeurIPS)* (2020)
8. French, G., Laine, S., Aila, T., Mackiewicz, M., aFinalyson, G.: Semi-supervised semantic segmentation needs strong, varied perturbations. *British Machine Vision Conference (BMVC)* (2020)
9. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. *Neural Information Processing Systems (NeurIPS)* (2004)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *International Conference on Learning Representation (ICLR)* (2015)
11. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. *ICLR* (2014)
12. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial examples in the physical world. *International Conference on Learning Representation Workshop* (2017)
13. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. *Neural Information Processing System (NeurIPS)* (2017)
14. Lee, D.H.: Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML workshop on Challenges in Representation Learning* (2013)
15. Li, Y., Chen, J., Xie, X., Ma, K., Zheng, Y.: Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation. *MICCAI* (2020)
16. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., et al: The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transaction on Medical Imaging* (2015)
17. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. *International Conference on 3D Vision (3DV)* (2016)
18. Ouali, Y., Hudelot, C., Tami, M.: Semi-supervised semantic segmentation with cross-consistency training. *Computer Vision and Pattern Recognition (CVPR)* (2020)
19. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J.V., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems (NeurIPS)* (2019)
20. 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.: Pytorch: An imperative style, high-performance deep learning library. *Neural Information Processing System (NeurIPS)* (2019)
21. Rizve, M.N., Duarte, K., Rawat, Y.S., Shah, M.: In defense of pseudo-labelling: An uncertainty-aware pseudo-label selective framework for semi-supervised learning. *International Conference on Learning Representation (ICLR)* (2021)
22. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentations. *International conference on medical image computing and computer assisted intervention (MICCAI)* (2015)
23. Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Neural Information Processing Systems (NeurIPS)* (2020)
24. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Neural Information Processing Systems (NeurIPS)* (2017)

25. Wang, G., Zhai, S., Lasio, G., Zhang, B., Yi, B., Chen, S., Macvittie, T.J., Metaxas, D., Zhou, J., Zhang, S.: Semi-supervised segmentation of radiationinduced pulmonary fibrosis from lung ct scans with multi-scale guided dense attention. *IEEE Transactions on Medical Imaging* (2014)
26. Wu, Y., Xu, M., Ge, Z., Cai, J., Zhang, L.: Semi-supervised left atrium segmentation with mutual consistency training. *MICCAI* (2021)
27. Xu, M.C., Zhou, Y.K., Jin, C., Wilson, F., Blumberg, S., de Groot, M., Alexander, D.C., Oxtoby, N.P., Jacob, J.: Learning morphological feature perturbations for calibrated semi supervised segmentation. *International Conference on Medical Imaging with Deep Learning (MIDL)* (2022)
28. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. *International Conference on Learning Representation (ICLR)* (2018)

Extended Appendix for Bayesian Pseudo Labels

Mou-Cheng Xu¹, Yu-Kun Zhou¹, Chen Jin¹, Marius de Groot², Daniel C. Alexander¹, Neil P. Oxtoby^{*1}, Yipeng Hu^{*1}, and Joseph Jacob^{*1}

¹ Centre for Medical Image Computing, University College London, UK

² GSK R & D, Stevenage, UK

1 Training Hyper-Parameters

Table 1. Hyper-parameters used across experiments. Different data might need different α . 3D models need longer time to converge.

Data	Batch Size	Learning rate	Steps	α	Unlabelled/labelled
BRATS	2	0.03	200	0.05	5
CARVE	2	0.01	800	1.0	4

2 Results of Ablation Studies

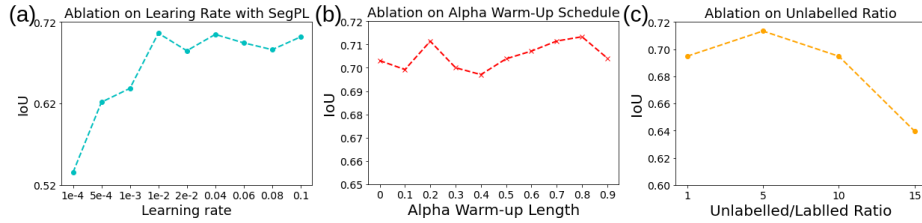


Fig. 1. Ablation studies are performed on hyper-parameters trained on BRATS with 150 labelled slices. a) Ablation on learning rate. Learning rate should be at least 0.01. We found that other baselines also needed large learning rate. b) Ablation on warm-up schedule of α from 0 to final α value. x axis 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 . c) Ablation on 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.

* Joint Senior Authorships.

3 Brier Score

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

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.

4 Differences between SegPL and Other methods

SegPL vs CPS CPS uses two models and cross-pseudo-supervision, as illustrated in Fig2, in which the first pseudo label is from the second model and the second pseudo label is from the first model. SegPL uses a single model that requires a single forward pass.

SegPL vs FixMatch The main difference between SegPL and FixMatch relates to the augmentation of the input images. In FixMatch (see Fig.2), the unlabelled image is randomly augmented twice, once weakly and once strongly. Each augmented image requires one forward pass and the pseudo label of FixMatch is generated from the weakly augmented image. SegPL however does not use data augmentation. $Pseudo\ Label = \mathbb{1}(Model(x) > T = 0.5)$

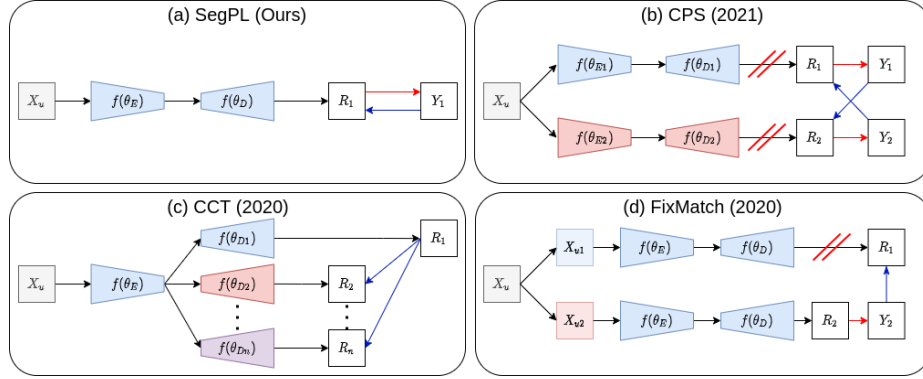


Fig. 2. Comparisons between SegPL and popular self-ensemble semi-supervised segmentation methods on unlabelled images. (a) Proposed SegPL. (b) CPS (cross-pseudo-supervision) (c) CCT (cross-consistency training). (d) FixMatch uses different augmentation at the input level. $\rightarrow$: data flow. $\rightarrow$: generate pseudo label. $\rightarrow$: supervision with loss function. $//$: stopping gradients.

5 Pre processing Of the Labels For Multi-Class Segmentation

The pre-processing of the labels of BRATS has two steps: 1) label fusion to mitigate the severe class imbalance between the minority tumour classes and the majority background healthy tissue class; 2) turning a multi-class label into multiple binary labels, for each binary prediction, we can use Sigmoid followed by confidence thresholding for pseudo labelling. Here we show two examples of label pre-processing, one is an abstract example (Fig.3) and the other one is a real example (Fig.4).

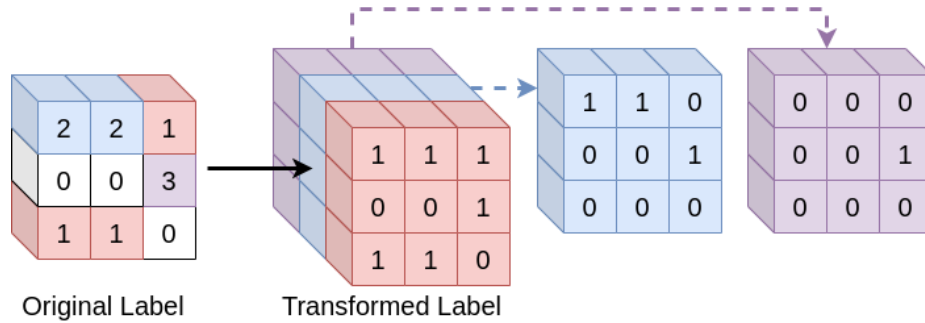


Fig. 3. An example of the pre-processing of one label of BRATS. 3: enhancing tumour core. 2: tumour core containing enhancing tumour core. 1: whole tumour containing class 2 and 3. 0: healthy tissues. Different colours represent different classes.

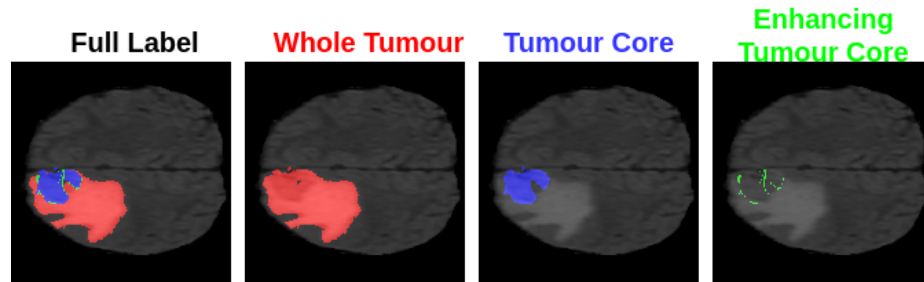


Fig. 4. Label fusion and binarized labels. Red: whole tumour including tumour core. Blue: tumour core including enhancing tumour core. Green: enhancing tumour core. Segmentation of each tumour class is a binary segmentation.

6 Segmentation Performance and Sizes of Foreground Areas

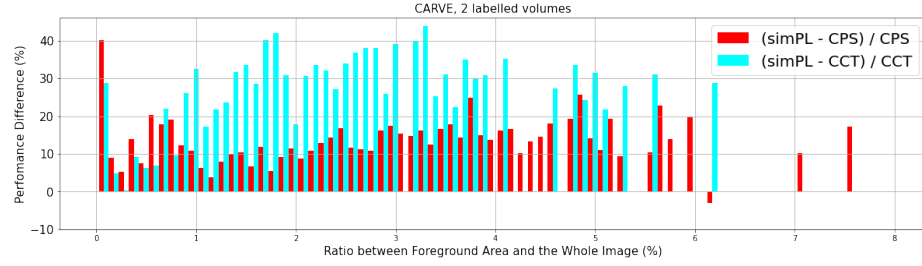


Fig. 5. How SegPL outperforms baselines at each size of foreground area. When the foreground area is extremely small (very left bin, less than 0.1%), SegPL outperforms CPS by 40%. It is also easy to see that, segmentation of vessels is a very hard task in general, the regions of the interest is less than 8% of the whole images. The segmentation of vessels suffers from severe class imbalance between the foreground and the background.

7 Temperature Scaling

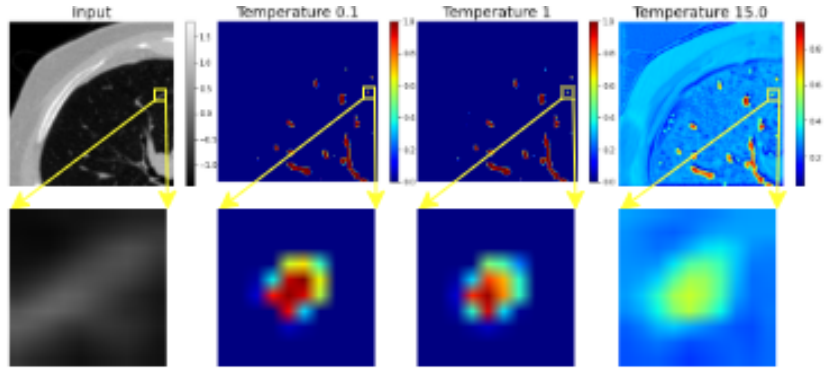


Fig. 6. Column 1: input. Column 2: temperature 0.1. Column 3: temperature 1.0. Column 4: temperature 15.0. The colour bar has deep red as most confident (1) and deep blue as least confident (0).

8 More Visual Results

9 Results Analysis

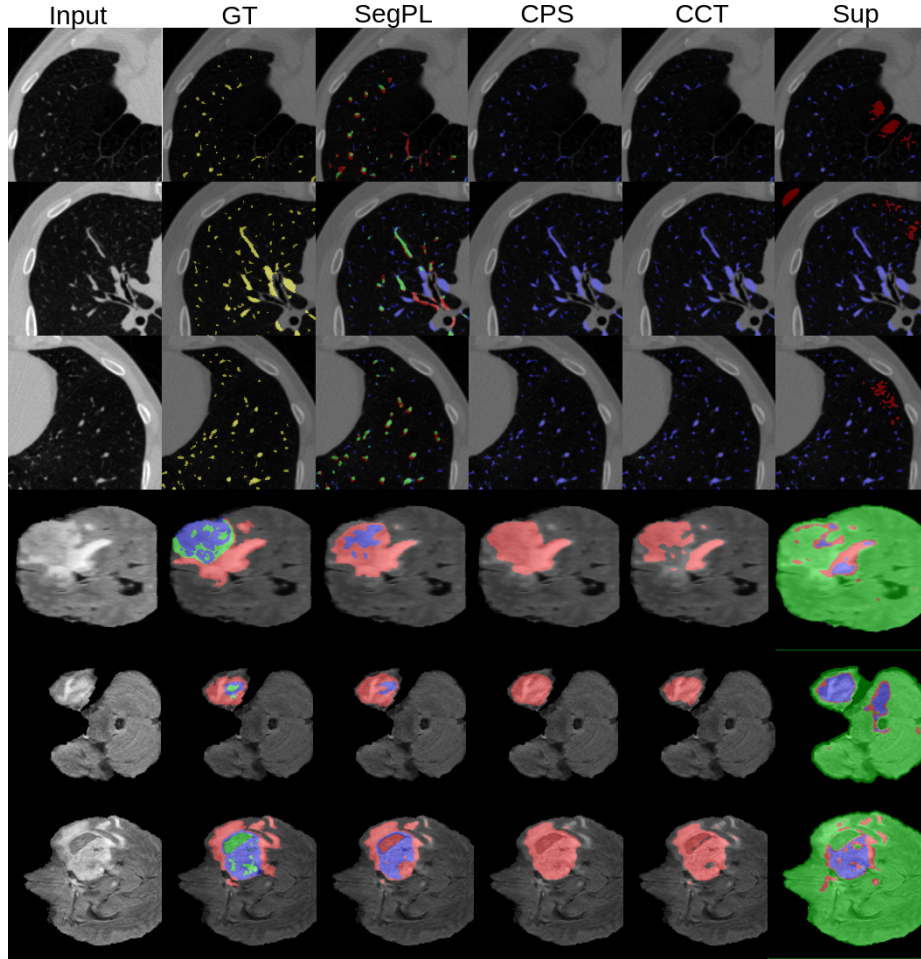


Fig. 7. Row1-3: CARVE trained with 5 labelled volumes. Red: false positive. Green: true positive. Blue: false negative. Yellow: ground truth. Row4-6: BRATS trained with 300 labelled slices. Red: whole tumour. Green: tumour core. Blue: enhancing tumour core. GT: Ground truth. CPS: cross pseudo labels (CVPR 2021). CCT: cross consistency training (CVPR 2020). Sup: supervised training.

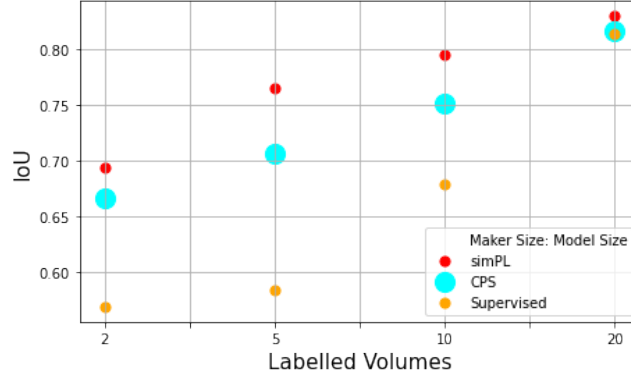


Fig. 8. Main results on the model size and the performances among different methods on CARVE data set. segPL outperforms baselines while uses less computational power.

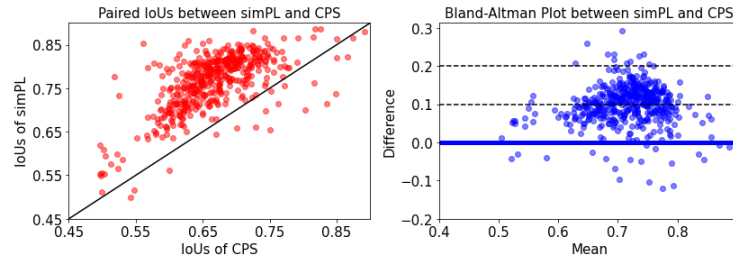


Fig. 9. Left: SegPL statistically outperforms the best performing baseline CPS when trained on 2 labelled volumes from the CARVE dataset. Each data point represents a single testing image. Right: Ablation study on the α which is the weight on the unsupervised learning part. Different data might need different α .