Reducing Annotator's Burden: Cross-Pseudo Supervision for Brain Tumor Segmentation

Lidia Luque*†[1, 2], Jon André Ottesen†[1, 2], Atle Bjørnerud[1, 2], Kyrre Eeg Emblem[3], and Bradley J MacIntosh[1, 4, 5]

[1]Computational Radiology & Artificial Intelligence (CRAI) unit, Department of Physics and Computational Radiology, Division of Radiology and Nuclear Medicine, Oslo University Hospital, Oslo, Norway
[2]Department of Physics, Faculty of Mathematics and Natural Sciences, University of Oslo, Oslo, Norway
[3]Department of Physics and Computational Radiology, Division of Radiology and Nuclear Medicine, Oslo University Hospital, Oslo, Norway
[4]Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada
[5]Sandra E Black Centre for Brain Resilience & Recovery, Sunnybrook Research Institute, Toronto, ON, Canada

Abstract

Deep learning is proven to help with common medical image processing procedures, namely segmentation. Labeling data is a core requirement for training a deep learning model; this is time-consuming and expert annotators are in short supply. Strategies that lower data annotation requirements are highly desirable. In this study, we adapt cross-pseudo supervision (CPS) for 3D medical segmentation, a state-of-art semi-supervised deep-learning method where labeled and unlabeled data are used in conjunction to further improve the resulting model. Using the 2021 BraTS dataset, a fully labeled publicly available brain tumor dataset, we train CPS-based networks using a varying number of labeled and unlabeled samples and compare the resulting models against the fully-supervised baseline. The results show that CPS improves performance scores across all combinations of dataset sizes, with an increase in the Dice similarity coefficient (DSC) of 2.6-4.2% and a decrease in the 95th percentile Hausdorff distance (95% HD) of 24-27%.

1 Introduction

Manual segmentation plays an important role in the annotation of focal pathology or tissues of interest. Annotation of medical images is a time- and labor-intensive process done by time-strained physicians. Developing automatic annotative methods is thus of high relevance to reduce clinician workloads.

Advances in deep learning have shown great capability in segmenting a variety of different organs and pathology, such as the liver [11, 20, 1], tumor masses [12, 22, 13] or abnormal cells [18, 14, 19, 2]. Still, large datasets labeled by experts are a prerequisite for achieving the segmentation accuracy and robustness needed for clinical implementation. The availability of clinical datasets and the expertise needed to annotate them are often limited. On the other hand, unlabeled data coming from similar distributions as the labeled data are often readily available. Semi-supervised deep-learning models offer the possibility of keeping labeling costs down by making use of these large quantities of unlabeled data.

Perhaps the most intuitive approach to semi-supervised learning, pseudo-labeling [15] uses a su-

*Corresponding Author: lidialuquef@gmail.com
†Shared first co-authorship

https://doi.org/10.7557/18.6815

pervised model to make predictions on an unlabeled dataset and then refines the model by using a selection of those predictions. The simplicity of this method has made it widely used for segmentation of medical imaging, including cardiac MRI [4], COVID-19 pneumonia lesions in CT scans [23, 9], and brain tumors in MRI [21, 24] amongst others. However, confirmation bias, where the model overfits to incorrect pseudo-labels, is a common pitfall.

Consistency regularization [3, 17] is another approach to semi-supervised learning that has become increasingly popular in recent years. This method is based on the assumption that applying a small perturbation, be it to the input data or the network, should not change the model's prediction. Thus, networks can be trained with standard supervised loss combined with a loss enforcing consistency of the predictions from the same sample. Chen et al. recently proposed a consistency regularization technique with network perturbation referred to as cross pseudo supervision (CPS) which achieved state-of-the-art accuracies in semi-supervised segmentation on Cityscapes and PASCAL VOC 2012 [6]. Two networks with the same architecture but different initializations are fed the same set of labeled and unlabeled images. Each network is trained separately with the available ground truth labels in a standard supervised manner. In addition, each network is also trained with the other network's prediction on a common image, which does not need to be labeled.

In this study, we implement CPS on two different architectures: a 3D UNet-Transformer (UNETR) network and a regular 3D UNet, and apply it to a glioma segmentation dataset. We perform experiments to understand how the size of the labeled and unlabeled datasets affects the model's accuracy and to which extent CPS can be used to reduce the amount of labeled training data required.

2 Methods

2.1 Dataset

The 2021 Brain Tumor Segmentation (BraTS) dataset [5] consists of 1251 multimodal brain MRIs from patients diagnosed with glioma across multiple institutions. Each image-series consists of precontrast T1-weighted, post-contrast T1-weighted, T2-weighted, and fluid-attenuated inversion recovery (FLAIR) scans. Expert annotators have delineated the ground truth regions of the three main tumor sub-components for all 1251 patients. The three tumor sub-components are: enhancing tumor, edema, and necrotic tumor core.

The dataset was split into three parts: a test dataset with 176 samples, a validation dataset with 25 samples, and a training dataset with 1050. Note that although the training dataset contains 1050 samples, the number used for supervised and semi-supervised during training will vary.

2.2 Network and training

2.2.1 Network Architecture

In this study, the vision transformer and UNet-based architecture UNETR [10] and a 3D UNet [16, 8] were used as the segmentation networks. The UNETR architecture consists of a vision transformer-based [7] encoder which is connected to a CNN decoder via skip connections. The input 3D volumes are divided into fixed-sized, non-overlapping 3D patches and embedded using a linear layer. Positional embeddings are added and the resulting sequence of vectors is used as the input to a standard transformer. Drawing inspiration from U-Net [16], skip connections are used to pass on both high- and low-level features to the decoder. In the skip connections, the outputs of the different layers of the transformer are reshaped back to the 3D input space with convolutional layers and concatenated with the output of the upsampling deconvolution in the CNN decoder. The MONAI framework UNETR and UNet implementations were used with standard pre-defined parameters.

2.2.2 Semi-Supervised Training

Given a set of two architecturally identical networks, the output predictions for an input $\boldsymbol{X}$ are given by

$$\boldsymbol{P}_1 = f(\boldsymbol{X}; \boldsymbol{\theta}_1) \quad \text{and} \quad \boldsymbol{P}_2 = f(\boldsymbol{X}; \boldsymbol{\theta}_2), \tag{1}$$

where $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are the predicted outputs after softmax normalization and $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are the model weights. Note that the input $\boldsymbol{X}$ is identical for both networks. The pseudo-labels $\hat{\boldsymbol{Y}}_1$ and $\hat{\boldsymbol{Y}}_2$ are found from the class with the highest confidence from the predicted confidence maps $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$, respectively.[1]

From these pseudo-labels and the ground-truth labels, denoted as $\boldsymbol{Y}$, two separate losses can be defined: first, a supervised loss for both models; second, a semi-supervised loss that utilizes the pseudo-labels generated by a network to train the predicted confidence map of the other network. The supervised loss, $\mathcal{L}_{sup}$, is given by

$$\mathcal{L}_{sup} = \frac{1}{|\mathcal{D}^l|} \sum_{X \in \mathcal{D}^l} Dice(\boldsymbol{P}_1, \boldsymbol{Y}) + Dice(\boldsymbol{P}_2, \boldsymbol{Y}), \quad (2)$$

where $\mathcal{D}^l$ is the labeled set of volumes in $\boldsymbol{X}$ and $Dice(\boldsymbol{P}_{1(2)}, \boldsymbol{Y})$ is the standard Dice loss between the predicted confidence map and the corresponding ground-truth label. The cross pseudo supervision loss $\mathcal{L}_{cps}$ is written as:

$$\begin{aligned}\mathcal{L}_{cps} = & \frac{1}{|\mathcal{D}^u|} \sum_{X \in \mathcal{D}^u} Dice(\boldsymbol{P}_1, \hat{\boldsymbol{Y}}_2) + Dice(\boldsymbol{P}_2, \hat{\boldsymbol{Y}}_1) \\ & + \frac{1}{|\mathcal{D}^l|} \sum_{X \in \mathcal{D}^l} Dice(\boldsymbol{P}_1, \hat{\boldsymbol{Y}}_2) + Dice(\boldsymbol{P}_2, \hat{\boldsymbol{Y}}_1). \end{aligned} \quad (3)$$

Note that while $\mathcal{L}_{cps}$ is defined on both the labeled and the unlabeled data, $\mathcal{D}^l$ and $\mathcal{D}^u$ respectively, $\mathcal{L}_{cps}$ is computed without using the ground-truth labels.

The total loss is the sum of both losses with a trade-off weight ϕ:

$$\mathcal{L} = \mathcal{L}_{sup} + \phi \mathcal{L}_{cps}. \quad (4)$$

In this study, we enforced $\phi = 0.5$ after some initial testing, however, there is room for further optimization.

2.3 Implementation details

From the entire training dataset of 1050 samples, we define four labeled subsets with 12, 25, 50, and 100 samples. Note that the subsets have overlapping samples, e.g., 12 of the samples in the 25-sample subset correspond to the subset of 12 samples. Supervised models (UNETR and UNet) were trained for each labeled dataset subset and for 1050 labeled samples. Semi-supervised CPS training was performed by defining a total dataset size containing the labeled subset. For example, for a total dataset size of 1050 samples, a model was trained with 12 (25, 50, and 100) of those samples labeled, whilst the remaining were unlabeled. In total, three total dataset sizes of 150, 600, and 1050 were used when training the different semi-supervised CPS models. Lastly, four supervised models were trained with CPS but without additional unlabeled data. In this case, $\mathcal{L}_{sup}$ and $\mathcal{L}_{cps}$ are both calculated on the same -labeled- input data, although $\mathcal{L}_{cps}$ still makes no use of the ground truth labels. In summary, the following models were trained: 10 (5 UNETR and 5 UNet) supervised models, 16 semi-supervised CPS models (12 UNETR and 4 UNet), and 4 supervised CPS models without unlabeled data. A visual overview of all trained models is given in Table 1. The semi-supervised training pipeline is illustrated in Figure 1.

		Total dataset size						
	Data	12	25	50	100	150	600	1050
Labeled Data	12	x*	-	-	-	x	x	x*
	25	-	x*	-	-	x	x	x*
	50	-	-	x*	-	x	x	x*
	100	-	-	-	x*	x	x	x*
	1050	-	-	-	-	-	-	x*

Table 1: A summary of the different 🟧 semi-supervised and 🟦 supervised models for UNETR (denoted by ”x”) and UNet (starred cells ”*”) as a function of the total amount of labeled data and the total dataset size.

All models were implemented in the PyTorch framework and trained on an Nvidia A100 (40GB) and an Nvidia V100 Volta (32GB) for 1000 epochs where one epoch iterates through 250 randomly drawn training examples. Standard data augmentation techniques were used: scaling, rotation, flipping, contrast adjustment, histogram shifting, intensity shifting, and intensity scaling. Gradient descent was performed with the Adam optimizer with weight decay set to $2e^{-5}$. The initial learning rate was initially set to $5e^{-4}$ and subsequently reduced with the cosine annealing scheduler. Each sample is randomly cropped to a patch size of $128 \times 128 \times 128$ and a batch size of 4 is used, 2 samples of which are unlabeled during semi-supervised training. We

[1] Also known as the *argmax* operation.

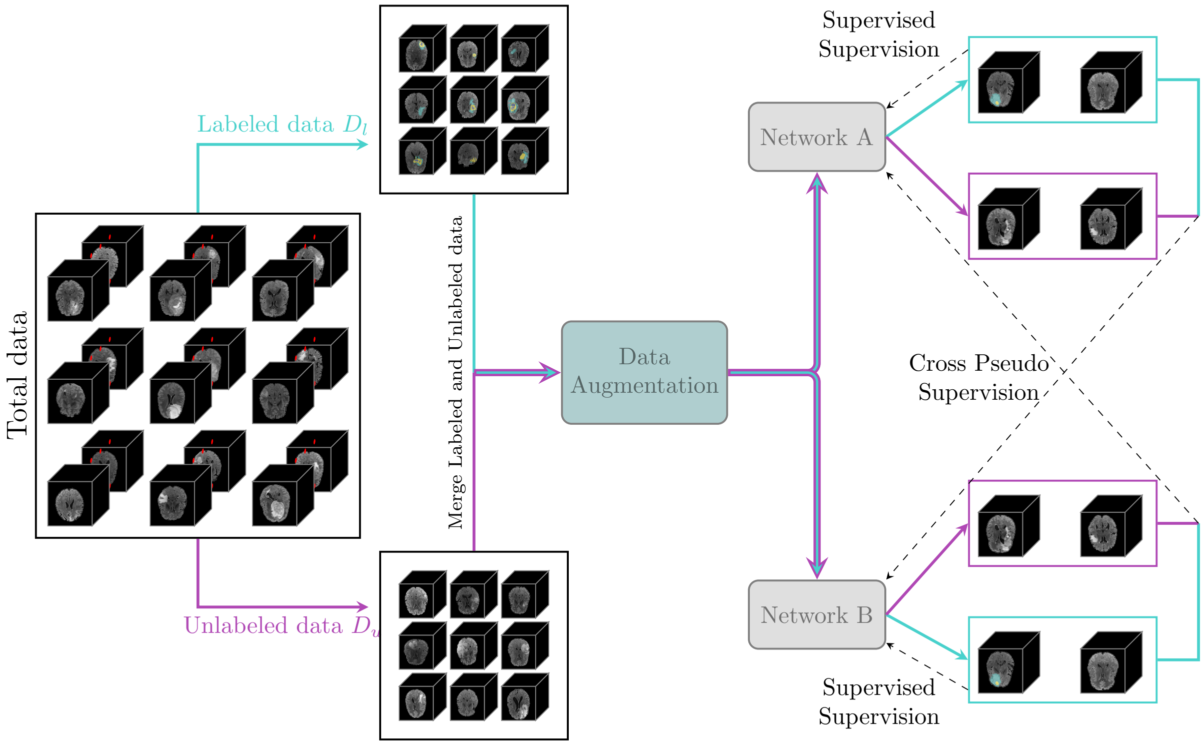

Figure 1: An illustration of the semi-supervised training pipeline. Given a total dataset, a subset is selected to be used as labeled and unlabeled data. Data augmentation is performed on a batch that consists of both subsets. The augmented batch is sent to the two networks, where the labeled data is used for supervised training. The predicted labels from the labeled and unlabeled subset are used for semi-supervised training of the other model.

validate on $n = 25$ samples at each epoch and pick the model with the lowest validation loss to use for inference. Lastly, a "wait period" of 15 epochs was used during semi-supervised training where the semi-supervised models were trained in a fully-supervised manner. This wait period helped convergence. We note that these hyperparameters worked well for both the UNETR and UNet architecture.

At test time, sliding window inference with overlapping patches was performed. When using a semi-supervised model, the inference was done with both networks and the final prediction was the class with the highest confidence for the networks' ensemble. Model performance was evaluated using the Dice similarity coefficient (DSC) and the 95th percentile Hausdorff distance (95% HD) between the prediction and ground truth excluding the background.

3 Results

Figures 2 and 3 display the Dice similarity coefficient and the 95th percentile Hausdorff distance for 12, 25, 50, and 100 labeled samples for supervised and semi-supervised training with a total dataset size of 1050 samples for both UNETR and UNet. Note that when using unlabeled data, the total amount of labeled data can be halved without considerably reducing the DSC nor the 95% HD. Both supervised and CPS exhibit similar improvements in the DSC and the 95% HD when increasing the amount of labeled data for both UNETR and UNet backbones.

Table 2 shows the breakdown of the DSCs for each class and for all models trained in supervised and semi-supervised manner with a total of 1050 samples. We obtain an average relative improvement in the mean DSC of 4.2% and 2.6%, equivalent to a mean absolute increase of 0.03 and 0.02 for

n_l	En.Tumor	%	Edema	%	Necrosis	%	Mean	%
				UNETR				
12	0.72/0.76	**5.3%**	0.69/0.68	-1.5%	0.52/0.57	**9.7%**	0.64/0.67	**4.1%**
25	0.74/0.77	**4.1%**	0.72/0.75	**3.2%**	0.56/0.60	**7.2%**	0.67/0.70	**4.6%**
50	0.78/0.80	**1.9%**	0.78/0.80	**2.3%**	0.61/0.68	**10.9%**	0.72/0.75	**4.6%**
100	0.80/0.82	**2.2%**	0.79/0.82	**4.0%**	0.68/0.71	**4.0%**	0.76/0.78	**3.4%**
				UNet				
12	0.78/0.79	**1.2%**	0.76/0.77	**1.5%**	0.63/0.67	**5.0%**	0.72/0.74	**2.4%**
25	0.79/0.81	**2.4%**	0.79/0.80	**2.0%**	0.67/0.69	**2.0%**	0.75/0.77	**2.2%**
50	0.81/0.83	**2.3%**	0.81/0.83	**2.7%**	0.69/0.72	**5.0%**	0.77/0.80	**3.3%**
100	0.83/0.84	**1.3%**	0.82/0.84	**2.6%**	0.71/0.74	**4.1%**	0.79/0.81	**2.6%**

Table 2: The test dice similarity coefficient (DSC) scores for supervised and semi-supervised models trained on a total dataset size of 1050 samples. The supervised and semi-supervised DSC are seperated by the "/" annotation (supervised/semi-supervised). The relative difference in the DSC are given for each class, bold typefont is used to emphasize an improvement in DSC after semi-supervised training. En.Tumor stands for enhancing tumor.

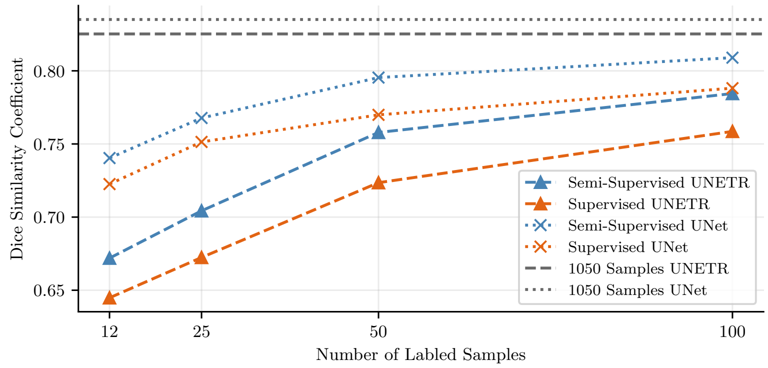

Figure 2: The Dice similarity coefficient for the supervised and semi-supervised training regimes with 12, 25, 50, and 100 for UNet and UNETR. A total of 1050 samples were used when training the semi-supervised models. Note, the two uppermost lines correspond to a fully-supervised UNet and UNETR model with 1050 labeled samples.

UNETR and UNet, respectively. Across all three classes, the necrotic tumor core shows the largest average gains of 8.0% with UNETR and 4.0% with UNet respective to the supervised baseline. The enhancing tumor shows an average increase in DSC of 3.4% (UNETR) and 1.8% (UNet), and the average increase in the edema scores is 2.0% and 2.2% with UNETR and UNet respectively. We note that among all experiments and for all classes there were only two occurrences of decreased DSC between the supervised and the semi-supervised model: both occurred for the edema class when training with the UNETR architecture with $n_l = 12$, one for a total dataset size of 1050 and the other one for a total dataset size of 600 (data now shown).

Table 3 show the mean 95% Hausdorff distance between all classes and for all models trained in a supervised and semi-supervised manner with a total of 1050 samples. We note an average improvement (decrease) of 27% and 24% for UNETR and UNet, respectively.

The training losses and the validation average DSC for a semi-supervised and a supervised training are plotted in Figure 4 for the UNETR architecture. The training is shown for a total dataset size of 1050 of which 25 samples were labeled. Both the supervised and the semi-supervised model's validation scores quickly increase during the first epochs. By epoch $\sim$ 50, the supervised validation scores flatten out, remaining at the same level for the rest of the training. On the other hand, the semi-supervised validation scores continue increasing until the end of the training, albeit at a much slower pace.

Figure 5 depicts the DSC for the supervised training scheme with 12, 25, 50, and 100 labeled

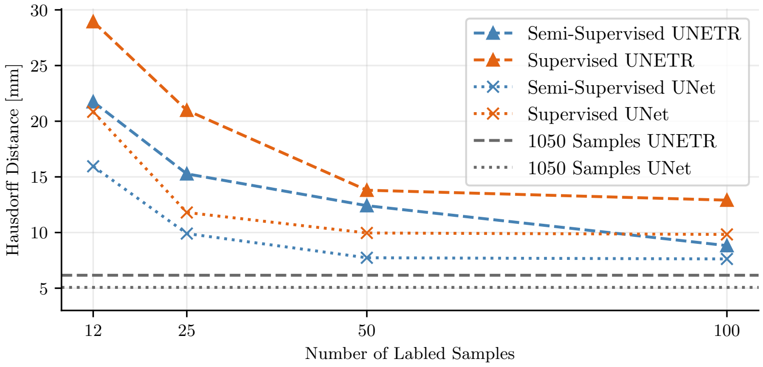

Figure 3: The 95th percentile Hausdorff distance for the supervised and semi-supervised training regimes with 12, 25, 50, and 100 for UNet and UNETR. A total of 1050 samples were used when training the semi-supervised models. Note, the two uppermost lines correspond to a fully-supervised UNet and UNETR model with 1050 labeled samples.

	UNETR		UNet	
n_l	Mean	%	Mean	%
12	28.9/21.7	**28.4**	20.8/15.0	**26.6**
25	21.0/15.3	**31.5**	11.8/9.9	**17.4**
50	13.8/12.4	**10.6**	10.0/7.7	**25.2**
100	12.9/8.8	**37.6**	9.8/7.6	**25.2**

Table 3: The test 95th percentile Hausdorff distance [mm] for supervised and semi-supervised models trained on a total dataset size of 1050 samples. The supervised and semi-supervised distances are separated by the "/" annotation (supervised/semi-supervised). The relative difference in the distances are also given, with bold typefont indicating an improvement (decrease) in the 95th percentile Hausdorff distance after semi-supervised training.

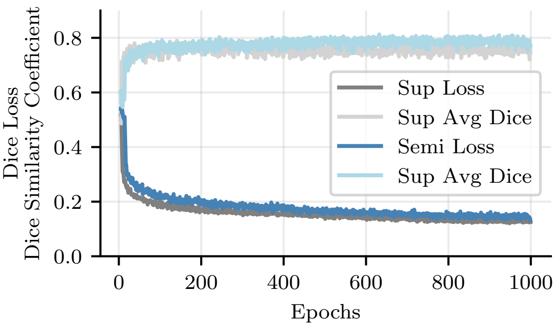

Figure 4: The training losses for the semi-supervised and the supervised training with the corresponding dice similarity coefficient on the validation set.

samples and the corresponding semi-supervised training scheme as a function of the total number of samples used during the training process (150, 600, and 1050). The first data point for each number of labeled samples shows the DSC for the supervised CPS training without unlabeled data. This was done to isolate the importance of the training method (CPS vs fully-supervised) from the contributions of unlabeled data. We observe that only using cross-pseudo supervision without adding unlabeled samples does not improve the model, with the resulting average DSCs being about equal or lower than the DSCs from the fully-supervised baseline. We can further observe that the unlabeled sample size does not have a substantial impact on the performance, and a total dataset size of 150 samples is enough to saturate further gains in accuracy.

4 Discussion

We demonstrate that semi-supervised CPS can achieve average relative gains in the DSC of 4.2% using UNETR and 2.6% using UNet compared to fully-supervised training. CPS also leads to a relative reduction in the 95% HD of 24% when using UNet and 27% when using UNETR. We thus consider CPS to be a suitable option for semi-supervised glioma segmentation. Moreover, the fact that the original CPS model showed similar improvements on non-medical 2D images [6] makes it likely that CPS will prove useful in a range of medical applications beyond segmenting brain tumors. Other segmentation tasks are clear candidates, but classification problems could also be well-suited for semi-supervised training with CPS.

Models trained on multi-class datasets often have substantial differences in accuracy between the classes. High variability within a class or imbal-

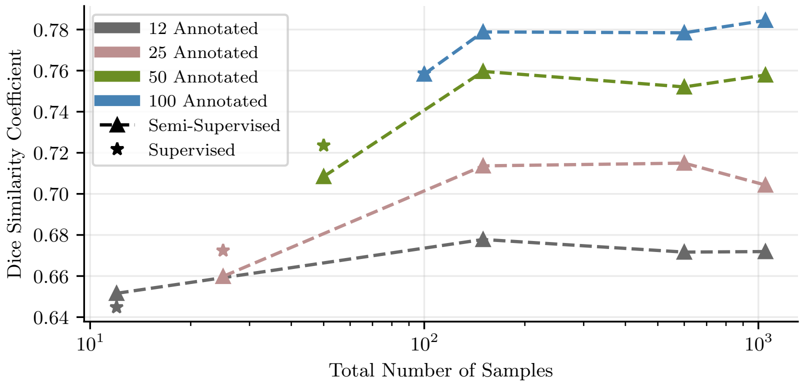

Figure 5: The Dice similarity coefficient (DSC) for the semi-supervised models with 12, 25, 50 and 100 labeled training samples as a function of the total number of samples used during the training process (150, 600, and 1050). The first data point refers to the model trained in a semi-supervised manner but without unlabeled samples. The starred datapoints show the DSC for the corresponding supervised models.

anced datasets can lead to low relative accuracies for a particular class. In our case, the necrotic tumor core is the class with the lowest DSC, and it is also the class where we see the largest improvements by using CPS, both in relative and absolute values (8.0% and 0.05 respectively for UNETR, 4.0% and 0.03 for UNet). This is important because, when faced with unacceptably low accuracy for a specific class, using CPS can help boost its score enough that further labeling is not needed.

We demonstrate that semi-supervised training with CPS is robust; CPS improves scores across all tested combinations of labeled and total dataset size. For example, adding only 50 unlabeled samples to an already large labeled dataset of 100 labeled samples still increases the average test DSC by $\sim$ 3%. We show that models trained on just half as many labeled samples as the fully-supervised model only perform 1.2% worse on average as measured by the DSC.

We note that CPS is easy to implement, making use of off-the-shelf networks and requiring only an adaptation of the training loss function. There are, however, two major disadvantages with CPS: first, the total training time was almost doubled compared to fully-supervised training due to the additional model being trained; second, memory consumption was more than doubled, going from 12.4GB to 27.4 GB for supervised and semi-supervised training, respectively, on the UNETR model. Parallelizing the model so that each network is trained on one GPU would reduce training time and allow for a smaller memory footprint on each GPU. If only one GPU is available, alternating between training each of the networks (saving their predictions to memory) could also be an option to reduce memory usage, but would considerably increase training time. We have, however, not tested either of these options.

Even with the disadvantages of additional memory consumption and training time, we believe that training with CPS should be strongly considered whenever unlabeled samples and the needed hardware are available.

5 Conclusion

We show that semi-supervised cross pseudo-supervision generalizes to 3D medical imaging segmentation. For brain tumor segmentation using the 2021 BraTS dataset, CPS increases accuracy by between 2.6% (UNet) and 4.2% (UNETR) on average across different amounts of labeled and unlabeled dataset sizes. CPS makes it possible to significantly decrease the annotators' burden and should be strongly considered when unlabeled data are easy to obtain.

References

[1] M. Ahmad, D. Ai, G. Xie, S. F. Qadri, H. Song, Y. Huang, Y. Wang, and J. Yang. Deep belief network modeling for automatic liver segmentation. *IEEE Access*, 7:20585–20595, 2019. ISSN 21693536. doi: 10.1109/ACCESS.2019.2896961.

[2] F. H. Araújo, R. R. Silva, D. M. Ushizima, M. T. Rezende, C. M. Carneiro, A. G. C.

Bianchi, and F. N. Medeiros. Deep learning for cell image segmentation and ranking. *Computerized Medical Imaging and Graphics*, 72: 13–21, 3 2019. ISSN 0895-6111. doi: 10.1016/J.COMPMEDIMAG.2019.01.003.

[3] P. Bachman, O. Alsharif, and D. Precup. Learning with pseudo-ensembles. *Advances in Neural Information Processing Systems*, 27, 2014. doi: 10.48550/arXiv.1412.4864.

[4] W. Bai, W. Shi, D. P. O'Regan, T. Tong, H. Wang, S. Jamil-Copley, N. S. Peters, and D. Rueckert. A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: application to cardiac mr images. *IEEE transactions on medical imaging*, 32:1302–1315, 2013. ISSN 1558-254X. doi: 10.1109/TMI.2013.2256922.

[5] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati, L. M. Prevedello, J. D. Rudie, C. Sako, R. T. Shinohara, T. Bergquist, R. Chai, J. Eddy, J. Elliott, W. Reade, T. Schaffter, T. Yu, J. Zheng, A. W. Moawad, L. O. Coelho, O. McDonnell, E. Miller, F. E. Moron, M. C. Oswood, R. Y. Shih, L. Siakallis, Y. Bronstein, J. R. Mason, A. F. Miller, G. Choudhary, A. Agarwal, C. H. Besada, J. J. Derakhshan, M. C. Diogo, D. D. Do-Dai, L. Farage, J. L. Go, M. Hadi, V. B. Hill, M. Iv, D. Joyner, C. Lincoln, E. Lotan, A. Miyakoshi, M. Sanchez-Montano, J. Nath, X. V. Nguyen, M. Nicolas-Jilwan, J. O. Jimenez, K. Ozturk, B. D. Petrovic, C. Shah, L. M. Shah, M. Sharma, O. Simsek, A. K. Singh, S. Soman, V. Statsevych, B. D. Weinberg, R. J. Young, I. Ikuta, A. K. Agarwal, S. C. Cambron, R. Silbergleit, A. Dusoi, A. A. Postma, L. Letourneau-Guillon, G. J. G. Perez-Carrillo, A. Saha, N. Soni, G. Zaharchuk, V. M. Zohrabian, Y. Chen, M. M. Cekic, A. Rahman, J. E. Small, V. Sethi, C. Davatzikos, J. Mongan, C. Hess, S. Cha, J. Villanueva-Meyer, J. B. Freymann, J. S. Kirby, B. Wiestler, P. Crivellaro, R. R. Colen, A. Kotrotsou, D. Marcus, M. Milchenko, A. Nazeri, H. Fathallah-Shaykh, R. Wiest, A. Jakab, M.-A. Weber, A. Mahajan, B. Menze, A. E. Flanders, and S. Bakas. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. 7 2021. doi: 10.48550/arxiv.2107.02314.

[6] X. Chen, Y. Yuan, G. Zeng, and J. Wang. Semi-supervised semantic segmentation with cross pseudo supervision. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 6 2021. ISSN 10636919. doi: 10.48550/arxiv.2106.01226.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *The International Conference on Learning Representations (ICLR)*, 2021. doi: 10.48550/arXiv.2010.11929.

[8] T. Falk, D. Mai, R. Bensch, Özgün Çiçek, A. Abdulkadir, Y. Marrakchi, A. Böhm, J. Deubner, Z. Jäckel, K. Seiwald, A. Dovzhenko, O. Tietz, C. D. Bosco, S. Walsh, D. Saltukoglu, T. L. Tay, M. Prinz, K. Palme, M. Simons, I. Diester, T. Brox, and O. Ronneberger. U-net: deep learning for cell counting, detection, and morphometry. *Nature Methods*, 16:67–70, 12 2018. ISSN 1548-7105. doi: 10.1038/S41592-018-0261-2.

[9] D. P. Fan, T. Zhou, G. P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE Transactions on Medical Imaging*, 39:2626–2637, 8 2020. ISSN 1558254X. doi: 10.1109/TMI.2020.2996645.

[10] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu. Unetr: Transformers for 3d medical image segmentation. *Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022*, pages 1748–1758, 3 2021. doi: 10.48550/arxiv.2103.10504.

[11] P. Hu, F. Wu, J. Peng, P. Liang, and D. Kong. Automatic 3d liver segmentation based on deep learning and globally optimized surface evolution. *Physics in Medicine Biology*, 61: 8676, 11 2016. ISSN 0031-9155. doi: 10.1088/1361-6560/61/24/8676.

[12] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods 2020 18:2*, 18:203–211, 12 2020. ISSN 1548-7105. doi: 10.1038/S41592-020-01008-Z.

[13] J. Jiang, Y. C. Hu, C. J. Liu, D. Halpenny, M. D. Hellmann, J. O. Deasy, G. Mageras, and H. Veeraraghavan. Multiple resolution residually connected feature streams for automatic lung tumor segmentation from ct images. *IEEE Transactions on Medical Imaging*, 38:134–144, 1 2019. ISSN 1558254X. doi: 10.1109/TMI.2018.2857800.

[14] Kurnianingsih, K. H. S. Allehaibi, L. E. Nugroho, Widyawan, L. Lazuardi, A. S. Prabuwono, and T. Mantoro. Segmentation and classification of cervical cells using deep learning. *IEEE Access*, 7:116925–116941, 2019. ISSN 21693536. doi: 10.1109/ACCESS.2019.2936017.

[15] D.-H. Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *International Conference on Machine Learning Workshops (ICMLW)*, 2013.

[16] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9351:234–241, 2015. ISSN 16113349. doi: 10.1007/978-3-319-24574-4_28.

[17] M. Sajjadi, M. Javanmardi, and T. Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in Neural Information Processing Systems*, pages 1171–1179, 6 2016. ISSN 10495258. doi: 10.48550/arxiv.1606.04586.

[18] K. Sirinukunwattana, S. E. Raza, Y. W. Tsang, D. R. Snead, I. A. Cree, and N. M. Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Transactions on Medical Imaging*, 35:1196–1206, 5 2016. ISSN 1558254X. doi: 10.1109/TMI.2016.2525803.

[19] Y. Song, E. L. Tan, X. Jiang, J. Z. Cheng, D. Ni, S. Chen, B. Lei, and T. Wang. Accurate cervical cell segmentation from overlapping clumps in pap smear images. *IEEE Transactions on Medical Imaging*, 36:288–300, 1 2017. ISSN 1558254X. doi: 10.1109/TMI.2016.2606380.

[20] X. Tang, E. J. Rangraz, W. Coudyzer, J. Bertels, D. Robben, G. Schramm, W. Deckers, G. Maleux, K. Baete, C. Verslype, M. J. Gooding, C. M. Deroose, and J. Nuyts. Whole liver segmentation based on deep learning and manual adjustment for clinical use in sirt. *European Journal of Nuclear Medicine and Molecular Imaging*, 47:2742–2752, 11 2020. ISSN 16197089. doi: 10.1007/S00259-020-04800-3.

[21] B. H. Thompson, G. D. Caterina, and J. P. Voisey. Pseudo-label refinement using superpixels for semi-supervised brain tumour segmentation. *Proceedings - International Symposium on Biomedical Imaging*, 2022-March, 10 2021. ISSN 19458452. doi: 10.1109/ISBI52829.2022.9761681.

[22] A. Vakanski, M. Xian, and P. E. Freer. Attention-enriched deep learning model for breast tumor segmentation in ultrasound images. *Ultrasound in Medicine Biology*, 46: 2819–2833, 10 2020. ISSN 0301-5629. doi: 10.1016/J.ULTRASMEDBIO.2020.06.015.

[23] X. Wang, Y. Yuan, D. Guo, X. Huang, Y. Cui, M. Xia, Z. Wang, C. Bai, and S. Chen. Ssa-net: Spatial self-attention network for covid-19 pneumonia infection segmentation with semi-supervised few-shot learning. *Medical Image Analysis*, 79:102459, 7 2022. ISSN 1361-8415. doi: 10.1016/J.MEDIA.2022.102459.

[24] Z. Zheng, X. Wang, X. Zhang, Y. Zhong, X. Yao, Y. Zhang, and Y. Wang. Semi-supervised segmentation with self-training based on quality estimation and refinement. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12436 LNCS:30–39, 2020. ISSN 16113349. doi: 10.1007/978-3-030-59861-7_4.