A MuST for Consistency Regularization in Semi-Supervised Medical Image Segmentation

No Author Given

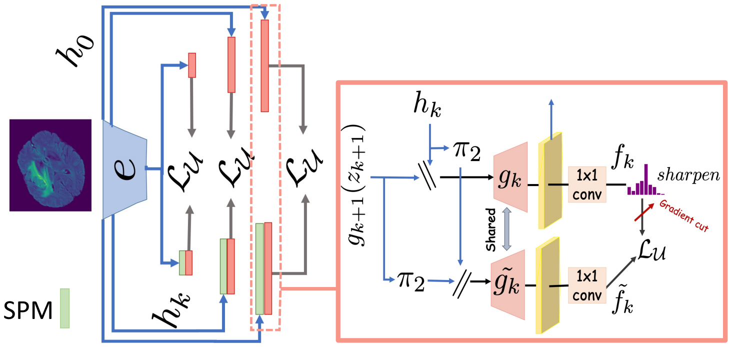
No Institute Given

Abstract. Deep neural networks have achieved significant performance in semantic segmentation especially when labeled data is abundant. However, acquiring large amounts of labeled data is a non-trivial, expensive, and time-consuming task in medical imaging that requires experts' input. Since unlabeled examples are easier to acquire, it is desirable to exploit available unlabeled data to improve the semantic segmentation network's performance. In this paper, we propose a novel Multi-Scale Training (MuST) semi-supervised framework based on Consistency Regularization for medical image segmentation task. Deep neural networks have achieved significant performance in semantic segmentation especially when labeled data is abundant. However, acquiring large amounts of labeled data is a non-trivial, expensive, and time-consuming task in medical imaging that requires experts' input. Since unlabeled examples are easier to acquire, it is desirable to exploit available unlabeled data to improve the semantic segmentation network's performance. In this paper, we propose a novel Multi-Scale Training (MuST) semi-supervised framework based on Consistency Regularization for medical image segmentation task.To effectively leverage the unlabeled examples, we specifically apply the Consistency Regularization technique on intermediate decoder layers independently. This simple and general framework can be applied to any encoder-decoder neural network such as U-net. The proposed framework was evaluated on white matter hyperintensity segmentation and brain tumor segmentation. Our framework, trained on a small number of labeled samples and a relatively abundant unlabeled samples, outperformed supervised baselines and achieved comparable results when all the labeled samples were available. We perform an extensive ablation study to evaluate the effectiveness of each part of our framework as well as its effectiveness when different labeled data sample sizes.

Keywords: Semi-supervised · Feature Space Perturbation · Layerwise Consistency Regularization

1 Introduction

Image segmentation is a fundamental task in medical image analysis as it provides a tool to enable computer-aided diagnostic and quantify anatomical structures, lesions, and diseases [22, 16]. Despite the significant improvement obtained by modern deep neural networks in various tasks, training a strong segmentation



Unsupervised Training Computation at an Arbitrary Layer

Fig. 1: **Overview of Multi-Scale Consistency Training (MuST)**

network requires a significant amount of pixel-wise labeled data [10]. However, large-scale labeled data acquisition is challenging and time-consuming, especially in the medical domain, which requires domain experts' input [27]. Consequently, the medical imaging community has increasingly begun investigating approaches such as semi-supervised learning (SSL). SSL is appealing since it takes advantage of available unlabeled data to alleviate the need for labeled examples.

There exist many approaches for SSL which have been proposed in the literature. One of the most common methods for SSL is pseudo-labeling, which attempts to generate pseudo-labels for unlabeled images and combines these with labeled data to update the network [23, 19, 2, 25, 13, 1, 14]. Entropy minimization is similar to pseudo-labeling, but minimizes the entropy of the model outputs to produce more confident predictions [4, 9, 21]. Performance of these approaches heavily relies on the quality of model outputs (i.e., for generating pseudo-labels or computing entropy), which can be noisy and affect future predictions. Other approaches employ generative adversarial networks (GANs) for generative modeling in SSL [26, 5]. However, effective GAN training is difficult and requires extensive hyperparameter tuning. In this work, we focus on Consistency Regularization (CR), which is a popular approach, achieving state-of-the-art performance in various tasks. Specifically, CR produces a robust model by enforcing consistency in model output across different perturbations of unlabeled input data [12, 24, 7, 17, 8].

The efficacy of CR is highly dependent on two conditions. First, the cluster assumption should hold; i.e., the decision boundary should reside in a low-density region, according to the data-distribution. While this may be violated in the input space, it has been shown to be preserved in the feature space [17, 8]. Second, a strong and diverse set of data-perturbations are vital for effective CR training [6, 17, 8]. Cross-Consistency training (CCT) [17] proposes to address both of these conditions by using feature space augmentation to enforce consistency between decoders in a semantic segmentation task. This method achieves state-of-the-art performance and has been successfully applied in whole-brain segmentation [3]. The approach is appealing for brain lesion segmentation such as White Matter Hyperintensity (WMH) segmentation, since augmentation occurs in the hidden space, rather than the image space. Naive augmentation in image space can eradicate lesions, which may present as small or even dots, and therefore alters the semantic of the brain image. Although CCT has been applied in whole-brain segmentation, more complex brain segmentation tasks, such as WMH segmentation, can be improved by multi-scale approaches which are beneficial for the segmenting fine-grained boundaries [18]. Until now, no method considers the benefit of multi-scale training for consistency regularization in brain segmentation tasks.

Accordingly, this paper proposes a novel Multi-Scale Training (MuST) for semi-supervised framework which enforces consistency at different resolutions. Concretely, each decoder layer of U-net receives an original and perturbed version of the feature representation and learns to produce a consistent prediction across layers. Hence, the shared encoder learns to produce robust feature representations at different resolutions, and each decoder layer is updated independently to learn perturbation invariant features. This way, the model can leverage unlabeled examples to better learn to identify small lesions and lesion boundaries. We evaluate this proposal using two publicly available datasets for brain tumor and WMH segmentation tasks. We perform extensive ablation study to show that feature-space augmentation at different scales is beneficial compared to augmentation at only a single scale.

2 Approach

2.1 Problem Definition

Let $\mathcal{D}_{\mathcal{L}} = \{(x_1^l, y_1^l), (x_2^l, y_2^l), ..., (x_m^l, y_m^l)\}$ and $\mathcal{D}_{\mathcal{U}} = \{x_1^u, x_2^u, ..., x_n^u\}$ be the labeled and unlabeled set of training examples respectively, where x_i^u and x_i^l are the i^{th} unlabeled and labeled image with spatial dimension of $M \times H \times W$ and y_i^l is a target image corresponding to x_i^l with dimensions $C \times H \times W$. Here, M is the number of input modalities, C is the number of classes, and $H \times W$ gives the pixel dimensions. The goal is to train a network for medical image segmentation, leveraging a large set of unlabeled training examples and only a few labeled examples (i.e., $n >> m$) so that the network generalizes well on unseen images.

We define an encoder network $e = (e_K, e_{K-1}, ..., e_0)$ and a main decoder network $g = (g_0, g_1, ..., g_K, g_{K+1})$. In the experiments, we use a 2D U-net with symmetric encoder-decoder as the backbone of our framework [20]. The U-net encoder-layers are connected using down-sampling functions. Therefore, the output of the encoder-layer at layer k is is $h_k = \downarrow (e_k(h_{k-1}))$ where $\downarrow (\cdot)$ is a 2×2 max-pooling layer and the recursive base case h_{-1} is set to be the input image. The decoder g_k is connected to the previous decoder g_{k+1} using an up-sampling function and its corresponding encoder-layer e_k via skip connection. The input of the decoder-layer g_k (i.e., at layer k) is written:

$$z_k = \begin{cases} h_{k-1} & \text{if } k = K+1 \\ h_k \parallel \uparrow (g_{k+1}(z_{k+1})) & \text{otherwise} \end{cases} \tag{1}$$

where K is the number of pairwise encoder-decoder layers, g_{K+1} is the first decoder (i.e., the bottleneck layer), $\parallel$ is channel-wise concatenation, and $\uparrow(\cdot)$ is an up-sampling function. For simplicity, we do not distinguish between the bottleneck layer and other decoder-layers in the remainder of this paper.

Our goal is to enforce consistency on the output of each decoder g_k. For this, an auxiliary classifier a_k is inserted after decoder-layer to produce segmentation predictions at layer k. Specifically, $f_k = (f_{k,1}, \dots, f_{k,C}) = (a_k \circ g_k)(z_k)$ gives the pixel-wise scores of the network at layer k for each of the C classes, and $f = f_0$ gives the final scores of the network, which is used for segmenting new examples at test-time. In particular, from the input image x (i.e., the recursive base case) f can be computed using the recursive equation given previously. Occasionally, to emphasize this dependence on x we abuse notation and write $f(x)$. Similarly, we write $f_k(x)$ for the output of the network at layer k with x as the input. In order to perform consistency regularization at different scales, we produce augmented version of z_k by applying a stochastic feature-space perturbation π and enforcing consistency on the produced feature maps f_k and $\tilde{f_k} = (a_k \circ g_k)(\tilde{z_k})$ where $\tilde{z_k} = \pi(z_k)$. In other words, the decoder g_k can be interpreted as a shared weight teacher-student model. Throughout the paper, g_k and $\tilde{g_k}$ refer to the teacher (i.e., $g_k(z_k)$) and student (i.e., $g_k(\tilde{z_k})$) decoders, respectively.

Our proposed framework has two main components: (i) multi-scale consistency regularization training and (ii) a Stochastic Perturbation Module (SPM). The SPM is introduced first, followed by the multi-scale consistency training. Figure 1 illustrates the overall architecture and training procedure of the proposed framework.

2.2 Stochastic Perturbation Module

As shown in Figure 1, before each decoder g_k, an SPM layer is inserted. It is responsible for generating $\tilde{z_k}$ from z_k and providing them to the teacher and student decoders, respectively. Given a set of predefined perturbation functions Π, with a uniform probability, SPM selects two perturbation functions, $\pi_1, \pi_2 \in \Pi$ to obtain $\tilde{z_k} = \pi_1(h_k) \parallel \uparrow (\pi_2(g_{k+1}(z_{k+1})))$. In our experiments, we consider Uniform noise, Gaussian noise, feature dropout and spatial dropout.

2.3 Multi-Scale Consistency Training

Multi-Scale Consistency Regularization. As previously noted, in order to leverage unlabeled examples we enforce consistency regularization at each decoder-layer. At each layer, z_k and $\tilde{z_k}$ are computed as explained in Section 2.2 and passed to the teacher and student decoders and auxiliary classifiers to produce predictions f_k and $\tilde{f_k}$, respectively. As shown in Figure 1 the teacher output

$g_k(z_k)$ is used as an input to the next decoder g_{k-1}. Following [4], pseudo-labels for consistency regularization are produced by sharpening the initial teacher predictions:

$$\forall j \in [1, ..., C] \;:\; f_{k,j} \leftarrow Sharpen(f_k, T) = \frac{f_{k,j}^{1/T}}{\sum_{c=1}^{C} f_{k,c}^{1/T}} \tag{2}$$

where T is temperature parameter. Next, using the pseudo-labels, we define the multi-scale consistency loss function $\mathcal{L}_{MSC}$ as follows:

$$\mathcal{L}_{MSC} = \frac{1}{|\mathcal{D}_{\mathcal{U}}|} \sum_{x_u \in \mathcal{D}_{\mathcal{U}}} \sum_{k=0}^{K+1} \mathbf{d}(f_k(x_u), \tilde{f}_k(x_u)) \tag{3}$$

where $\mathbf{d}(\cdot, \cdot)$ is a distance function. In this project we have adopted mean squared error distance measure. The objective $\mathcal{L}_{MSC}$ is optimized in combination with a supervised loss described next.

Supervised Training. In addition to the multi-scale consistency objective $\mathcal{L}_{MSC}$, we consider a multi-scale supervised training approach [18]. This should better leverage the supervised examples [18] and is important for multi-scale consistency regularization, since the procedure depends on the output of each auxiliary classifier a_k. The multi-scale supervised training approach below instills some prior knowledge in the auxiliary classifiers, according to the small amount of labeled data available. Specifically, we adapt the standard pixel-wise cross-entropy loss $\ell_{ce}(\cdot, \cdot)$ to a multi-scale supervised loss function by aggregating over all auxiliary predictions and the final prediction as follows:

$$\mathcal{L}_s = \frac{1}{|\mathcal{D}_{\mathcal{L}}|} \sum_{(x_i^l, y_i^l) \in \mathcal{D}_{\mathcal{L}}} \sum_{k=0}^{K+1} \ell_{ce}(f_k(x_i^l), y_{i,k}^l). \tag{4}$$

The target $y_{i,k}$ is constructed from the main ground truth label by applying a 2×2 function consecutively: $y_{i,k}^l = \downarrow (y_{i,k-1}^l)$ with recursive base case $y_{i,0}^l = y_i^l$. Finally, the overall semi-supervised objective we optimize is $\mathcal{L}_{\text{semi}}$ defined:

$$\mathcal{L}_{\text{semi}} = \mathcal{L}_s + w_u \cdot \mathcal{L}_{MSC} \tag{5}$$

where w_u is the multi-scale consistency loss weight used to balance the role of consistency regularization in training. In our experiments, w_u is slowly increased from zero to diminish the negative impact of the initial noisy teacher predictions.

3 Experiments and Results

3.1 Datasets and Evaluation Metrics

We evaluate our proposed semi-supervised framework on brain tumor segmentation and white matter hyperintensity (WMH) segmentation using publicly available datasets: BraTS [15] and WMH challenge [11].

BraTS. BraTS2018 contains 285 training subjects and 66 subjects for validation. Additionally, BraTS2019 introduces 50 extra subjects and BraTS2020 introduces 34 more subjects. Each patient has FLAIR, T1, T2, and T1ce MRI images. Ground truth images have three labels: GD-enhancing tumor (ET), peritumoral edema (ED), necrotic and non-enhancing tumor core (NCR/NET). These labels are combined to create overlapping classes: Whole tumor (WT = ET + ED + NET), enhancing tumor (ET), and tumor core (TC = ET + NET). In our experiments, we used BraTS2018, BraTS2019 (50 newly introduced subjects), and BraTS2020 (34 newly introduced subjects) as our train, validation, and test sets, respectively.

WMH Challenge. WMH Challenge dataset includes 60 FLAIR and T1 MRI images with corresponding binary segmentation of ground-truth WMH and brain tissue mask. In our experiments, we held out 12 subjects for testing while ensuring that an equal number of subjects are chosen from each site.

3.2 Experimental Setup

Implementation Details. Our framework is implemented using PyTorch and trained on 4 x NVIDIA Quadro RTX 5000 GPU for 50 epochs (BraTS) or 100 epochs (WMH). In all experiments, the batch size for supervised and unsupervised training is 32. We use SGD for optimizing loss functions with initial rate 0.01 for all experiments. We decrease the learning rate by a factor of 2 every 5, 10, 20, or 30 steps for BraTS (supervised), BraTS (unsupervised), WMH (supervised), and WMH (unsupervised) losses respectively. Parameter T in Eq. (2) is 0.5 (BraTs) or 0.8 (WMH). We use a linear ramp-up function to increase w_u in Eq. (5) from 0 to 20 (BraTs) or 30 (WMH).

Experimental Setting. Three-fold Monte-Carlo sampling is used in all experiments to ensure a fair comparison. Specifically, three supervised and unsupervised datasets are randomly generated for each setting. Further, three seeds have been set to ensure reproducibility. For BraTS, average testing performance is reported using the model with the best average validation result among seeds and classes. For the WMH Challenge dataset, the average performance after 100 epoch iterations is reported.

Evaluation Metric. Dice coefficient evaluation is used for the purpose of evaluation. Since the segmentation tasks aim for accurate subject-wise predictions, after the slice-wise predictions, slices are concatenated to obtain a subject-wise Dice score, and the subject-wise average score is reported.

Baselines. We compare our framework against recent semi-supervised related works on brain lesion segmentation. MASSL [5] used a 3D U-net to reconstruct synthetic labels generated by an attention mechanism for brain lesion and WMH segmentation task. MT [7] adapted the mean teacher-student framework for brain tumor segmentation by adding Gaussian noise to the student and teacher inputs. Unlike ours, the former is not a CR approach. While the latter is a CR approach, it performs augmentation in the input space, which can be problematic for fine-grained segmentation tasks (see Section 1).

Table 1: Comparison of our framework with supervised baseline (SB) on BraTS. L and U are number of labeled and unlabeled examples used in training, respectively.

Method	**L (U)**	**Dice % mean(std)**		
		WT	ET	TC
SB	285(0)	90.88(0.38)	86.94(0.24)	88.3(0.94)
SB	8(0)	66.84(6.06)	59.40(11.67)	54.01(13.62)
SB	14(0)	76.21(2.60)	70.95(1.61)	66.18(1.55)
SB	28(0)	83.85(0.83)	76.91(3.20)	73.94(3.88)
Ours	8(277)	71.02(5.93)	61.23(11.3)	56.58(13.3)
Ours	14(271)	79.88(0.32)	73.59(0.76)	70.43(1.26)
Ours	28(257)	**83.93(0.54)**	**78.22(0.97)**	**75.01(1.57)**

Practical Consideration. We sample the same amount of labeled and unlabeled examples in each iteration. However, unlike many recent works [17] that define the number of iterations based on the size of the bigger dataset (i.e., between labeled and unlabeled datasets), we define the number of iterations based on the size of the smaller dataset and alternate among supervised and unsupervised training. The former can lead to a biased interpretation as the model iterates over the labeled data $\mathcal{D}_{\mathcal{L}}$ more than the unlabeled data $\mathcal{D}_{\mathcal{U}}$. When strong data augmentation is used, this can lead to a higher performance regardless of the unsupervised training, and therefore, can lead to incorrect interpretation of improvement. Also alternate training is beneficial as different hyper-parameters can be leveraged for supervised and unsupervised objectives.

3.3 Results

Improvement Over the Supervised Baselines. To examine the effectiveness of the proposed framework, we compare the semi-supervised performance with the supervised baseline (SB) using different data partitions on two datasets. Table 1 illustrates that our method constantly improves the baseline on all classes and partitions. Specifically, our model enhances baseline's performance by 4.2%, 1.8%, and 2.6% with only using 8 labeled subjects on WT, ET, and TC, respectively. Furthermore, our framework obtains even more improvement over the supervised baseline on TC when using 14/48 labeled examples. As TC and ET are smaller lesions than WT, this observation supports our hypothesis that performing consistency regularization in different scales is beneficial for the smaller lesions and boundaries. We also see similar positive results on the WMH Challenge dataset, where we improve 4% Dice score over the supervised baselines when using 3 out 48 labeled examples (provided in the supplement). In general, our approach consistently improves upon the supervised baseline.

Table 2: Comparison of MuST when applying consistency regularization on different layers on BraTS

Layer k	WT	ET	TC
0	69.92(5.17)	**61.92(9.19)**	55.66(10.91)
1	68.27(5.90)	59.32(12.56)	53.77(14.46)
$K+1$	69.98(4.61)	60.77(10.92)	54.63(13.25)
$0, 1, (K+1)$	70.94(6.19)	59.91(12.31)	54.69(13.82)
all layers	**71.02(5.93)**	61.23(11.3)	**56.58(13.3)**

Stability. As depicted in Table 1, the Dice standard deviation is significantly lower for the semi-supervised model, especially for the smaller tumors (i.e., ET and TC) when sufficient data is available.

Effectiveness of Multi-scale Consistency. To further explore the impact of performing consistency regularization at different layers, we quantitatively compare our proposed framework when consistency regularization is applied at different layers (i.e., 0; 1; $K+1$; $0, 1, (K+1)$; and all layers). Here, $k = 0$ corresponds to the final decoder-layer (i.e., f_0), $K+1$ corresponds to the first decoder-layer (i.e., the bottleneck layer), and $0, 1, (K+1)$ means all three of these layers combined. As shown in Table 2, consistency training on all layers brought the best performance, which suggests the importance of multi-scale consistency training. Interestingly, applying CR to the bottleneck layer achieves higher performance than applying CR to the first or second layer on WT. This makes sense because consistency training at lower resolutions should assist identification of larger areas (i.e., the whole tumor). Likewise, applying CR to the final layer achieves higher performance than applying CR to the other layers for ET and TC, which makes sense because the final layer has the highest resolution and these classes correspond to smaller regions. Our proposed multi-scale consistency training reaps the benefits of both extremes.

Comparison with Related Works. In this section, we compare our method with previous works on the WMH dataset when only 3/48 labeled examples are used for training. In addition to the baselines mentioned in Section 3.2, multi-scale MT is compared with our proposed method. Table 3 shows that MuST outperforms baselines when trained using same supervision as well as multi-scale MT. In fact, outperforming multi-scale MT confirms the importance of feature-space perturbation. Furthermore, Table 3 shows the effectiveness of sharpening the teacher's outputs. The mean Dice score has improved by 1.67 %, while Dice standard deviation has decreased by 1.86 points (i.e., increased stability) when sharpening temperature of $T = 0.8$ is utilized compared to when no sharpening has been performed.

Table 3: Comparison with related works. T is the temperature in Eq. (2). L and U are number of labeled and unlabeled examples.

Method	**L (U)**	**mean(std)**
MASSL	3(45)	56.66 (7.84)
MT	3(45)	51.11(8.5)
Multi-scale MT	3(45)	51.84(4.7)
Ours (no sharpening)	3(45)	65.11(8.62)
Ours (T = 0.5)	3(45)	65.05(7.98)
Ours (T = 0.6)	3(45)	65.24(7.95)
Ours (T = 0.8)	3(45)	**66.78(6.76)**
Ours (T = 1)	3(45)	65.91(7.19)

4 Conclusion

In this paper we propose MuST, a novel semi-supervised framework for medical image segmentation. This framework exploits consistency regularization technique at multi-scales to leverage unlabeled data as well as available labeled data. Particularly, feature-space perturbations are used to produce different versions of the same input at feature-space in order to enforce consistency on all decoders independently. The proposed framework is evaluated on two public datasets for brain tumor and white matter hyper intensity segmentation tasks.

References

1. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2020)
2. 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. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 253–260. Springer (2017)
3. Bdair, T., Wiestler, B., Navab, N., Albarqouni, S.: Roam: Random layer mixup for semi-supervised learning in medical imaging. arXiv preprint arXiv:2003.09439 (2020)
4. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. Advances in Neural Information Processing Systems **32** (2019)
5. Chen, S., Bortsova, G., García-Uceda Juárez, A., Tulder, G.v., Bruijne, M.d.: Multi-task attention-based semi-supervised learning for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 457–465. Springer (2019)
6. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2613–2622 (2021)

7. Cui, W., Liu, Y., Li, Y., Guo, M., Li, Y., Li, X., Wang, T., Zeng, X., Ye, C.: Semi-supervised brain lesion segmentation with an adapted mean teacher model. In: International Conference on Information Processing in Medical Imaging. pp. 554–565. Springer (2019)
8. French, G., Laine, S., Aila, T., Mackiewicz, M., Finlayson, G.: Semi-supervised semantic segmentation needs strong, varied perturbations. arXiv preprint arXiv:1906.01916 (2019)
9. Grandvalet, Y., Bengio, Y., et al.: Semi-supervised learning by entropy minimization. CAP **367**, 281–296 (2005)
10. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al.: Recent advances in convolutional neural networks. Pattern Recognition **77**, 354–377 (2018)
11. Kuijf, H.J., Biesbroek, J.M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M.J., Casamitjana, A., Collins, D.L., Dadar, M., Georgiou, A., Ghafoorian, M., Jin, D., Khademi, A., Knight, J., Li, H., Lladó, X., Luna, M., Mahmood, Q., McKinley, R., Mehrtash, A., Ourselin, S., Park, B.Y., Park, H., Park, S.H., Pezold, S., Puybareau, E., Rittner, L., Sudre, C.H., Valverde, S., Vilaplana, V., Wiest, R., Xu, Y., Xu, Z., Zeng, G., Zhang, J., Zheng, G., Chen, C., van der Flier, W., Barkhof, F., Viergever, M.A., Biessels, G.J.: Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. IEEE Transactions on Medical Imaging **38**(11), 2556–2568 (2019). https://doi.org/10.1109/TMI.2019.2905770
12. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 (2016)
13. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3, p. 896 (2013)
14. Li, X., Sun, Q., Liu, Y., Zhou, Q., Zheng, S., Chua, T.S., Schiele, B.: Learning to self-train for semi-supervised few-shot classification. Advances in Neural Information Processing Systems **32**, 10276–10286 (2019)
15. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
16. Monteiro, M., Newcombe, V.F., Mathieu, F., Adatia, K., Kamnitsas, K., Ferrante, E., Das, T., Whitehouse, D., Rueckert, D., Menon, D.K., et al.: Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head ct using deep learning: an algorithm development and multicentre validation study. The Lancet Digital Health **2**(6), e314–e322 (2020)
17. Ouali, Y., Hudelot, C., Tami, M.: Semi-supervised semantic segmentation with cross-consistency training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12674–12684 (2020)
18. Park, G., Hong, J., Duffy, B.A., Lee, J.M., Kim, H.: White matter hyperintensities segmentation using the ensemble u-net with multi-scale highlighting foregrounds. NeuroImage **237**, 118140 (2021)
19. Peng, J., Wang, P., Desrosiers, C., Pedersoli, M.: Self-paced contrastive learning for semi-supervised medical image segmentation with meta-labels. Advances in Neural Information Processing Systems **34** (2021)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)

21. Seibold, C., Kleesiek, J., Schlemmer, H.P., Stiefelhagen, R.: Self-guided multiple instance learning for weakly supervised thoracic diseaseclassification and localizationin chest radiographs. In: Proceedings of the Asian Conference on Computer Vision (2020)
22. Sharma, N., Aggarwal, L.M.: Automated medical image segmentation techniques. Journal of medical physics/Association of Medical Physicists of India **35**(1), 3 (2010)
23. Tang, Y., Cao, Z., Zhang, Y., Yang, Z., Ji, Z., Wang, Y., Han, M., Ma, J., Xiao, J., Chang, P.: Leveraging large-scale weakly labeled data for semi-supervised mass detection in mammograms. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3854–3863 (2021). https://doi.org/10.1109/CVPR46437.2021.00385
24. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (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 radiation-induced pulmonary fibrosis from lung ct scans with multi-scale guided dense attention. IEEE transactions on medical imaging (2021)
26. Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z.: Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: International conference on medical image computing and computer-assisted intervention. pp. 408–416. Springer (2017)
27. Zhou, S.K., Greenspan, H., Davatzikos, C., Duncan, J.S., Van Ginneken, B., Madabhushi, A., Prince, J.L., Rueckert, D., Summers, R.M.: A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE (2021)

A Appendix

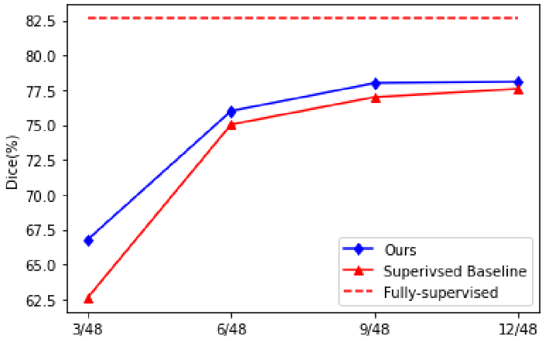


Fig. 2: Comparison of proposed semi-supervised framework with supervised baseline on WMH Challenge dataset. The proposed method achieves 4%, 1%, 1%, and 0.5% improvement over the supervised-baseline when using only 3, 6, 9, and 12 labeled examples out of 48 training examples.

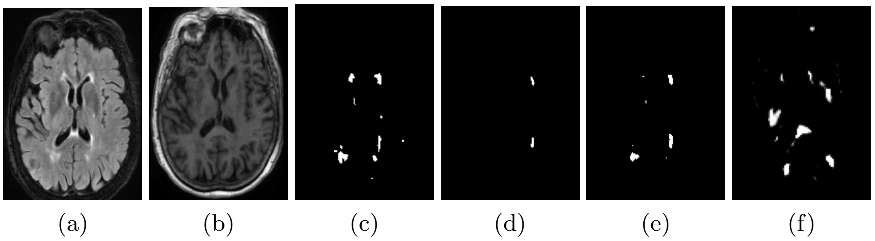


Fig. 3: Visualization of our method, supervised baseline, and related work on WMH challenge dataset based on using 8 labeled examples. (a) FLAIR, (b) T1 MRI modality, (c) Ground truth, (d) Supervised-baseline, (e) Our method, and (f) MASSL

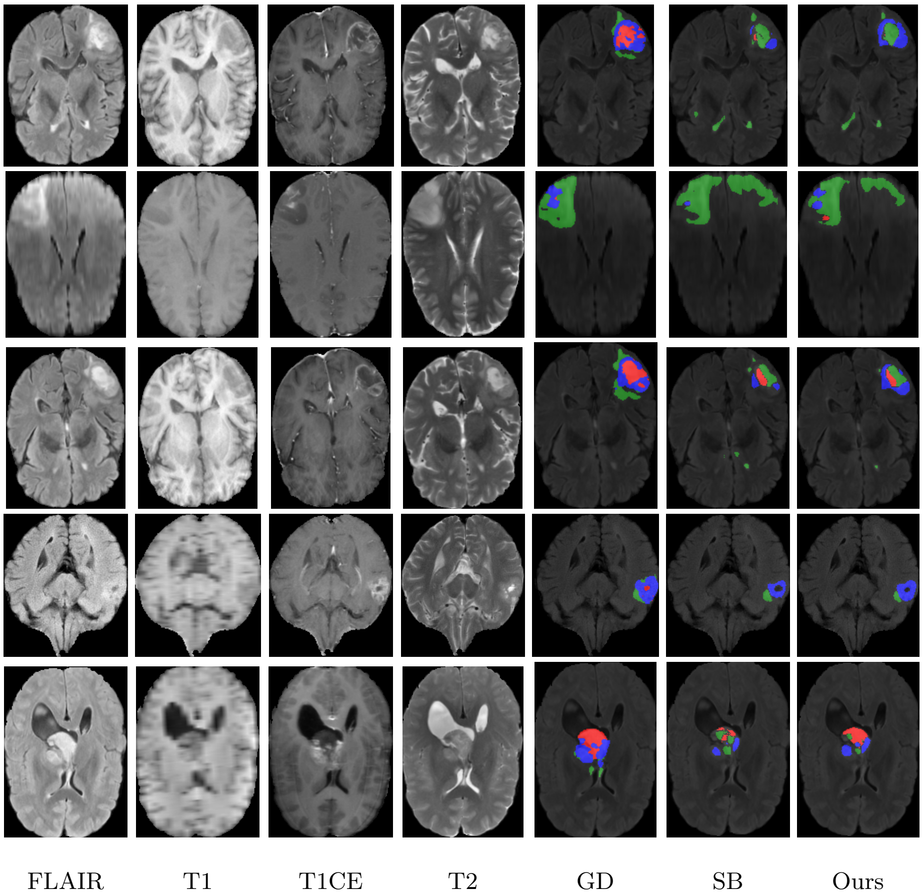


Fig. 4: Visualization of our method and supervised-baseline (SB) on BraTS dataset when trained using 8 and 14 labeled examples out 285 training examples. GD is the ground truth. Red is NET, green is ED, and blue is ET.