STRUDEL: Self-training with Uncertainty Dependent Label Refinement Across Domains

Fabian Gröger[1,2], Anne-Marie Rickmann[1(✉)], and Christian Wachinger[1]

[1] Artificial Intelligence in Medical Imaging (AI -Med), KJP, LMU München, Munich, Germany
arickman@med.lmu.de

[2] Computer Aided Medical Procedures, Technische Universität München, Munich, Germany

Abstract. We propose an unsupervised domain adaptation (UDA) approach for white matter hyperintensity (WMH) segmentation, which uses Self-TRaining with Uncertainty DEpendent Label refinement (STRUDEL). Self-training has recently been introduced as a highly effective method for UDA, which is based on self-generated pseudo labels. However, pseudo labels can be very noisy and therefore deteriorate model performance. We propose to predict the uncertainty of pseudo labels and integrate it in the training process with an uncertainty-guided loss function to highlight labels with high certainty. STRUDEL is further improved by incorporating the segmentation output of an existing method in the pseudo label generation that showed high robustness for WMH segmentation. In our experiments, we evaluate STRUDEL with a standard U-Net and a modified network with a higher receptive field. Our results on WMH segmentation across datasets demonstrate the significant improvement of STRUDEL with respect to standard self-training.

1 Introduction

Dementia presents a highly relevant societal challenge due to the ever-aging population. Research shows that aging-related structural and functional changes in the brain may manifest as cerebral small vessel disease (SVD), which is a major contributor to the risk of developing dementia [16]. A promising neuroimaging biomarker for SVD are white matter hyperintensities (WMHs) of presumed vascular origin. WMHs are visible in fluid-attenuated inversion recovery (FLAIR) magnetic resonance imaging (MRI) as diffuse regions of brighter intensity than surrounding white matter [14]. Convolutional neural networks (CNNs) have achieved remarkable performances for WMH segmentation [11]. However, CNNs are highly dependent on the training set and the performance can steeply

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-87589-3_32) contains supplementary material, which is available to authorized users.

C. Lian et al. (Eds.): MLMI 2021, LNCS 12966, pp. 306–316, 2021.
https://doi.org/10.1007/978-3-030-87589-3_32

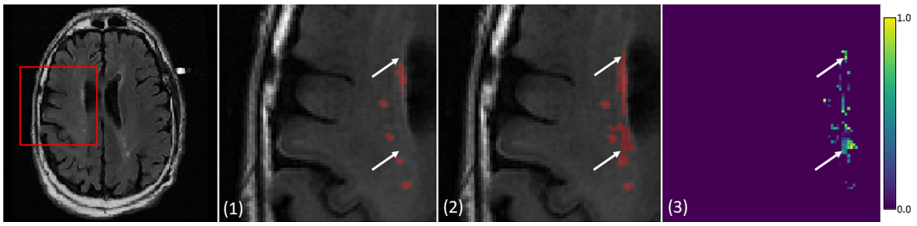

Fig. 1. Illustration of a FLAIR scan with (1) ground truth WMH, (2) noisy pseudo labels, and (3) corresponding uncertainty map. White arrows point to false positive predictions with higher uncertainty values (brighter pixels).

decrease on a target domain with a large domain shift. In a recent comparison, this resulted in traditional segmentation software producing higher quality WMH labels than CNNs [24]. Unsupervised domain adaptation (UDA) attempts to overcome the problem of domain shift without using target data annotations. Not relying on target annotations is a key benefit, as medical annotations are typically scarce and therefore will not cover the wide range of acquisition protocols, scanner types, artifacts, or patient statistics that make up domain differences in MRI. Self-training is a recent approach for UDA, where a segmentation model is first trained on annotated source data and then applied on target data to infer pseudo labels. These self-generated pseudo labels are then integrated into the network training to achieve the domain adaptation. However, pseudo labels tend to be noisy, as illustrated in Fig. 1, which necessitates estimating the reliability of pseudo labels to avoid propagating label errors.

In this work, we propose STRUDEL, a Self-TRaining approach with Uncertainty DEpendent Label refinement. It is motivated by earlier work on brain lesion segmentation [12], which demonstrated that uncertainty measures are an indicator for erroneous pixel-wise predictions. Following a Bayesian segmentation approach, we estimate the uncertainty for pseudo labels, see Fig. 1, and then integrate it in successive model refinements with an uncertainty-guided loss function. To further improve the initial pseudo label generation in STRUDEL, we propose to integrate the output of the lesion prediction algorithm (LPA) [19], which was reported to achieve robust results across domains [24]. In our experiments, we evaluate STRUDEL with a U-Net as the backbone and a modified network with a higher receptive field. Our results on WMH segmentation across datasets demonstrate the necessity for domain adaptation and further the significant improvement for integrating uncertainty and LPA in the training process.

1.1 Related Work

White Matter Hyperintensity Segmentation methods have recently been assessed in the WMH challenge [11]. All top-ranking methods were deep-learning-based and some have achieved superior performance to human observers. Specific inter-scanner robustness experiments showed there is still a need for improving the robustness of these methods which coincides with the findings in [24].

Unsupervised Domain Adaptation (UDA) approaches transfer a model from a source domain without direct supervision on the target domain and are commonly based on adversarial learning. These types of methods attempt to learn domain invariant features by minimizing the discrepancy between source and target domain or convert images from one domain to the other. Different approaches have demonstrated their effectiveness for medical applications [9,10]. However, the training process for adversarial networks can be a multi-faceted and complex endeavor [1,15].

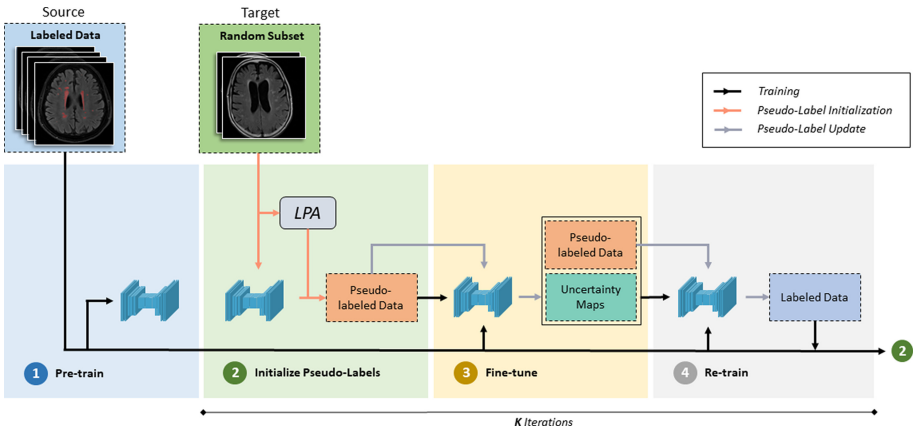

Fig. 2. Illustration of our Self-training pipeline for domain adaptation.

Self-training presents an alternative approach to UDA, which has recently been shown to be highly efficient [28]. It follows the principle that predictions generated in previous steps are used as pseudo labels for the next stage of network learning. The literature has addressed self-training for semantic segmentation in non-medical and medical applications and showed state-of-the-art performance on benchmark datasets [13,20,25,26,29,30]. A further categorization of methods for handling limited dataset annotations is available in the review by Tajbakhsh et al. [21]. The potential of integrating uncertainty guidance in self-training was recently demonstrated for segmenting sparsely annotated micro-CT scans [27].

2 Methods

2.1 Problem Definition

Given a labeled dataset from the source domain S with samples $X^S = \{x_i^S\}_{i=1}^N$ and labels $Y^S = \{y_i^S\}_{i=1}^N$, and an unlabeled dataset from the target domain T with samples $X^T = \{x_i^T\}_{i=1}^M$, the goal is to predict labels in the target domain. We want to achieve this goal by incorporating the large number of unlabeled target samples in the network training, where typically $M > N$. In self-training, this is achieved by inferring pseudo target labels $\tilde{y}^T$ and updating them iteratively to improve the label quality and consequently the learning process.

2.2 STRUDEL: Self-training with Uncertainty

Figure 2 provides a graphical overview of our proposed Self-TRaining with Uncertainty DEpendent Label refinement, with the pseudo-code in Algorithm 1. First, we pre-train a base model on the source dataset $\left(X^S, Y^S\right)$ with standard supervised learning. The base model is then applied to a random subset (drawn without replacement) of the target sample, $r(X^T)$, of size P to infer pseudo labels. However, these pseudo labels will initially not be of high quality due to the domain gap. Consequently, we propose to leverage existing WMH segmentation software, where we use LPA, to increase the quality of pseudo labels. To this end, we apply a pixel-wise OR operator between base model predictions and LPA prediction to obtain the pseudo target labels $\tilde{Y}^T = \{\tilde{y}_i^T\}_{i=1}^P$.

Algorithm 1: Self-Training with Uncertainty on Noisy Labels

input : Source data X^S, Source labels Y^S, Target data X^T
output: Output model $\mathcal{M}_K$

$r() \leftarrow$ random sampler;
$\mathcal{M}_0 \leftarrow$ train base model with (X^S, Y^S);
$D_{\text{fix}} \leftarrow (X^S, Y^S)$; `// initialize fixed training set`
for $k \leftarrow 1$ **to** K **do**
  $X_k^T \leftarrow r(X^T)$; `// sample random subset`
  $\tilde{Y}_k^T \leftarrow pixel-wise\_or(\mathcal{M}_{k-1}(X_k^T), LPA(X_k^T))$; `// init. pseudo labels`
  $D_k \leftarrow D_{\text{fix}} \cup (X_k^T, \tilde{Y}_k^T)$; `// merge training data`
  $\mathcal{M}_{k-1} \leftarrow$ fine-tune $\mathcal{M}_{k-1}$ with D_k;
  $\tilde{Y}_k^T, U_k \leftarrow \mathcal{M}_{k-1}(X_k^T)$; `// update labels and get uncertainty`
  $D'_k \leftarrow D_{\text{fix}} \cup (X_k^T, \tilde{Y}_k^T)$; `// merge training data`
  $\mathcal{M}_k \leftarrow$ re-train model with D'_k and uncertainty U_k;
  $D_{\text{fix}} \leftarrow D_{\text{fix}} \cup (X_k^T, \mathcal{M}_k(X_k^T))$; `// update fixed training set`
end
return $\mathcal{M}_K$;

In a third step, the base model is fine-tuned with $\tilde{Y}^T$. With this model, we segment again the same random sample drawn earlier, $r(X^T)$, producing pseudo labels of higher quality. Here, we assume that a model can generate better predictions than its noisy training labels [6]. Next to the labels, we also infer the segmentation uncertainty U at this stage. Finally, a new model is trained from scratch, where we add the updated pseudo labels $\tilde{Y}^T$ with the corresponding uncertainty U to the training set. Training a new model at this point has advantages over fine-tuning as also noted in [28]. The labels inferred from this model on the random subset are then added to the fixed training set, which initially only consists of the annotated source data. We then continue with the next iteration in step 2, where this model serves as a base model, another random subset is sampled and pseudo labels are inferred.

2.3 Uncertainty-Guided Pseudo Labels

Inferring pseudo target labels usually has the disadvantage of label noise. To increase the robustness of our method against label noise, we propose an uncertainty-guidance that strengthens regions of low uncertainty and penalizes regions of high uncertainty. To estimate the uncertainty, we follow a Bayesian machine learning approach with an estimation of Monte Carlo (MC) samples by dropout [5]. Accordingly, we train the backbone segmentation network with dropout layers and perform C stochastic forward passes at test time to obtain Monte Carlo samples. The expectation over the MC samples $\mathbb{E}(\hat{y})$ provides us a more robust label prediction, which we use to update the pseudo-label. Further, computing the variance across C MC samples gives us a pixel-wise measure of uncertainty of the predicted segmentation:

$$U(\hat{y}) = \{\sigma_1, ..., \sigma_{H \times W}\} = \frac{1}{C} \sum_{i=1}^{C} \left(\widehat{y}_i - \mathbb{E}(\hat{y})\right))^2, \tag{1}$$

where $U(\hat{y})$ denotes the uncertainty map, σ_i the pixel-wise variance, H, W the images height and width, and $\hat{y}_i$ the model prediction from the ith MC sample. Anticipating small values for the uncertainties, we rescale the values into the range [0, 1]. We integrate the uncertainty into network training by the definition of an uncertainty-aware binary cross entropy (UBCE) loss:

$$\mathcal{L}_{\text{UBCE}} = -\frac{1}{H \times W} \sum_{n=1}^{H \times W} (1 - \sigma_n) \left[\tilde{y}_n \cdot log\left(\hat{y}_n\right) + (1 - \tilde{y}_n) \cdot log\left(1 - \hat{y}_n\right)\right]. \tag{2}$$

Note that the uncertainty-aware cross entropy $\mathcal{L}_{\text{UBCE}}$ is only applied to pseudo-labeled data within the re-training step (see Algorithm 1 line 11), whereas the standard cross entropy $\mathcal{L}_{\text{BCE}}$ is applied to fixed data samples. The combined loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{UBCE}}. \tag{3}$$

2.4 Segmentation Backbone Architectures

For the segmentation model M, we evaluate two neural network architectures. First, the U-Net [17], which has proven its performance in difficult segmentation tasks and has been adapted and improved for various applications since then. Second, to explore the effects of a superior model architecture within the framework, we include a novel network architecture, the OctSE-Net, which introduces two modifications to the U-Net. The first modification is to replace all convolution layers with octave convolutions [4], which factorize feature maps by their frequencies. Octave convolutions can increase segmentation performance, by offering a wider context, while reducing memory consumption. The second modification is to integrate squeeze & excitation (SE) blocks [8], more precisely the channel and spatial SE block [18], which can boost accuracy by re-calibrating

feature maps. Both modifications increase the receptive field without substantially increasing model parameters. This can improve the segmentation accuracy without encouraging overfitting and therefore help the generalization across domains. For uncertainty estimation, we insert dropout layers after each convolutional block in both architectures.

3 Experiments and Results

3.1 Datasets

As source dataset, we use data from the WMH challenge (https://wmh.isi.uu.nl), and, as target dataset, we use data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (http://adni.loni.usc.edu). The WMH segmentation challenge dataset provides manual annotations for 60 subjects from 3 sites. For each subject, co-registered 3D T1-weighted and 2D multi-slice FLAIR scans are available, where we work with bias-field corrected T1 scans and original FLAIR scans as suggested by [7,22]. The large ADNI-2 dataset [3] with over 3,000 scans from 58 sites serves as our multi-domain target data. For each subject, T1-weighted and 2D FLAIR scans are available, which we have linearly aligned within a session with ANTs [2]. T1 scans have further been bias field corrected with N4 normalization [23], using the ANTs implementation. To quantitatively evaluate our methods on the ADNI dataset, we extracted a subset of 30 subjects based on scanner type and WMH lesion load for manual annotation[1]. 21 of these annotated scans serve solely as a test set and 9 are used to train alternative segmentation approaches, described in the next section.

3.2 Implementation Details

We implemented the proposed framework and all baseline experiments in Pytorch v.1.6.0 (see footnote 1). Experiments were performed employing the Adam optimizer with default parameters (betas = (0.9, 0.999), eps = 1e-08), learning rate 1e-4 and batch size 4. Image intensities were normalized to zero-mean and unit-variance and axial slices center cropped to a consistent size of 192×192 pixels across all datasets. Standard spatial augmentation techniques (flipping, rotation, scaling, and elastic transformation) were used during training for regularization. In self-training, the size of the random subset per iteration is set to $P = 35$. The thresholds to obtain binary segmentation maps for the creation of pseudo-labels were set to 0.5 and 0.75 for the network prediction and LPA, respectively. The high threshold for LPA mitigates hypersensitive responses. We use 80 epochs for training from scratch and 20 epochs for fine-tuning. We set the number of stochastic forward passes to $C = 10$. We found that increasing C further does not improve the segmentation performance. The drop-out rate was set to 0.2. No explicit post-processing was performed in any experiment. A Geforce Titan RTX GPU was predominantly used for training and testing.

[1] Code and manual segmentations are available under: https://github.com/ai-med/STRUDEL.

3.3 Experiments

We perform several experiments to evaluate the domain transfer performance of different approaches. First, as **Base Model**, we directly apply a model trained on the source data on the target domain data without any adaptation. Next, we evaluate two approaches that use a small labeled subset of the target domain during training. The **Joint Model** combines the target and source training data, and the **Fine-Tuning** model uses the labeled target data to fine-tune the Base Model. Finally, we evaluate two UDA approaches with pseudo labels: **Self-Training** without uncertainty guidance and the proposed **STRUDEL** with uncertainty guidance, both using LPA labels. We also report results for STRUDEL without LPA and for LPA itself, where we set the threshold parameter to 0.45, as suggested in [24] for ADNI. We evaluate the segmentation accuracy for all experiments by following evaluation metrics suggested by the WMH segmentation challenge [11]: (1) Dice Similarity Coefficient (DSC), (2) modified Hausdorff distance (95th percentile; H95), (3) absolute log-transformed volume difference (lAVD), (4) sensitivity for detecting individual lesions (Recall), and (5) F1-score for individual lesions (F1).

Table 1. Comparison of segmentation methods, network architectures, type of used data (S: Source manual labels, T: Target manual labels, P: target pseudo labels), and their mean performance ± standard deviation on the metrics: Dice Coefficient (DSC), 95th Percentile Hausdorff Distance (H95), log transformed absolute volume difference (lAVD), lesion Recall and F1.

Methods	S	T	P	DSC ↑	H95 [mm] ↓	lAVD ↓	Recall ↑	F1 ↑
LPA	✗	✗	✗	0.57 ± 0.16	23.1 ± 23.4	0.71 ± 0.49	0.81 ± 0.16	0.39 ± 0.18
U-Net								
Base Model	✓	✗	✗	0.45 ± 0.28	27.1 ± 37.5	1.09 ± 1.70	0.67 ± 0.32	0.48 ± 0.21
Joint Model	✓	✓	✗	0.64 ± 0.19	17.2 ± 25.0	0.60 ± 0.52	0.74 ± 0.29	0.52 ± 0.15
Fine-Tuning	✓	✓	✗	**0.73 ± 0.16**	**11.2 ± 23.0**	0.36 ± 0.41	**0.75 ± 0.22**	**0.65 ± 0.14**
Self-Training	✓	✗	✓	0.64 ± 0.20	17.8 ± 28.8	0.51 ± 0.68	0.51 ± 0.27	0.50 ± 0.23
STRUDEL	✓	✗	✓	0.69 ± 0.18	**11.2 ± 14.5**	**0.30 ± 0.32**	0.58 ± 0.27	0.64 ± 0.22
OctSE-Net								
Base Model	✓	✗	✗	0.60 ± 0.23	19.7 ± 29.5	0.77 ± 1.12	0.80 ± 0.26	0.61 ± 0.19
Joint Model	✓	✓	✗	0.73 ± 0.15	11.8 ± 24.7	0.34 ± 0.37	**0.89 ± 0.10**	0.59 ± 0.14
Fine-Tuning	✓	✓	✗	0.73 ± 0.15	11.4 ± 23.4	0.41 ± 0.38	0.77 ± 0.18	0.64 ± 0.17
Self-Training	✓	✗	✓	0.73 ± 0.13	14.7 ± 18.2	**0.25 ± 0.27**	0.56 ± 0.21	0.63 ± 0.17
STRUDEL	✓	✗	✓	**0.78 ± 0.10**	**7.79 ± 8.52**	0.27 ± 0.23	0.77 ± 0.16	**0.70 ± 0.15**
↪ w/o LPA	✓	✗	✓	0.67 ± 0.20	12.9 ± 13.4	0.63 ± 0.58	0.58 ± 0.23	0.66 ± 0.18

3.4 Results and Discussion

Table 1 reports the quantitative segmentation results on the ADNI target domain. Figure 3 shows the DSC in more details as boxplots. As a reference, the DSC of the Base Models on the source dataset are 0.73 for U-Net and 0.76 for OctSE-Net. We observe that the direct transfer of the Base Model on the ADNI dataset performs poorly, regardless of the backbone architecture. LPA beats the baseline U-Net by a large margin, which is in accordance with the results described in [24]. OctSE-Net outperforms LPA, which confirms our assumption that OctSE-Net is a more robust architecture. However, the more detailed results in Fig. 3 show that both base models produce some predictions with zero DSC, whereas LPA does not. These outliers can lead to poorly initialized pseudo labels, which is the reason for including LPA in pseudo-label initialization, confirmed by the bad results for STRUDEL w/o LPA. We report results for all self-training-based experiments after 5 iterations, as no further improvement was observed afterwards. STRUDEL outperforms all other methods in DSC and H95, and is best or second best in lAVD and lesion F1. LPA and the joint model perform best in terms of lesion recall. Both of these methods have relatively poor performance in lesion F1, which is the result of a high number of false positive predictions. STRUDEL performs strongly in both metrics, which we believe is due to the uncertainty capturing false positives reliably. Results of a Wilcoxon signed-rank test on DSC show that the improvement of STRUDEL with respect to Self-Training is significant for OctSE-Net ($p < 0.005$) and also that the improvement of OctSE-Net with respect to U-Net is significant for STRUDEL ($p < 0.001$).

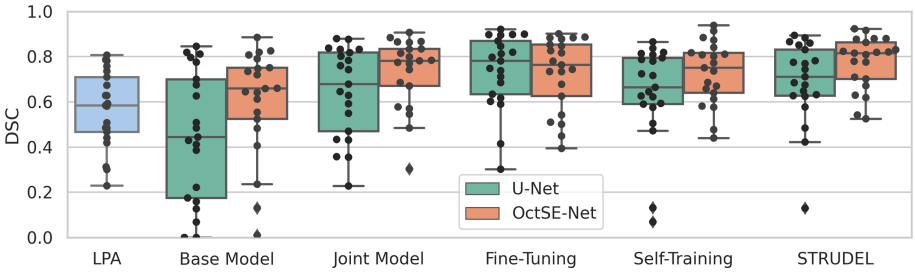

Fig. 3. Boxplot of Dice Similarity Coefficient for the different methods. Points outside the whiskers are determined as outliers based on the inter-quartile range.

4 Conclusion

Self-Training is a simple and effective approach for UDA, however noisy pseudo labels can limit its effectiveness. In this work, we proposed STRUDEL, an uncertainty-guided self-training method for unsupervised domain adaptation. We found that introducing uncertainty into the objective function can efficiently

guide the learning process in the presence of noisy labels. We further demonstrated that leveraging an existing algorithm (LPA) for pseudo label initialization can additionally boost performance. Our experimental results showed that Self-Training with uncertainty guidance is a strong approach for UDA, which in combination with a strong and robust network architecture, can even outperform supervised methods.

Acknowledgments. This research was partially supported by the Bavarian State Ministry of Science and the Arts and coordinated by the bidt, and the BMBF (DeepMentia,031L0200A).

References

1. Arjovsky, M., Bottou, L.: Towards principled methods for training generative adversarial networks. arXiv preprint arXiv:1701.04862 (2017)
2. Avants, B.B., Tustison, N., Song, G.: Advanced normalization tools (ants). Insight J. **2**(365), 1–35 (2009)
3. Beckett, L., Donohue, M., Wang, C., Aisen, P., Harvey, D., Saito, N.: The alzheimer's disease neuroimaging initiative phase 2: increasing the length, breadth, and depth of our understanding. Alzheimer's Dementia J. Alzheimer's Assoc. **11**, 823–31 (2015)
4. Chen, Y., et al.: Drop an octave: reducing spatial redundancy in convolutional neural networks with octave convolution, pp. 3434–3443 (2019)
5. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of The 33rd International Conference on Machine Learning (2015)
6. Guan, M., Gulshan, V., Dai, A., Hinton, G.: Who said what: modeling individual labelers improves classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
7. Hernández, M.D.C.V., et al.: On the computational assessment of white matter hyperintensity progression: difficulties in method selection and bias field correction performance on images with significant white matter pathology. Neuroradiology **58**(5), 475–485 (2016)
8. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018
9. Huo, Y., Xu, Z., Bao, S., Assad, A., Abramson, R.G., Landman, B.A.: Adversarial synthesis learning enables segmentation without target modality ground truth. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 1217–1220. IEEE (2018)
10. Kamnitsas, K., et al.: Unsupervised domain adaptation in brain lesion segmentation with adversarial networks, pp. 597–609 (2017)
11. Kuijf, H.J., et al.: Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. IEEE Trans. Med. Imaging **38**(11), 2556–2568 (2019)
12. Nair, T., Precup, D., Arnold, D., Arbel, T.: Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation, pp. 655–663 (2018)

13. Nie, D., Gao, Y., Wang, L., Shen, D.: ASDNet: attention based semi-supervised deep networks for medical image segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11073, pp. 370–378. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_43
14. Prins, N., Scheltens, P.: White matter hyperintensities, cognitive impairment and dementia: an update. Nature reviews. Neurology **11**, 157–165 (2015)
15. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
16. Raz, L., Knoefel, J., Bhaskar, K.: The neuropathology and cerebrovascular mechanisms of dementia. J. Cerebral Blood Flow Metabol. Official J. Int. Soc. Cerebral Blood Flow Metabol. **36**, 172–186 (2015)
17. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
18. Roy, A.G., Navab, N., Wachinger, C.: Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks. IEEE TMI **38**(2), 540–549 (2019)
19. Schmidt, P.: Bayesian inference for structured additive regression models for large-scale problems with applications to medical imaging. Ph.D. thesis, Ludwig-Maximilians-Universität München (2017)
20. Shin, I., Woo, S., Pan, F., Kweon, I.S.: Two-phase pseudo label densification for self-training based domain adaptation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12358, pp. 532–548. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58601-0_32
21. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X.: Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation. Med. Image Anal. **63**, 101693 (2020)
22. Tubi, M.A., et al.: White matter hyperintensities and their relationship to cognition: effects of segmentation algorithm. NeuroImage **206**, 116327 (2020)
23. Tustison, N.J., et al.: N4itk: improved n3 bias correction. IEEE Trans. Med. Imaging **29**(6), 1310–1320 (2010)
24. Vanderbecq, Q., et al.: Comparison and validation of seven white matter hyperintensities segmentation software in elderly patients. NeuroImage Clin. **27**, 102357 (2020)
25. Xia, Y., et al.: 3D semi-supervised learning with uncertainty-aware multi-view co-training. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3646–3655 (2020)
26. Yu, L., Wang, S., Li, X., Fu, C.-W., Heng, P.-A.: Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 605–613. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_67
27. Zheng, H., et al.: Cartilage segmentation in high-resolution 3D Micro-CT images via uncertainty-guided self-training with very sparse annotation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 802–812. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_78
28. Zoph, B., et al.: Rethinking pre-training and self-training. In: Advances in Neural Information Processing Systems, vol. 33 (2020)

29. Zou, Y., Yu, Z., Kumar, B.V., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 289–305 (2018)
30. Zou, Y., Yu, Z., Liu, X., Kumar, B.V., Wang, J.: Confidence regularized self-training. In: The IEEE International Conference on Computer Vision (ICCV), October 2019