Supplementary Material for: Graph-constrained Contrastive Regularization for Semi-weakly Volumetric Segmentation

Simon Reiß[1], Constantin Seibold[1], Alexander Freytag[2], Erik Rodner[3], and Rainer Stiefelhagen[1]

[1] Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany
[2] Carl Zeiss AG, 07745 Jena, Germany
[3] University of Applied Sciences Berlin, 12459 Berlin, Germany
{simon.reiss, constantin.seibold, rainer.stiefelhagen}@kit.edu,
alexander.freytag@zeiss.com, erik.rodner@htw-berlin.de

Abstract. In this supplementary document we complement the approach and experimental settings in the main paper. We outline additional implementation details, summarize how we adapt semi-supervised methods for our semi-weakly scenario and perform ablation studies for them in order to have strong competing models for testing our proposed approach. Finally, we present qualitative insights regarding our methods learned voxel-embeddings.

Keywords: Volumetric semantic segmentation, semi-weakly supervised learning, regularization, contrastive learning

1 Implementation details

In our experiments we use the tensor-processing framework PyTorch [6] for training models. Specifically, we base our implementation [1] for 3D UNet training on an implementation of 3D UNet in PyTorch which can be found online [2]. The basic building blocks of the UNet as we use it are a sequence of groupnorm with 8 groups [12], convolution with kernel size 3 and padding 1 and a final ReLU activation. These building blocks are used twice in each UNet layer of which we have 4, with channels as specified in the main paper $(64, 128, 256, 256)$.

2 Baseline ablation studies

Next, we will go into detail how we adapted the baseline semi-supervised methods to the semi-weakly supervised scenario as well as what we did for adapting 2D to 3D methods. To make sure that we compare our method against strong

[1] https://github.com/Simael/Con2R
[2] https://github.com/wolny/pytorch-3dunet/blob/master/pytorch3dunet/unet3d/model.py

baseline models, we performed ablation studies and hyperparameter tuning on the competing semi-supervised methods. Therefore we will now outline some of the ablation experiments and implementation details regarding the methods Pseudo-label [5], Mean-Teacher [11], FixMatch [10] and Uncertainty-aware Mean-Teacher [13]. Similarly to what we present in the main paper regarding *Con2R*-related ablation studies, we carry out all experiments on the RETOUCH dataset [2] in the 24 annotations setting.

2.1 Pseudo-label

For the pseudo-labeling method, we found the results to severely degrade the segmentation results when naively implementing [5]. Online pseudo-labeling, where the segmentation model which is currently optimized in training is used to generate pseudo-labels for the unlabeled samples led to models diverging completely. Thus, we implement an offline approach using a fixed pre-trained model to provide pseudo-labels. We noticed when we integrate the loss normalization from [14] we started getting better results. The normalization term considers the loss on labeled L_h and the loss on pseudo-labeled data L_p and weights these losses by:

$$\hat{L} = \frac{1}{1+\alpha}(L_h + \alpha \frac{\bar{L_h}}{\bar{L_p}} \cdot L_p) \ , \tag{1}$$

where $\bar{L_h}$ and $\bar{L_p}$ refer to exponential moving averages over the respective losses. We set $\alpha = 4.0$ as specified in [14] for the best performance.

2.2 Mean-Teacher

As described in the paper, for Mean-Teachers in segmentation scenarios, if we intend to use geometric augmentation (*e.g.* flipping) we have to reverse this augmentation when enforcing the consistency loss between pixel-wise predictions of the teacher and the student model. To accomplish this, we align the teachers predictions to the predictions of the student. We tested different values for the exponential moving average decay factor and in line with previous work [8] find a factor of 0.5 to make the Mean-Teacher approach work.

2.3 Uncertainty-aware Mean-Teacher

We also implement the Uncertainty-aware Mean-Teacher [13]. As our implementations are all based on the 3D UNet architecture, we simply integrate the dropout layers needed for Monte-Carlo [4] sampling-based uncertainty calculations into our architecture. Therefore, we integrate dropout layers directly after the encoder as well as before the final pixel-wise classification layer. The dropout probability is kept as in the original paper at 50%. Similarly to the Mean-Teacher in Section 2.2, we leverage an exponential moving average smoothing factor equal to 0.5. As we initialize all semi-supervised models with weights pre-trained on the

Table 1. Uncertainty-aware Mean-Teacher results with different thresholds

threshold	val mIoU
$\gamma = 0.1$	42.98 ± 5.06
$\gamma = 0.3$	42.96 ± 5.44
$\gamma = 0.5$	44.45 ± 4.06
$\gamma = 0.7$	41.98 ± 7.13

Table 2. Uncertainty-aware Mean-Teacher results with different numbers of stochastic forward passes in Monte-Carlo dropout

threshold	val mIoU
4 forwardpasses	44.16 ± 5.89
8 forwardpasses	44.45 ± 4.06

Table 3. Uncertainty-aware Mean-Teacher results with different dropout variants

threshold	val mIoU
classical dropout	44.45 ± 4.06
3D feature dropout	39.50 ± 5.10

available subset of annotations (*i.e.* the 3D UNet Baseline models), we omit the gaussian scheduling as well as the successive up-weighting of the semi-supervised loss term as specified in the implementation details of [13].

Further, we ablate which threshold γ for the voxel-wise uncertainty values U should be used in training in Table 1. The thresholding of the uncertainty of the Uncertainty-aware Mean-Teacher models selects confident enough portions of the volume predictions via $U < U_{\max} \cdot \gamma$ and uses those portions for the consistency regularization loss term. We find that a threshold of $\gamma = 0.5$ to work best. Calculating uncertainties with Monte-Carlo dropout has several hyperparameters, we compare using different numbers of stochastic forward passes in Tab. 2 and two dropout variants in Tab. 3 (classical stochastic dropout vs. dropping out full channels, *i.e.* zeroing out entire 3D features). We opt for 8 forward passes and classical stochastic dropout.

2.4 FixMatch

Generally, we train FixMatch [10] with a standard-cross entropy loss between the pseudo-labels derived from weakly augmented volumes and the predictions as obtained from strongly augmented volumes as input. Weak augmentation strategies are simple flipping operations, which we ablate in Tab. 4. There we see that using a conservative weak augmentation scheme of only flipping volumes in longitudinal- and vertical direction with a probability of 50% gave the best performance.

For the strong augmentation strategy we turn our attention to photometric augmentations in Tab. 5, which we add on top of the flipping augmentations. We tested three photometric perturbations: adjusting brightness, adjusting the gamma value and the sharpness, as provided in the PyTorch vision library [7]. Our experiments show, that FixMatch with brightness and sharpness perturba-

Table 4. FixMatch results in mean IoU and standard deviation when changing the weak augmentation strategy, only applying subset of flip augmentations

horizontal	vertical	longitudinal	val mIoU
✓	✓	✓	46.05 ± 5.03
-	✓	✓	47.17 ± 4.13
✓	-	✓	46.35 ± 6.72
-	-	✓	45.94 ± 6.00
-	-	-	43.06 ± 6.32

Table 5. FixMatch results in mean IoU and standard deviation when varying the photometric augmentation strategy

brightness	gamma	sharpness	val mIoU
✓	-	-	45.78 ± 7.75
-	✓	-	44.12 ± 8.14
-	-	✓	45.20 ± 8.06
✓	✓	-	28.38 ± 20.04
✓	-	✓	46.05 ± 5.03
-	✓	✓	42.47 ± 7.03
✓	✓	✓	23.12 ± 19.95

tions using a magnitude sampled uniformly from an interval of $[0, 2]$ as choice for strong augmentations work best. Further, in the original FixMatch publication CutOut [3] is used for semi-supervised classification as strategy for strong augmentations. We implement CutOut such that we cut out small volumes in the strongly augmented input volume and ignore the corresponding areas in the pseudo-labels. This is in line with previous implementations such as [9] which studied FixMatch for 2D segmentation. In Tab. 6 we see that best results are achieved by either not using CutOut at all or cutting out large chunks of size $16 \times 16 \times 16$. Originally, FixMatch uses a cosine annealing learning rate schedule in training. In Tab. 7 we find that with our network architecture and dataset, a simple constant learning rate of 0.01 produces better results. On the classification task for which FixMatch was originally designed, the threshold for obtaining confident enough pseudo-labels is chosen to be $\tau = 0.95$. For volumetric segmentation we find such thresholds generally to be too high and ablate the sensitivity for our setting in Tab. 8. As can be seen the best threshold is found to be $\tau = 0.5$. Finally, we also consider how down-weighting the loss for the unlabeled data by λ_u effects the performance (Tab. 9). Equal weight between the loss on labeled and pseudo-labeled data resulted in the best validation performance.

Table 6. FixMatch results when using CutOut [3], we vary the cube size that is cut out

CutOut size	val mIoU
none	47.17 ± 4.13
$4 \times 4 \times 4$	46.26 ± 5.57
$8 \times 8 \times 8$	45.72 ± 6.14
$16 \times 16 \times 16$	48.22 ± 5.65

Table 7. FixMatch results with different learning rate schedules

learning rate	val mIoU
constant $lr = 0.01$	48.22 ± 5.65
cosine	45.27 ± 5.73

Table 8. FixMatch results when tuning the Pseudo-label confidence threshold τ

confidence threshold τ	val mIoU
$\tau = 0$	45.54 ± 6.72
$\tau = 0.2$	43.84 ± 6.50
$\tau = 0.5$	46.05 ± 5.03
$\tau = 0.7$	43.74 ± 8.23
$\tau = 0.95$	43.77 ± 7.78

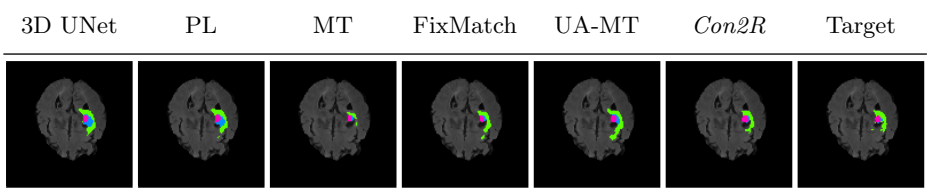
Table 9. FixMatch results with different weighting of unlabeled examples

weighting factor λ_u	val mIoU
$\lambda_u = 1.0$	46.05 ± 5.03
$\lambda_u = 0.5$	45.78 ± 6.13

3 Qualitative images for BraTS

As outlined in the main paper, our *Con2R* method did outperform the strong baseline models in all semi-weakly supervised scenarios. In Fig. 1 we can see the visual effects of tackling the task of brain tumor sub-region segmentation [1] explicitly. While all competing methods except for Mean-Teacher over segment **edema** severely, our method is even able to capture details such as the small split at the bottom of the segmentation target. Uncertainty-aware Mean-Teacher comes close to segmenting the thin **non-enhancing tumor** as well as our method, but still over segments the relevant portions.

Fig. 1. Segmentation results with 24 annotations in semi-weak brain tumor sub-region segmentation, results overlayed with first input channel of MRI scan



4 *Con2R* voxel-embedding visualization

In Fig. 2 we visualize what the voxel-wise embeddings learn when they are optimized via our graph-contrastive constraints that enforce *positional-* and *semantic proximity*. We can see, that the 64 channels in the voxel-embeddings learn very diverse and semantically relevant details such as focusing on the background, the retina deliniations, different fluid-types or different retinal layers. The input B-scan and associated target B-scan can be seen in Fig. 3.

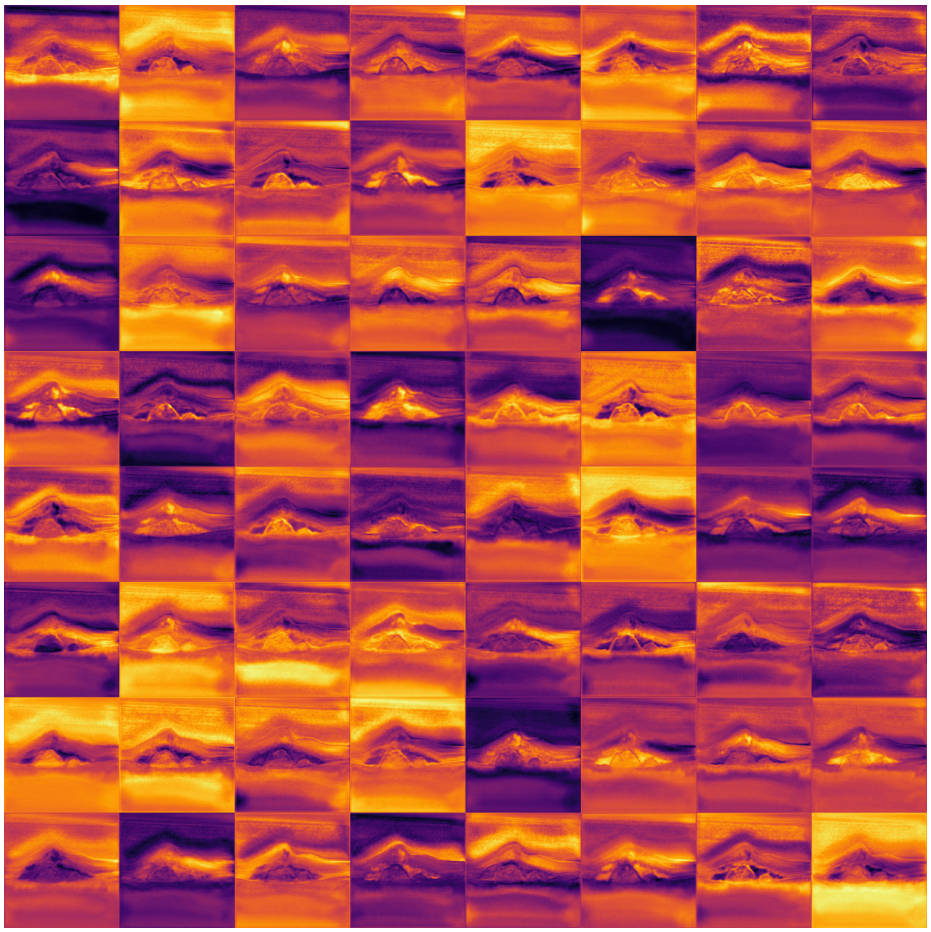


Fig. 2. 64 channels in the learned voxel-embeddings, for visualization purposes the values of one embedded slice $\in \mathbb{R}^{64\times W\times H}$ were feature-scaled channel-wise

Fig. 3. Input OCT B-scan alongside the associated ground-truth segmentation used to visualize the voxel-embeddings

References

1. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., van Ginneken, B., et al.: The medical segmentation decathlon. arXiv preprint arXiv:2106.05735 (2021) 5
2. Bogunović, H., Venhuizen, F., Klimscha, S., Apostolopoulos, S., Bab-Hadiashar, A., Bagci, U., Beg, M.F., Bekalo, L., Chen, Q., Ciller, C., et al.: Retouch: the retinal oct fluid detection and segmentation benchmark and challenge. IEEE transactions on medical imaging **38**(8), 1858–1874 (2019) 2
3. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017) 4, 5
4. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016) 2
5. 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) 2
6. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017) 1
7. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32**, 8026–8037 (2019) 3
8. Reiß, S., Seibold, C., Freytag, A., Rodner, E., Stiefelhagen, R.: Every annotation counts: Multi-label deep supervision for medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9532–9542 (2021) 2
9. Seibold, C., Reiß, S., Kleesiek, J., Stiefelhagen, R.: Reference-guided pseudo-label generation for medical semantic segmentation. arXiv preprint arXiv:2112.00735 (2021) 4
10. 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. arXiv preprint arXiv:2001.07685 (2020) 2, 3
11. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint arXiv:1703.01780 (2017) 2

12. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018) 1
13. 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: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 605–613. Springer (2019) 2, 3
14. Zoph, B., Ghiasi, G., Lin, T.Y., Cui, Y., Liu, H., Cubuk, E.D., Le, Q.V.: Rethinking pre-training and self-training. arXiv preprint arXiv:2006.06882 (2020) 2