Active Learning for Scribble-based Diffusion MRI Segmentation*

Jonathan Lennartz[1,2][0009−0003−0827−943X] and Golo Pohl[1][0009−0002−5636−2767] and Thomas Schultz[1,2][0000−0002−1200−7248]

[1] University of Bonn, Germany
[2] Lamarr Institute for Machine Learning and Artificial Intelligence

Abstract. Scribbles are a popular form of weak annotation for the segmentation of three-dimensional medical images, but typically require iterative refinement to achieve the desired segmentation map. The complexity of diffusion MRI (dMRI) poses additional challenges. Previous work addressed the high dimensionality of dMRI via unsupervised representation learning, and combined it with a random forest classifier that can be re-trained quickly enough to provide interactive feedback to the human annotator. Our work extends that framework in multiple ways. Our main contribution is to add an active learning component that suggests locations in which additional scribbles should be placed. It relies on uncertainty quantification via test time augmentation (TTA). Second, we observe that TTA increases segmentation accuracy even by itself. Moreover, we demonstrate that anomaly detection via isolation forests effectively suppresses false positives that arise when generalizing from sparse scribbles. Taken together, these contributions substantially improve the accuracy that can be achieved with various annotation budgets.

Keywords: Scribble-based Segmentation · Active Learning · Anomaly Detection.

1 Introduction

Voxel-wise annotation of 3D images is a huge effort when done fully manually. A popular alternative is to let a human annotator provide a few examples of the relevant classes in the form of sparse scribbles on individual slices within the volume, train a segmentation method to generalize them to the rest of the volume, and iteratively add scribbles in regions where this did not yet produce the correct classification. Deep neural networks represent the state of the art for image segmentation, and have shown a good ability to generalize from sparse

* Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 422414649. Data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University

annotations in 3D images [2], but their training is typically too slow to provide interactive feedback. Therefore, an alternative framework has been proposed in which a convolutional neural network is used for unsupervised representation learning, and the interactive session relies on a fast random forest classifier [10].

In particular, that framework addressed the segmentation of diffusion MRI, which probes tissue microstructure in the brain with a large number of diffusion-weighted measurements per voxel. Consequently, even visualizing that data, and evaluating the proposed segmentation with respect to it, becomes a non-trivial task [9]. For this reason, it is highly desirable to have an active learning component that guides the human annotator by suggesting slices, and regions in them, in which additional scribbles are expected to increase the accuracy the most.

Our work contributes such an active learning system. It is based on uncertainty quantification via test-time augmentation, which we find to be helpful for selecting slices and voxels for annotation, but which we also find to considerably improve the baseline accuracy of the initial segmentation.

As an additional contribution, we observe that the misclassifications in our system are partly due to the fact that the sparse scribbles only cover a small part of the overall feature space. Therefore, large regions within the 3D image fall outside the distribution on which the random forest has been trained. We introduce a novel mechanism for false positive elimination (FPE) that is based on reasoning that, when considering data that is dissimilar to any of the available examples, it is safest to assume that it is part of the background. We implement this idea using anomaly detection via an isolation forest [7].

We show that taking these building blocks together permits a substantial improvement in accuracy for various annotation budgets, resulting from increasing numbers of iterative refinement. We also evaluate each ingredient of our approach individually to clarify its relative contribution to this overall benefit.

2 Related Work

Scribble-based segmentation [1] and the use of active learning for medical image segmentation [3, 4] are active topics of research. Our approach is in line with recent findings regarding the usefulness of test-time augmentation [3] and random sampling [4] for active learning. To apply them to the challenging case of interactive diffusion MRI segmentation, we implement computationally efficient variants of these ideas. To our knowledge, our idea of further improving scribble-based segmentation by anomaly detection based false positive elimination is new.

Scribble-based segmentation is related to, but not the same as promptable segmentation [6]. Even though promptable segmentation methods are becoming increasingly powerful, they are currently only trained with prompts containing a limited number of points. Specialized models for medical images are being developed, but some are still limited to bounding box prompts [8], and we are not aware of any such methods for diffusion MRI. Therefore, we consider an integration of foundation models an interesting topic for future work, but build our current system on a more classical supervised classification approach [10].

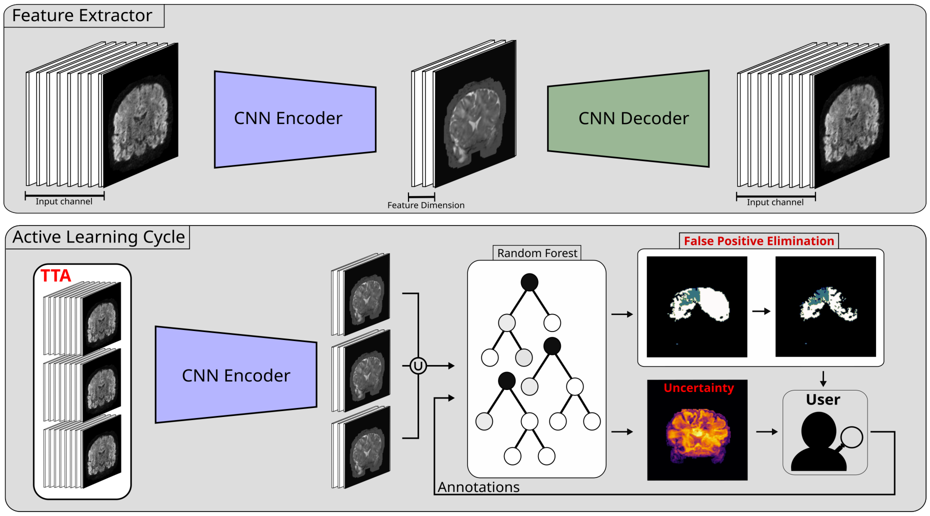

Fig. 1. Overview of our methodology with the newly introduced components highlighted in red: Test-time augmentation (TTA), false positive elimination, and uncertainty quantification for active learning.

3 Methodology

Our active learning system builds on an interactive pipeline for diffusion MRI segmentation that has been described previously [10] and that is summarized in Figure 1, with our novel contributions highlighted in red.

In an unsupervised pre-process that is depicted in the top row, it extracts a 44-dimensional feature representation for each voxel that can serve as a basis for its segmentation. It employs a dual-branch CNN whose local branch reduces the 288-dimensional data to a more compact 22-dimensional representation, and whose regional branch adds another 22 dimensions that capture spatial context. Both branches are trained jointly as an autoencoder. In the interactive loop, shown in the bottom row, a random forest classifier is trained on those features, with user-defined scribbles as labels. Within a few seconds, the remaining voxels are annotated automatically, a 3D segmentation result is visualized, and additional scribbles can be placed to correct it. Our work adds three new components to this: Test-time augmentation (Section 3.1), false positive elimination (Section 3.2), and active learning via uncertainty estimation (Section 3.3).

3.1 Data Augmentation

To apply data augmentation, we run the CNN encoder N times on copies of the original diffusion MRI data to which voxel- and channel-wise i.i.d. Gaussian noise have been added. We use this augmentation twice: First, for increasing the amount of training data that is available for the random forest, so that each annotated voxel provides N examples. Second, as a test-time augmentation, by evaluating the random forest on each voxel N times, and averaging the results.

Formally, given a transformation $T(\mathbf{x};\sigma)$ that adds noise with standard deviation σ to the original data $\mathbf{x}$, CNN encoder ϕ, and random forest predictor f, the final estimate $\hat{y}$ that involves test-time augmentation can be written as

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} f(\phi(T(\mathbf{x};\sigma))) \tag{1}$$

3.2 False Positive Elimination

When using the original system, we observed that false positive detections of foreground classes often occurred far away from any of the scribbles, without an obvious pattern. This can be explained by the sparseness of the scribbles in feature space, so that in substantial parts of the volume, evaluating the random forest amounts to a generalization beyond its training distribution.

We propose to resolve the somewhat random behavior that occurs in such cases with the prior assumption that voxels that are dissimilar to any labeled ones are most likely to be background. Since it would not be efficient to compute dissimilarities via pairwise distances in feature space, we instead use isolation forests [7] as an anomaly detection method that is fast and easy to integrate in our pipeline, similar to the random forests that we already use for segmentation.

Isolation Forests build an ensemble of isolation trees by randomly selecting attributes and splitting values to partition the data. Since anomalous points are few and distinct, they are isolated closer to the root of the tree, resulting in shorter average path lengths $E(h(x))$ compared to normal points. The anomaly score is defined as

$$s(x,\psi) = 2^{-\frac{E(h(x))}{c(\psi)}} \in [0,1] \tag{2}$$

with $c(\psi)$ being the average path length, depending on the sample size ψ. In our experiments, we set voxels to background once the anomaly threshold exceeds 0.5 [7], and find that this already improves the results. In practice, the human annotator could easily be given control over this parameter in case they should find the detected foreground region to be too large or too small overall.

3.3 Uncertainty Estimation and Active Learning

The uncertainty measure that serves as a basis of our active learning approach is the traditional Shannon entropy

$$H(\hat{y}) = -\sum_{i=1}^{C} \hat{y}_i \log(\hat{y}_i)$$

of the prediction $\hat{y}$ after test-time augmentation as per Equation 1. Since diffusion MRI segmentation involves non-exclusive labels [10], we compute the uncertainty per-class, considering each case separately as a foreground/background segmentation with $C = 2$ classes.

We restrict this voxel-wise uncertainty measure to the brain mask.To keep the annotation effort feasible, our active learning mechanism suggests a single slice per class for annotation, and only a small subset of voxels within it. The selection happens probabilistically based on aggregated uncertainty values at the slice level. For each slice, candidate voxels are sampled based on the same voxel weights and a fixed budget of 200 voxel annotations. If a class is underrepresented, the budget is shifted to the largest class, then to the second largest and so on, ensuring a consistent number of annotated voxels per experiment. If a selected voxel belongs to multiple classes, all of them are annotated.

4 Experimental Setup and Results

4.1 Dataset

We evaluate our extensions and the active learning loop on the Human Connectome Project (HCP) dataset [5] in a similar fashion as in prior work [10]. The HCP dataset provides diffusion MR images (dMRI) which have high spatial resolution, with 145^3 voxels after clipping and isotropic voxel spacing of 1.25 mm, and high q space resolution, resulting in 288 channels per voxel. Overall, we consider a total of 93 distinct subjects for this study. For the experiments, segmentation tasks for eight randomly selected HCP subjects have been taken from the ground truth data that has been created during the development of TractSeg [11], a state-of-the-art tool for white matter bundle segmentation in dMRI data. We curate two distinct segmentation tasks from these masks:

- **Task 1:** Classes of varying sizes, namely *cingulum* (CG), *corticospinal tract* (CST), *fornix* (FX), and *corpus callosum* (CC).
- **Task 2:** Classes divided into the left and right hemispheres of the brain, probing the ability to distinguish between homologous structures. We use the *inferior occipito-frontal fascicle* (IFO), *inferior longitudinal fascicle* (ILF), and *superior longitudinal fascicle* (SLF), each split into left and right.

In these experiments, the number of annotated voxels is a simple, but approximate proxy for annotation effort. In the future, we hope to account for more realistic models, and to validate them with a user study.

4.2 Improving on the Baseline

First, we quantify how much TTA and FPE improve the baseline. We report Dice scores for individual tracts across both tasks for eight randomly chosen subjects. All methods use the same dual branch CNN [10] as a feature extractor, which was trained on the 93 subjects in the HCP dataset for 20 epochs in an unsupervised manner. All methods share the same initial annotations, in which up to 200 foreground voxels per class were chosen at random from the slice with the largest number of voxels for the respective class, and the same number of voxels were taken in the same way for the background.

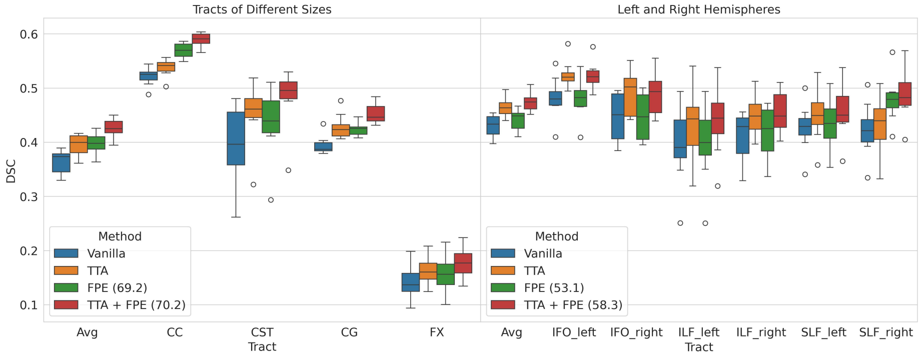

Fig. 2. Compared to the original method, test-time augmentation (TTA) and false positive elimination (FPE) increase the Dice score that is achieved on the initial set of scribbles, even before the active learning starts. The combination gives the best results.

Figure 2 compares the results of the proposed TTA + FPE pipeline to versions that use either TTA, FPE, or neither. For TTA we use $N = 4$ samples, which we found to be a good trade-off between performance and computational cost. For FPE, we use the Scikit-learn implementation of Isolation Forests [7] and its default anomaly threshold. Both mechanisms improve results on average, and combining both works best. The fraction of predicted foreground voxels which were kept after the FPE are reported in brackets, and is slightly higher when also using TTA, indicating that TTA reduces false positives by itself.

4.3 Effectiveness of Active Learning

After the initial annotations are generated and a prediction has been made by the random forest classifier, we perform multiple rounds of refinement, each time adding scribbles in one slice per class. Figure 3 summarizes how segmentation accuracy improves as more scribbles become available.

Vanilla and *random sampling* represent baselines that use random annotation, where each voxel within the brain mask has equal weight. In *random sampling,* our novel TTA and FPE are used, while they are omitted in *vanilla.* This comparison provides additional evidence for their effectiveness. We observed however that FPE becomes less relevant over iterations, as anomalies become less frequent, so that FPE eventually removes less than 5% of the predicted foreground. A benefit compared to *vanilla* remains, but is mostly due to TTA.

Error sampling uses a weighting based on the true error, which is not known to an active learning system. We still include it as a reference for what we could hope to achieve with an ideal uncertainty quantification and note that, in almost all cases, our proposed *entropy weighted sampling* (Section 3.3) closes part of the gap between *random* and *error sampling.*

To provide a qualitative impression of the overall benefit from our proposed modifications in a representative example subject after five iterations of active

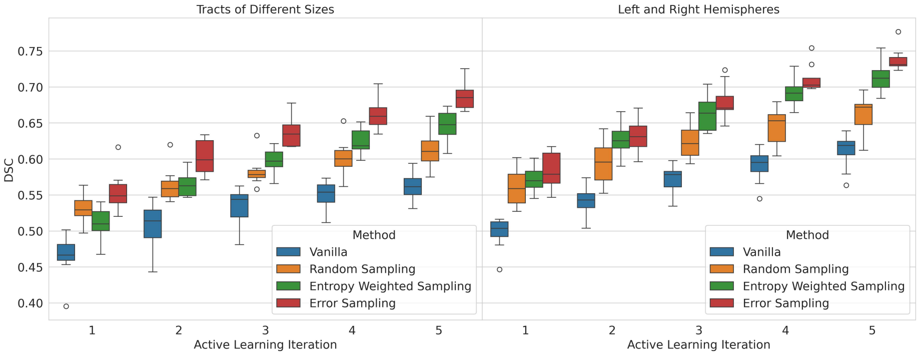

Fig. 3. Compared to random sampling, an uncertainty-weighted sampling of voxels for annotation produces more accurate segmentations in almost all cases. Also shown are the original "vanilla" method (random sampling without test-time augmentation and false positive elimination) and a sampling that makes use of the true error maps.

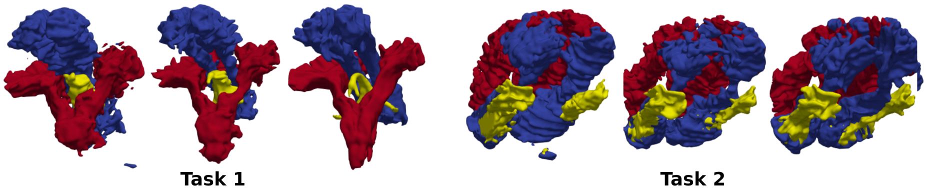

Fig. 4. Visual comparison between segmentations from the vanilla approach without TTA and FPE (left), our proposed active learning scheme (middle) and the ground truth (right) for a single subject after an interactive segmentation process of five steps. In Task 1, the corpus callosum is omitted to reduce occlusions.

learning, Figure 4 presents three-dimensional visualizations of some of the segmented tracts, and a comparison to the respective ground truth masks.

4.4 Sampling Approach

Only selecting the most uncertain voxels often leads to redundancy, i.e., it clumps together annotations in certain image regions [4]. Weighted random sampling is a simple strategy to increase the diversity of the selected voxels, while still focusing on uncertain ones. We evaluate three different sampling approaches for selecting voxels within a chosen slice: Taking the k highest entropy voxels, weighted sampling according to the normalized entropy, and weighted sampling according to normalized squared entropy. The idea behind the latter is that weighted sampling might select too few highly uncertain voxels if a much larger number of less uncertain ones are present. Squaring should reduce this problem by further reducing weights that are already close to zero, while still retaining more diversity than the top-k strategy.

AL Iteration	1	2	3	4	5
Task 1					
k Highest Values	.455±.023	.531±.017	.575±.021	.610±.018	.645±.022
Normalized Entropy	**.529±.033**	**.564±.036**	.598±.029	.612±.027	.631±.032
Squared Entropy	**.510±.024**	**.565±.018**	.597±.018	.625±.019	.647±.022
Task 2					
k Highest Values	.524±.031	.595±.032	.643±.027	.676±.023	**.701±.024**
Normalized Entropy	.566±.029	.612±.026	.640±.029	.668±.031	.686±.026
Squared Entropy	**.572±.020**	**.628±.024**	**.664±.026**	**.693±.023**	**.715±.023**

Table 1. Average DSC over 5 active learning iterations for each sampling strategy. Bold scores indicate a significant improvement over at least one competing strategy. Task 1: Classes of different size, Task 2: Similar classes in left and right hemisphere.

Results are shown in Table 1, where we report the mean DSC and its standard deviation across subjects over five active learning iterations for each sampling strategy. For each task and iteration, we first use a Friedman test to determine whether overall differences are significant, followed by a Nemenyi post-hoc test to investigate pairwise differences. Numbers are in bold if the Friedman test indicated significant overall differences ($\alpha = 0.05$) and the corresponding strategy was significantly better than at least one competitor. In all cases, one of the random strategies led to the best result. On the second task, squared entropy worked better than plain entropy, while both random strategies performed similarly well on the first task. For all experiments, we use TTA as well as FPE.

5 Discussion and Conclusion

In this work, we introduce general improvements on a previously described framework for interactive scribble-based diffusion MRI segmentation [10] by introducing data augmentation in training and inference, and false positive elimination via anomaly detection. Furthermore, we introduce an active learning scheme for that framework, based on uncertainty quantification via test-time augmentation.

Experiments on multiple subjects from the HCP dataset [5] across two segmentation tasks confirm that augmentation plays a crucial role during training, inference, and uncertainty estimation. They also show that false positive elimination is especially effective early in the annotation process, when the training data are still sparse, and that it turns itself off automatically once the feature space is sufficiently explored. Finally, we demonstrate an active learning approach whose suggested annotations are substantially better than random sampling and, in some cases, quite close to selecting annotations based on the ground truth. With all these components in place, we arrive at useful segmentation results within few annotation cycles, and minimal waiting times between them.

References

1. Chen, Q., Hong, Y.: Scribble2D5: Weakly-supervised volumetric image segmentation via scribble annotations. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI), Part VIII. LNCS, vol. 13438, pp. 234–243. Springer (2022). https://doi.org/10.1007/978-3-031-16452-1_23
2. Çiçek, O., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 424–432. Springer (2016)
3. Gaillochet, M., Desrosiers, C., Lombaert, H.: TAAL: test-time augmentation for active learning in medical image segmentation. In: Nguyen, H.V., Huang, S.X., Xue, Y. (eds.) Data Augmentation, Labelling, and Imperfections. pp. 43–53. Springer Nature Switzerland, Cham (2022)
4. Gaillochet, M., Desrosiers, C., Lombaert, H.: Active learning for medical image segmentation with stochastic batches. Medical Image Analysis **90**, 102958 (2023). https://doi.org/10.1016/j.media.2023.102958
5. Glasser, M.F., Smith, S.M., Marcus, D.S., Andersson, J.L.R., Auerbach, E.J., Behrens, T.E.J., Coalson, T.S., Harms, M.P., Jenkinson, M., Moeller, S., Robinson, E.C., Sotiropoulos, S.N., Xu, J., Yacoub, E., Ugurbil, K., Essen, D.C.V.: The human connectome project's neuroimaging approach. Nature Neuroscience **19**(9), 1175–1187 (2016)
6. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment Anything. In: Proc. IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3992–4003 (2023). https://doi.org/10.1109/ICCV51070.2023.00371
7. Liu, F.T., Ting, K.M., Zhou, Z.: Isolation-based anomaly detection. ACM Trans. Knowl. Discov. Data **6**(1), 3:1–3:39 (2012). https://doi.org/10.1145/2133360.2133363
8. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024). https://doi.org/10.1038/s41467-024-44824-z
9. Schultz, T., Vilanova, A.: Diffusion MRI visualization. NMR in Biomedicine **32**(4), e3902 (2019). https://doi.org/10.1002/nbm.3902
10. Torayev, A., Schultz, T.: Interactive Classification of Multi-Shell Diffusion MRI With Features From a Dual-Branch CNN Autoencoder. In: EG Workshop on Visual Computing for Biology and Medicine. pp. 1–11 (2020). https://doi.org/10.2312/vcbm.20201165
11. Wasserthal, J., Neher, P., Maier-Hein, K.H.: Tractseg - fast and accurate white matter tract segmentation. NeuroImage **183**, 239–253 (2018). https://doi.org/10.1016/j.neuroimage.2018.07.070