Spatial Aggregation for Semi-supervised Active Learning in 3D Medical Image Segmentation

Siteng Ma[1], Honghui Du[1(✉)], Dairui Liu[1], Kathleen M. Curran[2], Aonghus Lawlor[1], and Ruihai Dong[1(✉)]

[1] Insight Centre for Data Analytics, School of Computer Science, University College Dublin, Ireland
siteng.ma@ucdconnect.ie, {honghui.du,ruihai.dong}@ucd.ie
[2] School of Medicine, University College Dublin, Ireland

Abstract. Existing active learning (AL)-based 3D medical image segmentation methods often select images, slices, or patches as isolated entities, overlooking inter-slice spatial relationships in 3D images. Additionally, AL methods train the segmentation model on labeled data only and ignore valuable unlabeled data. Both factors limit its ability to further reduce labeled data needs. To address these problems, we propose a novel semi-supervised AL approach termed SpaTial AggRegation (STAR), which enables the model to learn from unlabeled data beyond annotated samples by leveraging spatial correlations between slices, reducing labeling costs. In each AL iteration, STAR employs a spatial cross-attention mechanism to transfer relevant knowledge from adjacent labeled slices to unlabeled ones by generating pseudo-labels. These pseudo-labeled slices and queried slices are used to train the segmentation model. The experimental results indicate that STAR outperforms other state-of-the-art AL methods, achieving fully supervised 3D segmentation performance with as little as 18%-19% of the labeled data. The code is available at https://github.com/HelenMa9998/STAR.

Keywords: 3D Medical Image Segmentation· Active Learning · Semi-supervised Learning · Label Cost

1 Introduction

Deep learning has shown strong performance in 3D medical image segmentation [11, 6, 19]. Yet, its effectiveness heavily depends on large amounts of labeled data and manual slice-by-slice annotation of 3D medical images is both costly and inefficient [24]. Active learning (AL) is an iterative algorithm where a model selectively queries labels for the most informative samples to improve learning efficiency with minimal labeled data [17]. However, AL-based 3D medical image segmentation methods [2, 13, 14, 20] train segmentation models exclusively on labeled data, overlooking the wealth of unlabeled data and missing valuable insights. Moreover, AL in 3D medical imaging can easily select redundant data (e.g., repeatedly selecting similar slices), leading to inefficient annotation and minimal performance gains.

Semi-supervised learning (SSL) allows models to learn from a small amount of labeled data and a larger amount of unlabeled data by pseudo-labeling or consistency regularization to improving performance [21, 28, 23]. Integrating SSL with AL helps reduce the number of training labels required [27], but only few studies [3, 12, 4] have explored semi-supervised active learning (SSAL) for 3D medical imaging. So far, these works have relied on a model initially trained on labeled data to pseudo-label unlabeled data for retraining. Since the reliability of pseudo-labels depends solely on the trained model, it can introduce noise (e.g., the model may confidently misclassify data and repeatedly make the same errors on similar inputs).

Consecutive slices within the same 3D medical image volume exhibit strong spatial correlations [9] (e.g., if a tumor in one slice is likely present in adjacent slices). Leveraging these correlations can minimize AL redundancy by reducing unnecessary, similar queries and enhance SSL by providing spatial context to correct mislabeled pseudo-labels. Nonetheless, current AL and SSL methods for 3D medical image analysis completely ignore spatial relationships. Therefore, this paper investigates the research question: *Can spatial correlations reduce labeling efforts with SSAL while maintaining performance? If so, how?*

To answer the question, we propose SpaTial AggRegation (STAR), the first SSAL algorithm that leverages spatial continuity between 2D slices within 3D medical images to reduce labeling costs while enhancing segmentation performance. STAR begins by selecting evenly spaced slices from each 3D volume for initial labeling to ensure sample diversity. It then employs a spatial cross-attention-based label aggregation module to generate pseudo-labels for unlabeled slices based on their similarities to nearby labeled ones while selecting dissimilar slices as queries for the oracle to label. At each iteration, confidently pseudo-labeled slices combine with manually labeled data to refine the segmentation model in an AL loop. Extensive experiments show STAR outperforms state-of-the-art methods across multiple datasets, achieving better segmentation performance with fewer labeled samples.

2 Proposed method

Consider a 2D slice sequence $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{d}$ extracted from a 3D medical image, where each slice $\mathbf{x}_i \in \mathbb{R}^{h \times w}$ has spatial dimensions $h \times w$. STAR (Fig. 1) sequentially partitions the full sequence into groups $\mathcal{X}_g$, where g represents the group index, by sampling every jth 2D slice for oracle annotation. Each group consists of two consecutive sampled slices and the intervening slices, totaling $j + 1$ slices [3]. The sampled slices form the labeled training pool, while the remaining unlabeled slices constitute the unlabeled set.

To ensure clinically valid pseudo-labels for unlabeled slices, STAR applies a spatial cross-attention mechanism (Section 2.1) to generate pseudo-labels by aggregating segmentation information from pathologically similar labeled slices

[3] If the last group has fewer than $j + 1$ slices, it is supplemented with slices from the previous group in practice.

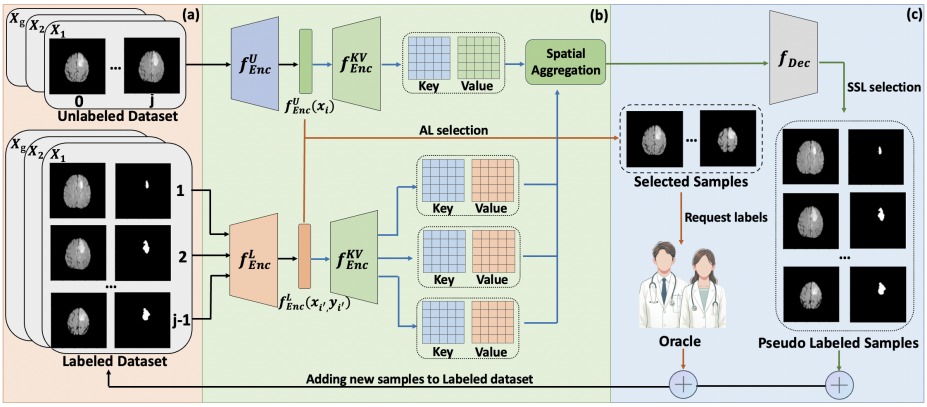


Fig. 1: The overview of STAR. (a) Slice Initialization; (b) Spatial Aggregation; (c) Label Generation via AL (orange flow) and SSL (green flow).

neighboring the unlabeled slices within the same group. This effectively mitigates sequential noise and cumulative errors common in traditional linear label propagation methods. Then, STAR queries the least neighbor-similar unlabeled slices and adds them, along with confidently pseudo-labeled slices to the training pool (Section 2.2). This repeats until a set number of iterations is reached.

2.1 Spatial Aggregation

STAR employs two encoders, f^U_{Enc} and f^L_{Enc}, to transform unlabeled and labeled slices into feature maps, respectively. f^U_{Enc} encodes a single unlabeled image, while f^L_{Enc} processes both image and its label via parallel convolutional layers and sums their outputs to incorporate label information. Both share the same backbone architecture (e.g., ResNet-50 [7]) but have different parameters. Subsequently, another encoder f^{KV}_{Enc} maps features into *key* and *value* pairs with different dimensions: the lower-dimensional *key* helps identify similar slices within spatially adjacent ones, while the higher-dimensional *value* stores detailed pathological structures for richer feature representation. Given an unlabeled slice $\mathbf{x}_i$ and a labeled slice $\mathbf{x}_{i'}$, the *key-value* matrices are computed as:

$$\mathbf{K}_i, \mathbf{V}_i = f^{KV}_{Enc}(f^U_{Enc}(\mathbf{x}_i)), \quad \mathbf{K}_{i'}, \mathbf{V}_{i'} = f^{KV}_{Enc}(f^L_{Enc}(\mathbf{x}_{i'}, \mathbf{y}_{i'})) \tag{1}$$

where $\mathbf{y}_{i'}$ is the corresponding segmentation label; $\mathbf{K} \in \mathbb{R}^{H\times W\times D/8}$ and $\mathbf{V} \in \mathbb{R}^{H\times W\times D/2}$; H, W, and D are the feature map's height, width, and depth.

For an unlabeled slice $\mathbf{x}_i \in \mathcal{X}_g$, STAR calculates its attention score by comparing its *key* with those of labeled slices in $\mathcal{X}_g$, where the score reflects the pathological similarity. Unlike the original cross-attention [25], STAR enhances it by concatenating the unlabeled slice's *value* with the sum of labeled slices' *values* weighted by attention scores. This ensures STAR refines representations

by aggregating contextual information from near-labeled data while preserving the unlabeled slice's intrinsic features. Formally:

$$Attention(\mathbf{x}_i) = \sum_{i' \in \mathcal{I}_g^L} Softmax(\frac{\mathbf{K}_{i'} \cdot \mathbf{K}_i}{\sqrt{D}})\mathbf{V}_{i'} \tag{2}$$

$$\mathbf{F}_i = Attention(\mathbf{x}_i) +\!\!+ \mathbf{V}_i \tag{3}$$

where $\cdot$ is dot product, $+\!\!+$ represents concatenation, $\mathcal{I}_g^L$ is the set of indices for labeled slices in group $\mathcal{X}_g$, and D is the depth dimension of $\mathbf{K}$. $\mathbf{F}_i$ is subsequently used to generate the pseudo-label $\hat{\mathbf{y}}_i \in \{0,1\}^{h \times w}$ for $\mathbf{x}_i$ through decoder (i.e., $\hat{\mathbf{y}}_i = f_{Dec}(\mathbf{F}_i)$).

2.2 Label Generation and Training Process

The training pool contains both labeled and pseudo-labeled slices. To optimize annotation selection, STAR exploits spatial correlations to minimize redundant queries in AL process. To avoid the noise and cost of raw image comparisons, we compute slice similarity in the embedding space where features are more meaningful and robust. Cosine similarity efficiently captures angular differences and is robust to appearance changes, making it well-suited for image embeddings.

Therefore, STAR selects the unlabeled slice with the lowest average cosine similarity to labeled slices within each group for annotation, as less similar slices are often more informative and improve segmentation performance. The spatial similarity of an unlabeled slice $\mathbf{x}_i$ is calculated as:

$$Sim(\mathbf{x}_i) = \sum_{i' \in \mathcal{I}_g^L} \frac{cos(f_{Enc}^U(\mathbf{x}_i), f_{Enc}^L(\mathbf{x}_{i'}, \mathbf{y}_{i'}))}{|\mathcal{I}_g^L|} \tag{4}$$

In addition, to ensure pseudo-labels reliability and reduce noise, STAR uses entropy to rank predictions and selects the top N as high-confidence pseudo-labels, which are updated iteratively during training.

Following the AL framework, STAR performs multi-round training with iterative selection. In each round, STAR adopts a two stage training strategy to enhance generalization and training efficiency. In the first stage, sampled slices are labeled and added to the labeled pool, where f_{Enc}^U and f_{Dec} are trained through supervised learning with focal loss [10]. The parameters of f_{Enc}^U are shared with f_{Enc}^L for initialization. In the second stage, STAR pairs every two closest labeled slices for training (e.g., $((\mathbf{x}_{i'}, \mathbf{y}_{i'}, \mathbf{x}_{i''}), \mathbf{y}_{i''})$). Concretely, STAR feeds $(\mathbf{x}_{i'}, \mathbf{y}_{i'})$ into f_{Enc}^L and $\mathbf{x}_{i''}$ into f_{Enc}^U. The model then predicts the segmentation label $\hat{\mathbf{y}}_{i''}$, which is subsequently compared with the actual label $\mathbf{y}_{i''}$ using propagation loss to update f_{Enc}^L, f_{Enc}^U, f_{Enc}^{KV} and f_{Dec}. This process repeats for all pairs across all the 3D images. The training continues until the annotation budget is met or a stopping criterion is reached, resulting in a model trained on a dataset that combines strategically selected labeled and unlabeled data.

Table 1: Model performance with varied label percentages on BraTS and Spleen. $max(p)$ is the highest performance. $S\%$ is the minimum data percentage needed to match supervised performance (0.834 on BraTS, 0.891 on Spleen). A dash (-) indicates the method does not reach the supervised performance. The best approach is highlighted in red.

BraTS	**AL**	**SSL**	**15%**	**20%**	**25%**	**30%**	**max(p)**	**S%**
Rand	✓	×	0.806 ± 0.03	0.819 ± 0.01	0.828 ± 0.01	0.833 ± 0.01	0.833 ± 0.01	-
Marg	✓	×	0.807 ± 0.02	0.823 ± 0.01	0.827 ± 0.01	0.819 ± 0.01	0.833 ± 0.01	-
Ent	✓	×	0.797 ± 0.01	0.807 ± 0.02	0.828 ± 0.01	0.831 ± 0.01	0.833 ± 0.01	-
MC-D	✓	×	0.807 ± 0.02	0.828 ± 0.01	0.825 ± 0.01	0.825 ± 0.01	0.830 ± 0.01	-
KCG	✓	×	0.800 ± 0.02	0.825 ± 0.02	0.830 ± 0.02	0.832 ± 0.01	0.832 ± 0.01	-
BAL	✓	×	0.793 ± 0.03	0.819 ± 0.01	0.809 ± 0.02	0.812 ± 0.03	0.831 ± 0.00	-
CEAL(Marg)	✓	✓	0.822 ± 0.01	0.816 ± 0.01	0.810 ± 0.03	0.818 ± 0.01	0.823 ± 0.01	-
CEAL(Ent)	✓	✓	0.807 ± 0.02	0.828 ± 0.01	0.829 ± 0.01	0.816 ± 0.01	0.833 ± 0.01	-
CEAL(MC-D)	✓	✓	0.808 ± 0.02	0.825 ± 0.01	0.805 ± 0.02	0.821 ± 0.03	0.831 ± 0.01	-
TAAL	✓	✓	0.815 ± 0.02	0.824 ± 0.01	0.814 ± 0.02	0.824 ± 0.01	0.832 ± 0.01	-
STAR	✓	✓	**0.828 ± 0.01**	**0.837 ± 0.01**	**0.830 ± 0.01**	**0.834 ± 0.01**	**0.836 ± 0.01**	**19.4%**
Spleen	**AL**	**SSL**	**15%**	**20%**	**25%**	**30%**	**max(p)**	**S%**
Rand	✓	×	0.862 ± 0.03	0.871 ± 0.01	0.869 ± 0.04	0.888 ± 0.01	0.904 ± 0.01	34.1%
Marg	✓	×	0.870 ± 0.01	0.854 ± 0.04	0.871 ± 0.02	0.869 ± 0.03	0.910 ± 0.02	32.1%
Ent	✓	×	0.857 ± 0.03	0.877 ± 0.02	0.861 ± 0.02	0.859 ± 0.00	0.890 ± 0.00	-
MC-D	✓	×	0.857 ± 0.02	0.859 ± 0.03	0.866 ± 0.03	0.890 ± 0.02	0.898 ± 0.02	36.0%
KCG	✓	×	0.853 ± 0.03	0.843 ± 0.04	0.880 ± 0.01	0.888 ± 0.02	0.897 ± 0.01	34.1%
BAL	✓	×	0.849 ± 0.03	0.876 ± 0.03	0.860 ± 0.03	0.891 ± 0.02	0.892 ± 0.02	-
CEAL(Marg)	✓	✓	0.805 ± 0.05	0.853 ± 0.02	0.838 ± 0.03	0.868 ± 0.01	0.883 ± 0.01	-
CEAL(Ent)	✓	✓	0.828 ± 0.03	0.874 ± 0.02	0.894 ± 0.01	0.855 ± 0.02	0.907 ± 0.02	32.1%
CEAL(MC-D)	✓	✓	0.862 ± 0.02	0.885 ± 0.01	0.865 ± 0.04	**0.893 ± 0.03**	0.904 ± 0.03	20.2%
TAAL	✓	✓	0.861 ± 0.04	0.876 ± 0.02	0.868 ± 0.01	0.868 ± 0.03	0.910 ± 0.01	24.2%
STAR	✓	✓	**0.873 ± 0.03**	**0.899 ± 0.02**	**0.902 ± 0.01**	0.884 ± 0.03	**0.913 ± 0.02**	**18.2%**

3 Experiments and Results

3.1 Experimental Setup

Two well-known public medical image segmentation datasets are selected: **BraTS** (MRI) [4] [1], **Spleen** (CT) [22]. Annotating 3D images is challenging due to their high dimensionality. A common approach is to divide them into 2D slices for easier labeling. To simulate this process, datasets are split by slices into training (7,750 BraTS, 2,525 Spleen), validation (2,170 BraTS, 1,860 Spleen), and test sets (644 BraTS, 481 Spleen) subject-wise.

Ten widely used methods, covering all major types of query strategies and SSAL methods, are selected for comparison. These include uncertainty-based methods (**Margin (Marg)** [15], **Entropy (Ent)** [18], **MC-Dropout (MC-D)** [5]), diversity-based methods (**Random (Rand)**, **K-Center Greedy (KCG)** [16]), and hybrid methods (**BAL** [8]), and SSAL methods (**CEAL(Marg)** [26], **CEAL(Ent)** [26], **CEAL(MC-D)** [26], **TAAL** [4]).

The initial labeling interval j is set to 10. To simulate AL's interactive selection, 100 slices are queried with 100 pseudo-labels per iteration for BraTS, while

[4] As FLAIR is the most common MRI sequence for pathology visualisation and segmentation, this study uses the LGG FLAIR sequence for tumor core segmentation.

50 slices are queried with 50 pseudo-labels per iteration for Spleen. The segmentation model is ResNet-50 [7], trained with the Adam optimizer at a learning rate of 0.0001 for supervised training and 0.00001 for incremental training. Batch sizes are 32 for BraTS and 16 for Spleen, with a maximum of 100 epochs. Each AL iteration continues until the validation loss converges, with early stopping patience of 5 epochs. Model parameters are fine-tuned incrementally for efficiency. All experiments are conducted on an NVIDIA GeForce RTX 4090 with consistent hyperparameters.

Segmentation performance is evaluated with the Dice coefficient [29]. To compare various AL methods, we track test performance at each step during AL process until the performance stabilizes, recording the peak performance achieved. The performance is reported at 5% data intervals, starting from 15% up to stabilization (e.g., 30%). All experiments were performed 5 times with different random seeds, and the average across these runs is reported. Furthermore, we measure the minimum labeled slices required to match fully supervised performance, assessing each method's efficiency in reducing annotation effort.

3.2 Experiment Results

The experimental results for BraTS and Spleen are shown in Table 1. STAR outperforms all methods across all percentage levels, except at 30% in Spleen, where it remains competitive. The fewer labeled samples available, the greater STAR's advantage, highlighting its effectiveness in selecting informative samples. With labeled data increases, performance differences across methods converge due to diminishing returns. Notably, STAR is the only method to reach supervised performance in BraTS with just 19.4% of the data and in Spleen with only 460 labeled slices (18.2%). In comparison, CEAL(MC-D) (second) requires 50 more labeled slices (510, 20.2%) and TAAL (third-best) requires 150 (610, 24.2%), demonstrating STAR's efficiency in reducing annotation effort.

To better assess segmentation performance, we visualize predictions and compare them to ground truth masks. Fig. 2 shows representative results from the five best-performing methods on BraTS and Spleen, featuring slices with large and small lesions. Large lesions are generally easier to segment, while small ones are the most challenging. Compared to other methods, STAR produces clearer, more accurate boundaries for large lesions (shown in Fig. 2(vii)). For small lesions, other methods often over- or under-segment or fail to detect the target (shown in Fig. 2(iii)-(vi)), while ours predicts accurately. These results highlight STAR's efficiency in leveraging limited labeled data.

To further investigate STAR's effectiveness in generating pseudo-labels via spatial aggregation, we compare pseudo-label accuracy with and without the spatial aggregation on BraTS. Fig. 3b shows that spatial aggregation substantially improves pseudo-label accuracy by approximately 3% across various labeled data percentages. Fig. 3a(i) and (iv) illustrate that spatial cross-attention accurately identifies useful anatomical information (tumor areas in red) in nearby slices. Fig. 3a(iv) shows STAR's pseudo-label closely matching the ground truth (Fig. 3a(iii)), demonstrating the accuracy of our proposed spatial aggregation.

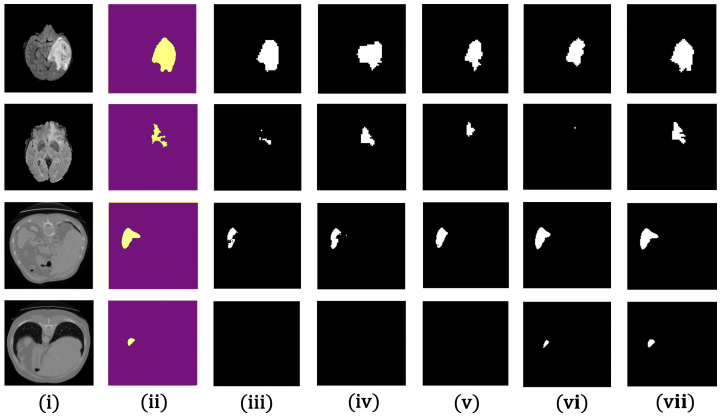


Fig. 2: Examples of segmentation results on BraTS (top two rows, 20% training data) and the Spleen (bottom two rows, 15% training data). (i) Original slice, (ii) Ground truth, (iii) MC-D (BraTS) / Rand (Spleen), (iv) KCG (BraTS) / Marg (Spleen), (v) CEAL(Ent) (BraTS) / CEAL(MC-D) (Spleen), (vi) CEAL(MC-D) (BraTS) / TAAL (Spleen), (vii) STAR.

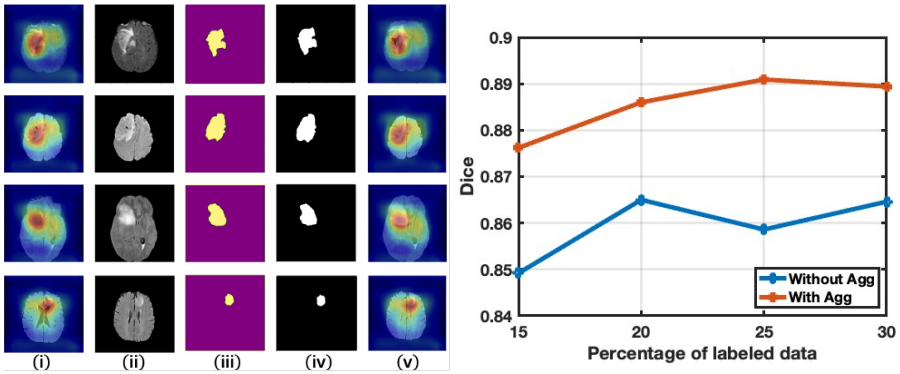
Table 2: Ablation study on BraTS. Agg denotes the proposed Spatial Aggregation, while Sim refers to the spatial similarity-based query strategy (Section 2.2). CEAL applies a general pseudo-labeling approach [26]. $max(p)$ is the highest performance. $S\%$ is the minimum data percentage needed to match supervised performance. The best approach is highlighted in red.

Agg	Sim	15%	20%	25%	30%	max(p)	S%
×	×	0.806 ± 0.025	0.819 ± 0.010	0.828 ± 0.014	0.833 ± 0.006	0.833 ± 0.006	-
×	✓	0.827 ± 0.014	0.825 ± 0.010	0.816 ± 0.031	0.827 ± 0.018	0.833 ± 0.018	-
✓	×	0.825 ± 0.021	0.821 ± 0.010	0.824 ± 0.012	0.821 ± 0.012	0.834 ± 0.012	1700 (21.9%)
CEAL	✓	0.812 ± 0.024	0.836 ± 0.013	0.819 ± 0.007	0.828 ± 0.008	0.836 ± 0.013	**1500 (19.4%)**
✓	✓	**0.828 ± 0.006**	**0.837 ± 0.007**	**0.830 ± 0.008**	**0.834 ± 0.005**	**0.837 ± 0.007**	**1500 (19.4%)**

3.3 Ablation Study

STAR utilizes spatial correlations in two components: spatial aggregation (Section 2.1) and the spatial-similarity-based query strategy (Section 2.2). To evaluate their impact and functionality, we conduct an ablation study by removing each component separately. Additionally, we replace spatial aggregation with another semi-supervised method (e.g., CEAL [26]) to further examine its effectiveness. Experiments are performed on BraTS, with the average performance reported over three runs due to the computational cost.

Table 2 presents the ablation study results. When the similarity-based query strategy is inactive, the model with spatial aggregation (row 3) outperforms the one without it (row 1) at smaller training sizes (15%–20%) and achieves similar performance at 25%–30%. This may be because, as labeled training data increases, the influence of pseudo-labels on training decreases. When the



(a) Pseudo-label examples on BraTS.

(b) Pseudo-label Accuracy.

Fig. 3: Pseudo-label Effectiveness. (a): (i) First slice in the group with attention map, (ii) Unlabeled slices, (iii) Ground truth, (iv) Pseudo-labels, (v) Last slice with attention map. Attention maps reflect similarity, where the blue regions indicate areas of high similarity. (b): Accuracy is measured by averaging dice score of predictions across all iterations. Without aggregation, pseudo-labels rely solely on the model's predictions.

similarity-based query strategy is active, the model with spatial aggregation (row 5) surpasses the one without it (row 2) across all data percentages, except at 15%, where the difference is minimal. This improvement is due to spatial aggregation refining segmentation, which enhances pseudo-label quality. A more accurate segmentation model further improves pseudo-label reliability, creating a reinforcing feedback loop that drives continuous performance gains. Additionally, comparing spatial aggregation (row 5) to CEAL (row 4) shows that spatial aggregation consistently outperforms CEAL, demonstrating its effectiveness in pseudo-label generation.

The spatial similarity-based query strategy also positively impacts model performance. When enabled (rows 2 and 5), it improves results across most labeled data percentages compared to when disabled (rows 1 and 3), except at 25% and 30%. This is because as the training data increases, the model has already learned sufficient information, reducing the impact of example selection and pseudo-labels on further improving performance. Furthermore, comparing rows 3 and 5 confirms that combining the similarity-based query strategy with memory aggregation reinforces both components, yielding better segmentation performance than either approach alone.

4 Conclusions

This paper focuses on an unexplored research question: how to leverage spatial information in 3D medical images to reduce labeling costs. To tackle this challenge, we propose a novel SSAL framework called STAR. STAR is the first

algorithm to use the spatial relationships between 2D slices in 3D medical images to effectively select instances for label querying and improve the accuracy of pseudo-labels, thereby reducing labeling costs and enhancing segmentation performance.

We evaluate STAR on two public 3D medical image datasets, demonstrating its superiority over existing methods. Notably, STAR enables the segmentation model to achieve fully supervised performance with only 18%–19% of labeled data. Currently, STAR incorporates only high-confidence pseudo-labels into training. Future work includes integrating more unlabeled data via consistency regularization or contrastive learning to enhance distribution understanding and reduce labeling costs, exploring random initialization for better generalization and applying the proposed method at the 3D patch level.

Acknowledgments. This research was conducted with the financial support of Science Foundation Ireland (SFI) to the Insight Centre for Data Analytics under Grant No. 12/RC/2289_P2.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data **4**(1), 1–13 (2017)
2. Chakravarthy, A.D., Abeyrathna, D., Subramaniam, M., Chundi, P., Gadhamshetty, V.: Semantic image segmentation using scant pixel annotations. Machine Learning and Knowledge Extraction **4**(3), 621–640 (2022)
3. Cheng, J., Liu, J., Kuang, H., Wang, J.: A fully automated multimodal mri-based multi-task learning for glioma segmentation and idh genotyping. IEEE Transactions on Medical Imaging **41**(6), 1520–1532 (2022)
4. Gaillochet, M., Desrosiers, C., Lombaert, H.: Taal: Test-time augmentation for active learning in medical image segmentation. In: MICCAI Workshop on Data Augmentation, Labelling, and Imperfections. pp. 43–53. Springer (2022)
5. Górriz, M., Giró Nieto, X., Carlier, A., Faure, E.: Cost-effective active learning for melanoma segmentation. In: ML4H: Machine Learning for Health NIPS, Workshop at NIPS 2017. pp. 1–5 (2017)
6. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. Li, G., Otake, Y., Soufi, M., Taniguchi, M., Yagi, M., Ichihashi, N., Uemura, K., Takao, M., Sugano, N., Sato, Y.: Hybrid representation-enhanced sampling for bayesian active learning in musculoskeletal segmentation of lower extremities. International Journal of Computer Assisted Radiology and Surgery pp. 1–10 (2024)

9. Li, S., Yin, H., Fang, L.: Group-sparse representation with dictionary learning for medical image denoising and fusion. IEEE Transactions on biomedical engineering **59**(12), 3450–3459 (2012)
10. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
11. Luo, X., Chen, J., Song, T., Wang, G.: Semi-supervised medical image segmentation through dual-task consistency. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 8801–8809 (2021)
12. Nath, V., Yang, D., Roth, H.R., Xu, D.: Warm start active learning with proxy labels and selection via semi-supervised fine-tuning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 297–308. Springer (2022)
13. Ozdemir, F., Peng, Z., Fuernstahl, P., Tanner, C., Goksel, O.: Active learning for segmentation based on bayesian sample queries. Knowledge-Based Systems **214**, 106531 (2021)
14. Peng, H., Lin, S., King, D., Su, Y.H., Abuzeid, W.M., Bly, R.A., Moe, K.S., Hannaford, B.: Reducing annotating load: Active learning with synthetic images in surgical instrument segmentation. Medical Image Analysis **97**, 103246 (2024)
15. Scheffer, T., Decomain, C., Wrobel, S.: Active hidden markov models for information extraction. In: International symposium on intelligent data analysis. pp. 309–318. Springer (2001)
16. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. In: International Conference on Learning Representations (2018)
17. Settles, B.: Active learning literature survey (2009)
18. Settles, B., Craven, M.: An analysis of active learning strategies for sequence labeling tasks. In: proceedings of the 2008 conference on empirical methods in natural language processing. pp. 1070–1079 (2008)
19. Shaker, A.M., Maaz, M., Rasheed, H., Khan, S., Yang, M.H., Khan, F.S.: Unetr++: delving into efficient and accurate 3d medical image segmentation. IEEE Transactions on Medical Imaging (2024)
20. Shen, M., Zhang, J.Y., Chen, L., Yan, W., Jani, N., Sutton, B., Koyejo, O.: Labeling cost sensitive batch active learning for brain tumor segmentation. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1269–1273. IEEE (2021)
21. Shi, L., Zhang, X., Liu, Y., Han, X.: A hybrid propagation network for interactive volumetric image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part IV. pp. 673–682. Springer (2022)
22. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019)
23. Su, J., Luo, Z., Lian, S., Lin, D., Li, S.: Mutual learning with reliable pseudo label for semi-supervised medical image segmentation. Medical Image Analysis **94**, 103111 (2024)
24. Top, A., Hamarneh, G., Abugharbieh, R.: Active learning for interactive 3d image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011: 14th International Conference, Toronto, Canada, September 18-22, 2011, Proceedings, Part III 14. pp. 603–610. Springer (2011)

25. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
26. Wang, K., Zhang, D., Li, Y., Zhang, R., Lin, L.: Cost-effective active learning for deep image classification. IEEE Transactions on Circuits and Systems for Video Technology **27**(12), 2591–2600 (2016)
27. Zhang, W., Zhu, L., Hallinan, J., Zhang, S., Makmur, A., Cai, Q., Ooi, B.C.: Boostmis: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20666–20676 (2022)
28. Zhou, T., Li, L., Bredell, G., Li, J., Unkelbach, J., Konukoglu, E.: Volumetric memory network for interactive medical image segmentation. Medical Image Analysis **83**, 102599 (2023)
29. Zijdenbos, A.P., Dawant, B.M., Margolin, R.A., Palmer, A.C.: Morphometric analysis of white matter lesions in mr images: method and validation. IEEE transactions on medical imaging **13**(4), 716–724 (1994)