Few Slices Suffice: Multi-Faceted Consistency Learning with Active Cross-Annotation for Barely-supervised 3D Medical Image Segmentation

Xinyao Wu*, Zhe Xu*(✉), and Raymond Kai-yu Tong(✉)

Department of Biomedical Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China
jackxz@link.cuhk.edu.hk; kytong@cuhk.edu.hk

Abstract. Deep learning-based 3D medical image segmentation typically demands extensive densely labeled data. Yet, voxel-wise annotation is laborious and costly to obtain. Cross-annotation, which involves annotating only a few slices from different orientations, has recently become an attractive strategy for labeling 3D images. Compared to previous weak labeling methods like bounding boxes and scribbles, it can efficiently preserve the 3D object's shape and precise boundaries. However, learning from such sparse supervision signals (aka. barely supervised learning (BSL)) still poses great challenges including less fine-grained object perception, less compact class features and inferior generalizability. To this end, we present a Multi-Faceted ConSistency (MF-ConS) learning framework for the BSL scenario. Our approach starts with an active cross-annotation strategy that requires only three orthogonal labeled slices per scan, optimizing the usage of limited annotation budget through a human-in-the-loop process. Building on the popular teacher-student model, MF-ConS is equipped with three types of consistency regularization to tackle the aforementioned challenges of BSL: (i) neighbor-informed object prediction consistency, which improves fine-grained object perception by encouraging the student model to infer complete segmentation from partial visual cues; (ii) non-parametric prototype-driven consistency for more discriminative and compact intra-class features; (iii) a stability constraint under mild perturbations to enhance model's robustness. Our method is evaluated on the task of brain tumor segmentation from T2-FLAIR MRI and the promising results show the superiority of our approach over relevant state-of-the-art methods.

Keywords: Barely-supervised · Cross-Annotation · Consistency.

1 Introduction

3D Medical image segmentation plays a pivotal role in computer-aided diagnosis. Developing high-performing segmentation models, however, hinges on the avail-

* Equal contribution

ability of extensively labeled voxel-wise data. This requirement is both resource-intensive and costly, necessitating significant time and expertise from radiologists. Semi-supervised learning (SSL) [22] has emerged as a promising approach to reduce the annotation burden via learning from a small densely labeled subset and an abundant unlabeled subset. Typically, the labeling strategy for SSL involves randomly selecting a few samples for dense labeling. However, this standard approach of budget allocation often results in annotation redundancy, with the limited labeling budget being allocated across too few samples. Intuitively, if we utilize more budget-friendly labeling methods for the selected scans, we can annotate more samples, thereby increasing the diversity of the labeled pool.

Regarding budget-friendly labeling, conventional methods such as image-level annotations [8], bounding boxes [14], scribbles [11, 30], and point annotations [10] are widely used. However, these strategies still result in a notable disparity in performance due to their inability to provide important boundary or inter-slice information, thereby hindering the establishment of the model's spatial object perception. Beyond these, cross-annotation, which entails annotating only a few slices from different orientations, has recently emerged as an attractive strategy for labeling 3D images, offering an efficient yet effective way to capture the shape and precise boundaries of 3D objects [3, 4]. Thus, this study investigates effective learning strategies under the cross-annotation paradigm, a scenario we refer to as barely supervised learning (BSL). Specifically, we propose annotating just three orthogonal slices for regions of interest from axial, sagittal and coronal views for each scan, as depicted in Fig. 1. This strategy considers the 3D spatial information and the differences between the three planes, providing efficient and effective supervision signals for model training. However, learning from such sparse supervision is more challenging than the typical SSL. As a result, this cross-labeling strategy necessitates a well-structured and optimized training framework to work in conjunction with human efforts.

Related Work. An intuitive approach to address slice-wise sparse supervision is to utilize the inter-slice similarity of 3D medical images, employing slice-to-slice registration to generate pseudo labels. The registration module can be jointly trained model [9] or off-the-shelf tools like ANTs [1], with the latest work, DeSCO [3], leveraging a co-training framework to exploit the registration-based pseudo labels and orthogonal labels. However, obtaining satisfactory registration results is a challenging task itself, especially for complex objects and the large variance in adjacent slices. From another view, sparse supervision can be interpreted as a challenging variant of semi-supervised learning (SSL) for extreme annotation scarcity. Here, samples traditionally considered on a scan-wise basis are instead investigated on a voxel-wise level, treating labeled voxels as individual labeled samples, and the rest as unlabeled. So far, consistency learning has become a mainstream SSL fashion [28, 6, 26, 27, 15, 24, 25, 21]. Yet, SSL inherently hinges on effective knowledge transfer from labeled to unlabeled data. Such sparse supervision poses great challenges to traditional SSL methods, including *less fine-grained object perception*, *less compact class features* and *inferior generalizability*. Thus, more effective designs for consistency regularization are called for BSL.

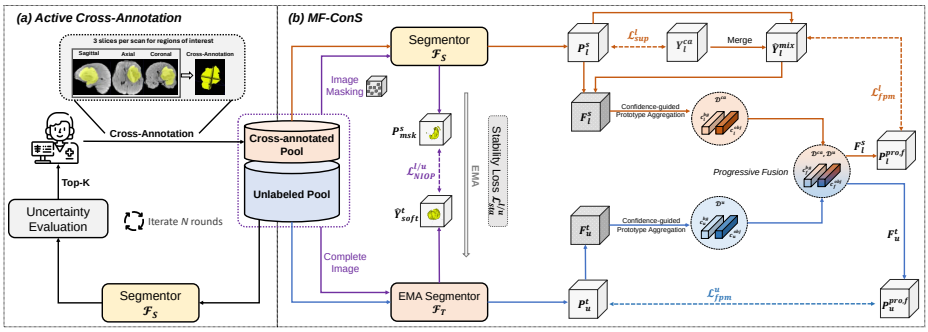

Fig. 1. Illustration of the (a) active cross-annotation strategy and (b) the proposed Multi-Faceted ConSistency (MF-ConS) learning framework.

To this end, we present a Multi-Faceted ConSistency (MF-ConS) learning framework for the BSL scenario. Our approach starts with a cross-annotation strategy that requires only three orthogonal labeled slices per scan and incorporates active learning to optimize the annotation budget allocation through a human-in-the-loop process. Building on the popular teacher-student model, MF-ConS features three types of consistency regularization to tackle the aforementioned challenges in BSL: (i) neighbor-informed object prediction consistency that encourages the student model to make associations and infer complete segmentation from limited visual cues (via masking) and then ensures that the outcomes are consistent in the segmentation space with those induced by the teacher model using the entire image. Intuitively, higher consistency indicates a stronger capability of fine-grained contextual object perception, serving as effective self-supervised signals given sparse supervision. (ii) non-parametric prototype-driven consistency that leverages the alignment between latent feature space (e.g., prototypes) and decision space (traditional inductive outputs by the model) to obtain more discriminative and compact intra-class features. Considering the scarcity of reliable labels in both sparsely-labeled data and unlabeled data for prototype generation, our strategy involves the fusion of prototypes derived from both data types to comprehensively represent the distribution of the feature space instead of separate usage in previous SSL [27] with relatively adequate ground truth. (iii) standard stability constraint that encourages consistent predictions under mild perturbations to enhance model's robustness. Our method is evaluated on brain tumor segmentation and the promising results show the superiority of our approach over relevant state-of-the-art methods.

2 Method

2.1 Active Cross-Annotation Strategy and Problem Formulation

Given a training set $\mathcal{D}$ of N volumes $\{X_1, X_2, \ldots, X_N\}$ of image dimensions $X \in \mathbb{R}^{H \times W \times D}$ (H, W and D denote the height, width and depth), MF-ConS

starts with a cross-annotation strategy. This strategy (Fig. 1 (a)) necessitates only three orthogonal labeled slices from the axial, sagittal, and coronal views for a subset of selected cases, effectively covering regions of interest (ROI) per scan. Additionally, it employs active learning [19] to optimize the allocation of the annotation budget via a human-in-the-loop process. This cross-annotation approach, by spanning multiple planes, has been proven effective in ensuring comprehensive coverage of the data distribution and efficiently capturing the 3D spatial information of objects [3]. We denote the cross-annotated label for image X_l as $Y_l^{ca} \in \mathbb{R}^{H\times W\times D}$. Note that the flexibility of cross-annotation allows for adjustments based on the complexity of the task and budget constraints. While our study uses the baseline that labels one slice per plane, the strategy can be adapted to label additional slices per plane for more challenging tasks or when fewer training samples are available. Regarding the active selection mechanism, we employ the classical entropy-based uncertainty sampling, which prioritizes the top-K cases that exhibit the highest average entropy in each selection round r to allocate resources to the samples that are most likely to enhance the model's learning. In general, our goal is to develop a segmentation model on a dataset $\mathcal{D}$, which includes a growing cross-annotated subset $\mathcal{D}^{ca} = \{(X_{l(i)}, Y_{l(i)}^{ca})\}_{i=1}^{n_{ca}}$, alongside the remaining unlabeled subset $\mathcal{D}^u = \{X_{u(i)}\}_{i=n_{ca}+1}^{N}$.

2.2 Multi-Faceted Consistency Learning

Basic Architecture. As depicted in Fig. 1 (b), our basic architecture includes a student segmentor $\mathcal{F}_S$ and a teacher segmentor $\mathcal{F}_T$. The weights θ of $\mathcal{F}_S$ are updated via standard back-propagation, while the weights $\tilde{\theta}_t$ for the t-th iteration of $\mathcal{F}_T$ are updated by the exponential moving average (EMA) of the student's weights, formulated as $\tilde{\theta}_t = \alpha\tilde{\theta}_{t-1}+(1-\alpha)\theta_t$, where α is the EMA decay rate and empirically set to 0.99 [28]. Besides efficiency, this EMA updating design can foster a self-ensembling slow-moving online teacher segmentor, which also assists in preventing representation collapse [17] due to sparse and potentially imbalanced supervision. Building on this architecture, we develop three types of consistency regularization to address the key challenges faced in BSL: less fine-grained object perception, less compact class features, and inferior generalizability.

Neighbor-Informed Object Prediction (NIOP) Consistency. To achieve fine-grained contextual perception, we encourage the student model to make associations [16] and infer complete segmentation from limited visual cues, as exampled in Fig. 2 (a). Specifically, we split image X into several $p \times p \times p$ patches without overlap. Each patch is randomly masked with a probability of δ, resulting in a masked input X_{msk}. X_{msk} is fed into the student segmentor, resulting in the predicted probability map $P_{msk}^{s} = \mathcal{F}_S(X_{msk};\theta)$. Then, the complete input X is fed into the EMA (teacher) segmentor, and we use a sharpening function to transform the probability output P^t into the soft pseudo label $\hat{Y}_{soft}^t$, formulated as $\hat{Y}_{soft}^t = \frac{(P^t)^{1/\beta}}{(P^t)^{1/\beta}+(1-P^t)^{1/\beta}}$, where β is the temperature of sharpening, empirically set to 0.1 [23]. We encourage consistent segmentation outcomes

between P^s_{msk} and $\hat{Y}^t_{soft}$ using mean absolute error (MAE), formulated as:

$$\mathcal{L}_{NIOP} = \mathcal{L}_{\text{MAE}}\left(P^s_{msk}, \hat{Y}^t_{soft}\right). \tag{1}$$

Distinct from [7] that performs mask-then-reconstruct for pre-training, our NIOP consistency advocates the task-specific mask-then-segment design. Intuitively, greater consistency reflects an enhanced ability for fine-grained contextual object perception because the model is compelled to execute implicit reconstruction of masked patches, followed by high-level object segmentation.

Prototype-driven Consistency. A significant challenge arising from sparse supervision is less compact and less discriminative class features [30], leading to an ambiguous embedding space. To this end, we introduce a non-parametric prototype-driven consistency learning scheme [20] to align the latent feature space (e.g., prototypes) and decision space (inductive outputs by the model) [27, 12]. Let F^s_l be the feature map of the cross-annotated image from the student segmentor, and F^t_u be the feature map of the unlabeled image from the teacher segmentor. The feature map is obtained from the layer preceding the penultimate convolution of the model and is then upsampled to match the image size using trilinear interpolation. Generally, for each class, the process of confidence-guided prototype aggregation [20] is defined as $c^{class} = \frac{\sum_v \left[\hat{Y}_v^{class} \cdot M_v^{class} \cdot F_v\right]}{\sum_v \left[\hat{Y}_v^{class} \cdot M_v^{class}\right]}$, where $\hat{Y}_v^{class}$ represents the one-hot class label; M_v^{class} denotes the reliability map, indicated by the prediction confidence. For cross-annotated data, $\hat{Y}_v^{class}$ is derived by merging cross-annotation with the argmax pseudo label generated by the student segmentor for unlabeled voxels, denoted as $\hat{Y}_l^{mix}$. For unlabeled data, $\hat{Y}_v^{class}$ is the argmax pseudo label generated by the teacher segmentor, denoted as $\hat{Y}_u^t$. Combining with the aforementioned feature maps (F^s_l and F^t_u), we can obtain the corresponding class prototypes $\{c_l^{bg}, c_l^{obj}\}$ from cross-annotated data and $\{c_u^{bg}, c_u^{obj}\}$ from unlabeled data. We anticipate that the prototypes can well represent central tendencies and variability within each class, making them robust representatives of entire class distributions. Empirically, the quality of $\{c_l^{bg}, c_l^{obj}\}$ can be higher at the early stage thanks to the sparse but effective supervision. Consequently, we progressively fuse prototypes from cross-annotated data and unlabeled data, formulated as $c_f^{class} = \gamma c_l^{class} + (1-\gamma) c_u^{class}$, where $\gamma = 1/(1+\lambda_{pro})$ and λ_{pro} is a Gaussian ramp-up value that incrementally transitions from 0 to 1 during training, progressively giving more attention to the unlabeled prototypes. As such, we obtain the fused prototypes $\{c_f^{obj}, c_f^{bg}\}$. Then, we conduct a voxel-by-voxel comparison of these fused prototypes with the features $\{F^s_l, F^t_u\}$ to derive the prototype-drive predictions, denoted as $P_l^{pro,f}$ for cross-annotated data and $P_u^{pro,f}$ for unlabeled data. Specifically, $P_l^{pro,f}$ can be obtained with the formulation $P_l^{pro,f} = \frac{\exp\left(20 \cdot \cos\left(F^s_l, c_f^j\right)\right)}{\sum_{j \in \{obj,bg\}} \exp\left(20 \cdot \cos\left(F^s_l, c_f^j\right)\right)}$, where 'cos' denotes cosine similarity and '20' is a scaling factor [20]. Similarly, we can obtain $P_u^{pro,f}$ using F^t_u and $\{c_f^{obj}, c_f^{bg}\}$ for the unlabeled input. We promote consistency

between the prototype-driven predictions and the model's inductive outputs:

$$\mathcal{L}_{fpm} = \mathcal{L}_{\text{MAE}}(P_l^{pro,f}, \hat{Y}_l^{mix}) + \mathcal{L}_{\text{MAE}}(P_u^{pro,f}, P_u^t), \tag{2}$$

where we use the one-hot format for $\hat{Y}_l^{mix}$ and P_u^t is the classical model-based prediction for the unlabeled input derived by $\mathcal{F}_T$.

Perturbed Stability Constraint. We utilize the standard stability constraint as outlined by [17], due to its simplicity and effectiveness in enhancing the model's local smoothness and generalizability. Specifically, for the identical image X subjected to different perturbations ξ and ξ' (e.g., mild Gaussian noises [26, 28]), we aim to align the pre-softmax predictions of the segmentor with those of the EMA segmentor, formulated as $\mathcal{L}_{sta} = \mathcal{L}_{\text{MAE}}\left(\mathcal{F}_T^{\tilde{\theta}}(X+\xi), \mathcal{F}_S^{\theta}(X+\xi')\right)$.

Total Loss. Overall, the final training loss $\mathcal{L}$ is summarized as:

$$\mathcal{L} = \mathcal{L}_{sup}^l(\mathcal{D}^{ca}) + \lambda[\mathcal{L}_{NIOP}(\mathcal{D}^{ca}, \mathcal{D}^u) + \mathcal{L}_{fpm}(\mathcal{D}^{ca}, \mathcal{D}^u) + \mathcal{L}_{sta}(\mathcal{D}^{ca}, \mathcal{D}^u)], \tag{3}$$

where $\mathcal{L}_{sup}^l$ is the supervised loss for annotated voxels using partial cross-entropy loss; λ is a trade-off weight scheduled by a time-dependent Gaussian function $\lambda(t) = 0.1 \cdot e^{-5\left(1-\frac{t}{t_{max}}\right)^2}$, with $t_{\max}$ denoting the maximum training iteration.

3 Experiments

Dataset. We evaluate our method on the brain tumor segmentation dataset [2], comprising 335 3D preoperative magnetic resonance images (MRI) from glioma patients with modalities of T1, T1Gd, T2 and T2-FLAIR. Considering that T2-FLAIR can well characterize the malignant tumors [29], we adopt the T2-FLAIR images with paired ground truth of the entire tumor. The images are preprocessed to the resolution of $1 \times 1 \times 1\ mm^3$. We follow the same data split as in [26], where the train/val/test sets include 250/25/60 cases. Here, we define two budget settings, allowing 10% or 20% samples for cross-annotation. Therefore, 25 or 50 scans will be ultimately chosen, leading to 75 or 150 labeled slices (with 3 labeled slices per scan), corresponding to merely 0.22% or 0.43% of the effort required by a fully dense labeling strategy across all training data.

Implementation and Evaluation Metrics. The framework is implemented on PyTorch using an NVIDIA GeForce RTX 3090 GPU. We adopt the 3D V-Net [13] as the backbone. During training, we randomly crop patches of $96 \times 96 \times 96$ voxels as the input and use the sliding window strategy with stride of $64 \times 64 \times 64$ voxels for testing. The batch size is set to 4 including 2 cross-annotated data and 2 unlabeled data. The masking probability δ transits from 0.25 to 0.5 via a Gaussian ramp-up function and the masking patch size p is set to 8 voxels. $t_{\max}$ is set to 20,000. The initial learning rate is set to 0.01 and decayed with a power of 0.9 after each iteration. We apply random flipping and rotating for weak data augmentation. For a comprehensive evaluation, we adopt region-based metrics, Dice score and Jaccard, and boundary-based metrics, average surface distance (ASD) and 95% Hausdorff distance (95HD).

Table 1. Quantitative comparison. $*$ denotes dense labeling. Cross-subject standard deviations are shown in parentheses. AL: Active Learning. The best results are in bold.

Method	Setting		Metrics			
	L/U (%)	Labeled Slices	Dice (%) ↑	Jaccard (%) ↑	95HD (voxel) ↓	ASD (voxel) ↓
SupOnly	10%/0	75	67.21 (17.05)	52.99 (18.65)	**14.12** (11.93)	4.63 (2.83)
MT [17]	10%/90%	75	77.91 (17.92)	66.86 (20.92)	21.63 (24.71)	2.50 (1.88)
UA-MT [28]	10%/90%	75	78.43 (19.71)	67.97 (21.58)	19.07 (16.70)	3.14 (3.57)
CPS [5]	10%/90%	75	77.98 (19.20)	67.27 (21.63)	18.81 (19.34)	2.75 (2.53)
ICT [18]	10%/90%	75	77.18 (19.26)	66.22 (21.84)	21.12 (25.38)	2.70 (2.65)
CPCL [27]	10%/90%	75	78.95 (17.68)	68.47 (20.48)	17.14 (17.67)	3.02 (2.59)
CAML [6]	10%/90%	75	77.87 (14.65)	65.92 (18.10)	19.06 (21.32)	**2.38** (1.59)
ACMT [26]	10%/90%	75	76.68 (19.83)	65.70 (22.16)	19.81 (18.14)	3.23 (2.96)
UPCoL [12]	10%/90%	75	78.06 (18.98)	67.36 (21.64)	14.96 (17.80)	3.33 (3.39)
DeSCO [3]	10%/90%	75	75.32 (16.35)	64.08 (22.37)	22.34 (19.22)	3.19 (3.05)
MF-ConS	10%/90%	75	80.37 (14.46)	70.01 (15.39)	18.44 (18.65)	2.49 (1.53)
MF-ConS (+AL)	10%/90%	75	**81.26** (13.59)	**70.32** (16.58)	16.19 (20.01)	2.45 (1.57)
SupOnly	20%/0	150	67.91 (18.79)	54.18 (19.74)	12.14 (9.21)	4.45 (3.42)
MT [17]	20%/80%	150	77.12 (18.34)	65.90 (21.26)	**13.18** (15.27)	3.14 (2.83)
UA-MT [28]	20%/80%	150	78.16 (13.07)	68.61 (17.15)	14.55 (16.13)	2.77 (1.91)
CPS [5]	20%/80%	150	79.83 (16.09)	69.08 (19.96)	15.62 (17.72)	2.52 (2.10)
ICT [18]	20%/80%	150	78.59 (13.28)	66.54 (16.63)	19.39 (24.70)	2.71 (1.81)
CPCL [27]	20%/80%	150	81.32 (14.01)	71.34 (17.91)	15.96 (21.32)	2.24 (1.64)
CAML [6]	20%/80%	150	76.95 (16.60)	65.16 (19.60)	18.88 (22.43)	2.88 (2.26)
ACMT [26]	20%/80%	150	79.76 (21.00)	67.77 (21.22)	19.02 (21.91)	2.86 (2.93)
UPCoL [12]	20%/80%	150	81.58 (12.49)	71.86 (16.19)	17.38 (20.25)	2.36 (1.57)
DeSCO [3]	20%/80%	150	78.03 (18.42)	66.43 (20.53)	17.97 (16.29)	4.33 (3.10)
MF-ConS	20%/80%	150	83.95 (11.43)	73.24 (15.28)	15.81 (21.38)	2.19 (1.12)
MF-ConS (+AL)	20%/80%	150	**84.20** (10.71)	**73.99** (13.97)	14.77 (22.19)	**2.14** (1.35)
SupOnly (upper bound)	100%$*$/0	34173	87.07 (7.90)	77.48 (11.45)	7.84 (8.09)	1.79 (1.49)

Comparison with State-of-the-art Methods. Table 1 presents the results of different approaches, wherein only 10% and 20% training samples undergo cross-annotation. We include recent state-of-the-art semi-/barely-supervised methods [17, 28, 5, 18, 27, 6, 26, 12, 3] for comparison. The backbone and training protocols are consistent to ensure fairness. The results are the average over three runs to mitigate the variability in results due to online sampling. As observed, our MS-ConS without active learning achieves $\{13.16\%, 16.04\%\}$ Dice improvements under $\{10\%, 20\%\}$ cross-annotated settings compared to supervised-only (SupOnly) baselines, showing its effectiveness in leveraging both cross-annotated images and unlabeled images. Compared to competing approaches, MS-ConS consistently yields notable improvements, showing the efficacy of multi-faceted consistency learning in regularizing model training under sparse supervision. Note that DeSCO [3] initially introduced cross-annotation as an efficient sparse labeling strategy, employing the off-the-shelf image registration technique [1] for label propagation. However, its performance in this task was found to be mediocre, potentially due to suboptimal registration quality when dealing with the heterogeneous characteristics of brain tumors. We can also notice that MF-ConS further improves when integrated with active learning for the identification and annotation of informative samples. Fig. 2 (e) presents exemplar 2D segmentation results under the 10% cross-labeled setting. Consistently, the results of our MF-ConS (+AL) align more precisely with the ground-truth masks. The highlighted boxes reflect the reduction in false positive rates, underscoring the practical effectiveness of our approach in tumor segmentation.

Ablation Analysis. To help better understand the multi-faceted consistency learning, we visualize each consistency type in Fig. 2 (a-c). Fig. 2 (d) presents the

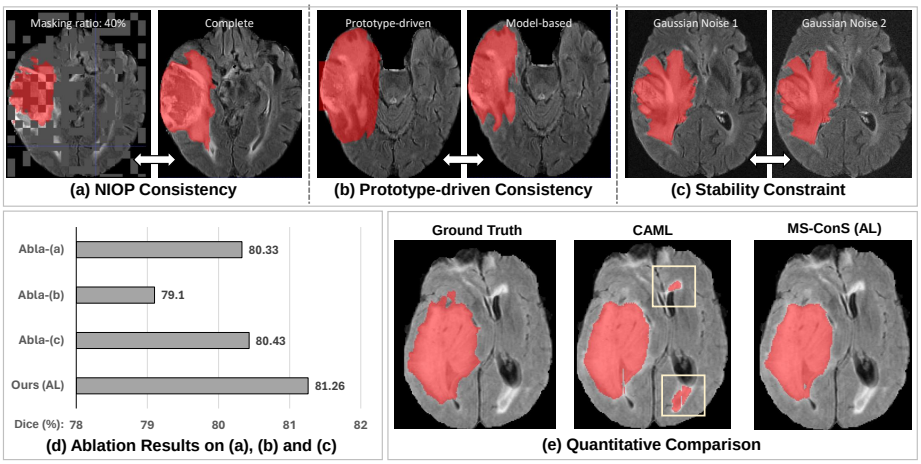

Fig. 2. (a-c) Visualization of three types of consistency regularization. (d) Corresponding ablation study on different types of consistency regularization. (e) Exemplar 2D segmentation results with 10% cross-annotation budget.

ablation results quantitatively under the setting of 10% cross-annotation budget, where Abla-(a), Abla-(b), and Abla-(c) denote the removal of the corresponding consistency in Fig. 2 (a-c), i.e., $\mathcal{L}_{NIOP}$, $\mathcal{L}_{fpm}$ and $\mathcal{L}_{sta}$. Note that the ablation experiments all correspond to our best version, i.e., MF-ConS (+AL), and thus all incorporate active learning. Specifically, Fig. 2 (a) illustrates how NIOP Consistency, even with a case of masking ratio of 40%, allows the model to capture and infer critical structural details for the high-level perception task, i.e., tumor segmentation. Similar to typical masked image modeling [7], this regularization also serves as self-supervised signals but with high-level task-specific nature. Meanwhile, it can be observed that prototype-driven consistency shown in Fig. 2 (b) plays an important role in enhancing discriminative and compact feature learning by closely aligning the model-based output with the non-parametric prototype-driven output. Fig. 2 (c) illustrates the typical stability constraint in maintaining the segmentation robustness to mild hand-crafted perturbations such as Gaussian noise. The ablation result also demonstrates its necessity. Our complete model, incorporating all three consistency mechanisms along with active learning, achieves the best performance, demonstrating the collective strength of the multi-faceted consistency learning approach in our sparse annotation setting.

4 Conclusion

In this paper, we presented a Multi-Faceted ConSistency (MF-ConS) learning framework for active barely-supervised 3D medical image segmentation. Starting with an efficient active cross-annotation strategy, MF-ConS exploits both limited cross-labeled data and abundant unlabeled data via three types of consis-

tency regularization, including neighbor-informed object prediction consistency, non-parametric prototype-driven consistency and perturbed stability constraint, effectively addressing the inherent challenges of sparse supervision including limited fine-grained object perception, less compact class features and inferior generalizability. We evaluated our method on brain tumor segmentation and demonstrated its superior performance compared to other state-of-the-art approaches.

Acknowledgments. This research was partly supported by General Research Fund (No. 14205419) and Hong Kong PhD Fellowship from Research Grants Council of Hong Kong.

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

References

1. Avants, B.B., Tustison, N., Song, G., et al.: Advanced normalization tools (ants). Insight j **2**(365), 1–35 (2009)
2. Bakas, S.: BraTS MICCAI brain tumor dataset (2020). https://doi.org/10.21227/hdtd-5j88
3. Cai, H., Li, S., Qi, L., Yu, Q., Shi, Y., Gao, Y.: Orthogonal annotation benefits barely-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3302–3311 (2023)
4. Cai, H., Qi, L., Yu, Q., Shi, Y., Gao, Y.: 3d medical image segmentation with sparse annotation via cross-teaching between 3d and 2d networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 614–624. Springer (2023)
5. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2613–2622 (2021)
6. Gao, S., Zhang, Z., Ma, J., Li, Z., Zhang, S.: Correlation-aware mutual learning for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 98–108. Springer (2023)
7. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
8. Lee, J., Kim, E., Lee, S., Lee, J., Yoon, S.: Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5267–5276 (2019)
9. Li, S., Cai, H., Qi, L., Yu, Q., Shi, Y., Gao, Y.: Pln: Parasitic-like network for barely supervised medical image segmentation. IEEE Transactions on Medical Imaging **42**(3), 582–593 (2022)
10. Li, Y., Zhao, H., Qi, X., Chen, Y., Qi, L., Wang, L., Li, Z., Sun, J., Jia, J.: Fully convolutional networks for panoptic segmentation with point-based supervision. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4552–4568 (2022)

11. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3159–3167 (2016)
12. Lu, W., Lei, J., Qiu, P., Sheng, R., Zhou, J., Lu, X., Yang, Y.: Upcol: Uncertainty-informed prototype consistency learning for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 662–672. Springer (2023)
13. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: Fourth International Conference on 3D Vision. pp. 565–571. IEEE (2016)
14. Oh, Y., Kim, B., Ham, B.: Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6913–6922 (2021)
15. Ouali, Y., Hudelot, C., Tami, M.: Semi-supervised semantic segmentation with cross-consistency training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12674–12684 (2020)
16. Pan, W., Xu, Z., Yan, J., Wu, Z., Tong, R.K.y., Li, X., Yao, J.: Semi-supervised semantic segmentation meets masked modeling: Fine-grained locality learning matters in consistency regularization. arXiv preprint arXiv:2312.08631 (2023)
17. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems. pp. 1195–1204 (2017)
18. Verma, V., Kawaguchi, K., Lamb, A., Kannala, J., Solin, A., Bengio, Y., Lopez-Paz, D.: Interpolation consistency training for semi-supervised learning. Neural Networks **145**, 90–106 (2022)
19. Wang, H., Chen, J., Zhang, S., He, Y., Xu, J., Wu, M., He, J., Liao, W., Luo, X.: Dual-reference source-free active domain adaptation for nasopharyngeal carcinoma tumor segmentation across multiple hospitals. IEEE Transactions on Medical Imaging (2024)
20. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: PANet: Few-shot image semantic segmentation with prototype alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9197–9206 (2019)
21. Wu, L., Fang, L., He, X., He, M., Ma, J., Zhong, Z.: Querying labeled for unlabeled: Cross-image semantic consistency guided semi-supervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(7), 8827–8844 (2023)
22. Wu, Y., Xu, M., Ge, Z., Cai, J., Zhang, L.: Semi-supervised left atrium segmentation with mutual consistency training. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 297–306. Springer (2021)
23. Xie, Q., Dai, Z., Hovy, E., Luong, T., Le, Q.: Unsupervised data augmentation for consistency training. Advances in Neural Information Processing Systems **33**, 6256–6268 (2020)
24. Xu, Z., Lu, D., Luo, J., Zheng, Y., Tong, R.K.y.: Separated collaborative learning for semi-supervised prostate segmentation with multi-site heterogeneous unlabeled mri data. Medical Image Analysis **93**, 103095 (2024)
25. Xu, Z., Lu, D., Yan, J., Sun, J., Luo, J., Wei, D., Frisken, S., Li, Q., Zheng, Y., Tong, R.K.y.: Category-level regularized unlabeled-to-labeled learning for semi-supervised prostate segmentation with multi-site unlabeled data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 3–13. Springer (2023)

26. Xu, Z., Wang, Y., Lu, D., Luo, X., Yan, J., Zheng, Y., Tong, R.K.y.: Ambiguity-selective consistency regularization for mean-teacher semi-supervised medical image segmentation. Medical Image Analysis p. 102880 (2023)
27. Xu, Z., Wang, Y., Lu, D., Yu, L., Yan, J., Luo, J., Ma, K., Zheng, Y., Tong, R.K.y.: All-around real label supervision: Cyclic prototype consistency learning for semi-supervised medical image segmentation. IEEE Journal of Biomedical and Health Informatics **26**(7), 3174–3184 (2022)
28. 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)
29. Zeineldin, R.A., Karar, M.E., Coburger, J., Wirtz, C.R., Burgert, O.: DeepSeg: deep neural network framework for automatic brain tumor segmentation using magnetic resonance FLAIR images. International Journal of Computer Assisted Radiology and Surgery **15**(6), 909–920 (2020)
30. Zhou, M., Xu, Z., Zhou, K., Tong, R.K.y.: Weakly supervised medical image segmentation via superpixel-guided scribble walking and class-wise contrastive regularization. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 137–147. Springer (2023)