LEARNING FROM UNCERTAIN AGREEMENT AND INFORMATIVE DISAGREEMENT FOR SEMI-SUPERVISED 3D MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Teacher–student learning is widely used for semi-supervised 3D medical image segmentation, yet effectively exploiting uncertain and conflicting teacher predictions remains challenging. Directly enforcing agreement may preserve ambiguity, whereas suppressing difficult regions can discard useful supervisory cues. To address this dilemma, we propose a reflective teacher–student framework that interprets teacher guidance according to predictive state and disagreement behavior. Specifically, Consensus Sharpening and Disagreement Reconciliation (CSDR) distinguishes uncertain agreement from informative disagreement and applies state-adaptive supervision to refine tentative consensus and reconcile meaningful conflicts. Teacher-Perturbation Disagreement Probing (TPDP) further characterizes disagreement through its responses to controlled teacher-probability perturbations and uses these responses to modulate teacher-guided learning. Experiments on BraTS2019, LA, and Pancreas-NIH demonstrate consistent improvements under limited-label settings, achieving the highest Dice score across all six evaluated labeling settings. The code will be publicly released upon acceptance.

1 INTRODUCTION

Three-dimensional (3D) medical image segmentation is essential for computer-aided diagnosis and treatment planning, yet high-performing models still depend heavily on costly voxel-level annotations Azad et al. (2024); Niyas et al. (2022). The burden is amplified in 3D images, where complex structures and ambiguous boundaries require labor-intensive expert delineation Zhang et al. (2025b). Semi-supervised learning (SSL) alleviates this bottleneck by exploiting abundant unlabeled volumes alongside limited labeled data.

The success of SSL largely depends on how effectively supervision is extracted from unlabeled data. Teacher–student learning is widely used for this purpose, with teacher predictions serving as supervisory targets for the student Tarvainen & Valpola (2017); Yu et al. (2019). However, teacher guidance is often unreliable in difficult regions such as ambiguous boundaries and small anatomical structures, where predictions tend to be uncertain or conflicting. Existing methods therefore either suppress unreliable teacher signals through uncertainty-aware selection or emphasize difficult but potentially informative regions Yu et al. (2019); Xu et al. (2023). A key dilemma remains: enforcing agreement with uncertain teacher predictions may preserve ambiguity, whereas suppressing difficult regions may discard useful supervisory cues. This motivates supervision that adapts to different predictive states rather than treating teacher guidance uniformly.

The relation between student and teacher predictions, whether they agree or disagree, and their associated uncertainty jointly characterizes the predictive state of an unlabeled voxel. Considering these two dimensions together reveals two particularly learning-relevant states that call for different forms of supervision. When the two networks predict the same class but one or both remain uncertain, the resulting *uncertain agreement* represents a tentative consensus that remains probabilistically ambiguous. Simply preserving such agreement may retain its ambiguity; the tentative consensus should instead be sharpened to reduce ambiguity and form a more confident prediction. Conversely, disagreement is not necessarily uninformative. When at least one network retains a relatively confident prediction, the conflict may still preserve useful discriminative cues Zhao et al.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(2025), giving rise to *informative disagreement*. Such disagreement should be reconciled progressively by accounting for the relative confidence of the two predictions, rather than being simply suppressed.

Beyond determining how different predictive states should be treated, a complementary question remains: is a single student-teacher comparison sufficient to characterize disagreement? In conventional teacher-student supervision, disagreement is typically measured from the current student-teacher prediction pair, providing only an instantaneous view of their predictive discrepancy. Yet two voxels with similar instantaneous discrepancies may respond very differently when the teacher probability distribution is perturbed. Such response differences reveal complementary information about the persistence and sensitivity of disagreement under variations in the teacher probability distribution, which cannot be captured by a single comparison alone. This motivates us to characterize disagreement not only by its instantaneous magnitude, but also by its response to variations in the teacher probability distribution.

Together, these observations raise a broader question: *How can a student learn from its teacher without simply imitating it?* We argue that an effective student should not only refine uncertain consensus and reason about informative disagreement, but also reconsider the teacher's guidance from multiple plausible perspectives. To this end, we propose a reflective teacher–student learning framework for semi-supervised 3D medical image segmentation. Its central idea is to combine state-adaptive supervision with disagreement probing. Specifically, we introduce a Consensus Sharpening and Disagreement Reconciliation (CSDR) mechanism that distinguishes predictive states based on student–teacher agreement/disagreement and predictive confidence. For uncertain agreement, CSDR constructs a shared student–teacher target and sharpens it to consolidate tentative consensus; for informative disagreement, it progressively shifts from entropy-based ambiguity reduction to confidence-weighted soft alignment. We further introduce a Teacher-Perturbation Disagreement Probing (TPDP) mechanism that perturbs the teacher probability distribution and aggregates the resulting disagreement responses to adaptively modulate supervision from the original teacher prediction.

The main contributions of this work are summarized as follows:

(1) We formulate semi-supervised teacher–student learning as a reflective process, where teacher guidance is interpreted according to predictive state and disagreement behavior rather than being uniformly imitated or discarded.

(2) We propose Consensus Sharpening and Disagreement Reconciliation (CSDR), a state-adaptive supervision mechanism that distinguishes uncertain agreement from informative disagreement, consolidating tentative consensus while progressively reconciling meaningful conflict.

(3) We propose Teacher-Perturbation Disagreement Probing (TPDP), which characterizes disagreement through its responses to teacher perturbations rather than a single prediction pair, enabling response-aware emphasis on discrepancies that remain pronounced or vary sensitively across perturbations.

2 Related Work

2.1 Semi-Supervised Medical Image Segmentation

Semi-supervised medical image segmentation seeks to reduce the reliance on dense voxel-level annotations by exploiting unlabeled data together with limited labeled samples He et al. (2023); Jiao et al. (2024); Wu & Zhuang (2022). Most existing methods build upon pseudo-labeling and consistency regularization, often combined with complementary strategies such as co-training, perturbation, prototype learning, and data mixing. Pseudo-labeling methods Chaitanya et al. (2023); Liu et al. (2024); Su et al. (2024); Peng et al. (2023) convert model predictions on unlabeled images into additional training targets, whereas consistency-based approaches Li et al. (2024); Zhang et al. (2023a;b) exploit prediction invariance across different perturbations, views, or model states. Despite substantial progress, the effectiveness of these approaches remains strongly dependent on the quality and utilization of supervisory signals derived from unlabeled data. This has motivated increasing attention to how uncertain, ambiguous, or conflicting predictions should be interpreted and exploited during semi-supervised learning.

2.2 Uncertainty- and Disagreement-Aware Teacher–Student Learning

Teacher–student learning is widely used to exploit unlabeled data, with Mean Teacher and its variants providing temporally aggregated predictions to guide the student Deng et al. (2025); Zhang et al. (2025a); Ge et al. (2025). Existing methods improve teacher supervision through uncertainty estimation, confidence-aware weighting, ambiguity selection, multi-view prediction, and cross-model collaboration Zeng et al. (2025c); Huang et al. (2025); Wang et al. (2022); Wu et al. (2026). Uncertainty-aware approaches typically suppress unreliable predictions, whereas ambiguity-oriented methods emphasize difficult but potentially informative regions Yu et al. (2019); Xu et al. (2023). Disagreement offers complementary information, and recent methods exploit prediction diversity or conflicting outputs across models and perturbations Zhao et al. (2025). However, uncertainty and disagreement are still mainly used for target selection, weighting, or consistency enforcement, with limited attention to their joint role in characterizing predictive states or to how disagreement responds to variations in teacher predictions. Our work addresses these aspects by differentiating supervision for uncertain agreement and informative disagreement and by probing disagreement responses under teacher-probability perturbations.

3 Methodology

3.1 Overview

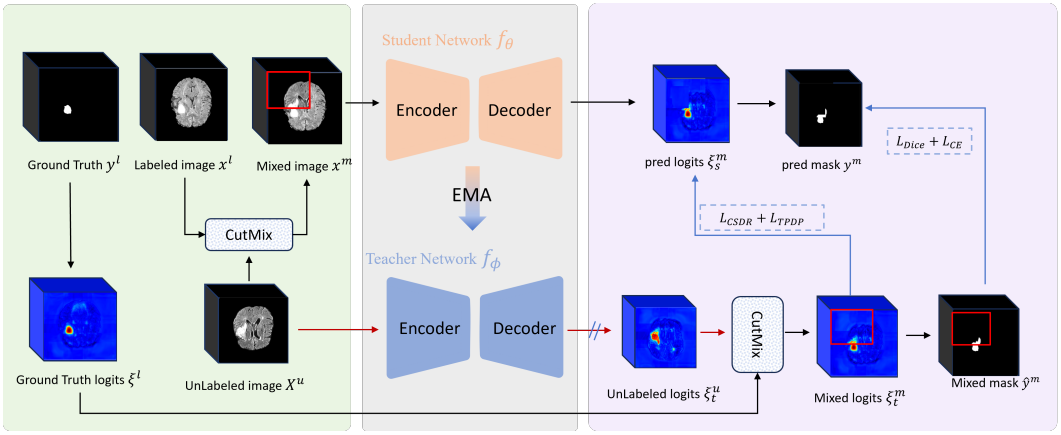


Figure 1: Overview of the proposed reflective teacher–student learning framework.

As illustrated in Fig. 1, the proposed framework follows a teacher–student architecture consisting of a student network f_θ and a teacher network f_ϕ, with the teacher updated as an exponential moving average (EMA) of the student Tarvainen & Valpola (2017). Both networks are initialized from labeled-data pretraining before semi-supervised learning. During semi-supervised training, a labeled volume and an unlabeled volume are combined through 3D CutMix Yun et al. (2019) to construct the mixed input for the student. Using the same mixing mask, the mixed target combines the ground-truth annotation in the labeled region with the teacher prediction in the unlabeled region. The resulting target supervises the student prediction through the Dice and cross-entropy losses, thereby integrating labeled and teacher-guided supervision within a unified mixed sample.

On unlabeled predictions, we introduce two complementary mechanisms for exploiting teacher guidance. Consensus Sharpening and Disagreement Reconciliation (CSDR) provides state-adaptive supervision by jointly considering student–teacher prediction relations and predictive confidence. Teacher-Perturbation Disagreement Probing (TPDP) complements this process by probing how disagreement responds to variations in the teacher probability distribution and using these responses to modulate teacher-guided learning. As illustrated in Fig. 1, the supervisory signals produced by CSDR and TPDP are incorporated into student optimization through the corresponding objectives $\mathcal{L}_{\text{CSDR}}$ and $\mathcal{L}_{\text{TPDP}}$. Their detailed formulations are presented in Sections 3.2 and 3.3, with the overall training objective summarized in Section 3.4.

3.2 Consensus Sharpening and Disagreement Reconciliation

We introduce Consensus Sharpening and Disagreement Reconciliation (CSDR) to adapt teacher–student supervision to different predictive states. In particular, CSDR focuses on uncertain agreement, where the two networks predict the same class but remain probabilistically ambiguous, and informative disagreement, where their predicted classes differ while at least one prediction retains meaningful confidence.

3.2.1 Predictive-State Characterization

For an unlabeled voxel v, let $p^s(v), p^t(v) \in \mathbb{R}^C$ denote the class-probability predictions of the student and teacher, respectively, where C is the number of classes. Their hard predictions are given by

$$\hat{c}^s(v) = \arg\max_c p_c^s(v), \qquad \hat{c}^t(v) = \arg\max_c p_c^t(v). \tag{1}$$

Accordingly, the student–teacher agreement map is defined as

$$A(v) = \mathbb{I}\left[\hat{c}^s(v) = \hat{c}^t(v)\right], \tag{2}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function.

We quantify voxel-wise predictive uncertainty for the student and teacher using normalized Shannon entropy:

$$H\big(p^k(v)\big) = -\frac{1}{\log C}\sum_{c=1}^{C} p_c^k(v)\log\big(p_c^k(v)+\epsilon\big), \qquad k \in \{s,t\}, \tag{3}$$

where $p_c^k(v)$ denotes the predicted probability of class c at voxel v from network k, and ϵ is used for numerical stability. This yields voxel-wise uncertainty maps for both the student and teacher, with larger values indicating greater predictive uncertainty.

We first construct two joint uncertainty masks for the student–teacher pair. M_{LL} identifies voxels where both predictions have low uncertainty (i.e., low entropy), whereas M_{HH} identifies voxels where both predictions have high uncertainty (i.e., high entropy):

$$M_{\mathrm{LL}}(v) = \mathbb{I}[H(p^s(v)) < \tau_L(e)]\,\mathbb{I}\big[H\big(p^t(v)\big) < \tau_L(e)\big], \tag{4}$$

and

$$M_{\mathrm{HH}}(v) = \mathbb{I}[H(p^s(v)) > \tau_H(e)]\,\mathbb{I}\big[H\big(p^t(v)\big) > \tau_H(e)\big], \tag{5}$$

where $\tau_L(e)$ and $\tau_H(e)$ denote epoch-dependent entropy thresholds for identifying low- and high-uncertainty predictions, respectively.

Combining the agreement map $A(v)$ with these joint uncertainty masks, we define the uncertain-agreement and informative-disagreement masks as

$$M_{\mathrm{UA}}(v) = A(v)\,(1 - M_{\mathrm{LL}}(v)), \tag{6}$$

and

$$M_{\mathrm{ID}}(v) = (1 - A(v))\,(1 - M_{\mathrm{HH}}(v)), \tag{7}$$

respectively. Thus, M_{UA} selects voxels where the student and teacher agree in their hard predictions but are not both confident, whereas M_{ID} selects disagreement voxels where the two predictions are not simultaneously highly uncertain.

3.2.2 Uncertain Agreement

For voxels identified by M_{UA}, the student and teacher produce the same hard prediction, but at least one prediction remains uncertain. Such agreement therefore represents only a tentative consensus rather than a confident decision. We therefore construct a shared student–teacher target and sharpen it to reinforce their common prediction tendency.

Let $p_T^s(v)$ and $p_T^t(v)$ denote the temperature-scaled student and teacher probability distributions, respectively, where temperature scaling moderates overly confident individual predictions before their fusion. Their shared prediction is defined as

$$\bar{p}(v) = \frac{1}{2}\left(p_T^s(v) + p_T^t(v)\right). \tag{8}$$

We then sharpen the shared distribution using a power transformation followed by normalization:

$$q_c^{\mathrm{UA}}(v) = \frac{\bar{p}_c^2(v)}{\sum_{j=1}^{C} \bar{p}_j^2(v)}, \tag{9}$$

where $q^{\mathrm{UA}}(v)$ denotes the sharpened target. This transformation amplifies the relative differences among class probabilities, producing a more peaked distribution while preserving the class ordering jointly expressed by the student and teacher.

The uncertain-agreement objective is formulated as

$$\mathcal{L}_{\mathrm{UA}} = \frac{1}{|M_{\mathrm{UA}}|} \sum_v M_{\mathrm{UA}}(v)\, \mathrm{KL}\left(q^{\mathrm{UA}}(v) \parallel p_T^s(v)\right), \tag{10}$$

which encourages the student prediction to approach the sharpened shared target.

3.2.3 Informative Disagreement

For voxels identified by M_{ID}, the student and teacher predict different classes, while the two predictions are not simultaneously highly uncertain. Such disagreement may still contain useful discriminative information rather than representing purely unreliable conflict. We therefore reconcile the competing predictions through a progressive transition from ambiguity reduction to confidence-aware soft alignment.

We first estimate the voxel-wise confidence of the student and teacher predictions from their normalized entropy:

$$r_s(v) = 1 - H(p^s(v)), \qquad r_t(v) = 1 - H\left(p^t(v)\right). \tag{11}$$

Their relative confidence weights are then defined as

$$w_s(v) = \frac{r_s(v)}{r_s(v) + r_t(v) + \epsilon}, \qquad w_t(v) = \frac{r_t(v)}{r_s(v) + r_t(v) + \epsilon}, \tag{12}$$

where ϵ is used for numerical stability. Based on these weights, a confidence-aware soft target is constructed from the temperature-scaled predictions:

$$q^{\mathrm{ID}}(v) = w_s(v) p_T^s(v) + w_t(v) p_T^t(v). \tag{13}$$

This target preserves information from both networks while assigning greater influence to the more confident prediction. The corresponding alignment objective is

$$\mathcal{L}_{\mathrm{ID}}^{\mathrm{align}} = \frac{1}{|M_{\mathrm{ID}}|} \sum_v M_{\mathrm{ID}}(v)\, \mathrm{KL}\left(q^{\mathrm{ID}}(v) \parallel p_T^s(v)\right). \tag{14}$$

At early training stages, however, the confidence estimates themselves may still be unreliable. We therefore additionally reduce prediction ambiguity by minimizing the entropy of the average student–teacher distribution:

$$p^{\mathrm{avg}}(v) = \frac{1}{2}\left(p_T^s(v) + p_T^t(v)\right), \tag{15}$$

yielding the entropy-based ambiguity-reduction term

$$\mathcal{L}_{\mathrm{ID}}^{\mathrm{ent}} = \frac{1}{|M_{\mathrm{ID}}|} \sum_v M_{\mathrm{ID}}(v) H(p^{\mathrm{avg}}(v)). \tag{16}$$

The two objectives are progressively balanced according to the training progress:

$$\mathcal{L}_{\mathrm{ID}} = (1-\gamma)\mathcal{L}_{\mathrm{ID}}^{\mathrm{align}} + \gamma \mathcal{L}_{\mathrm{ID}}^{\mathrm{ent}}, \tag{17}$$

where

$$\gamma = 1 - \frac{e}{E}, \tag{18}$$

and e and E denote the current and total training epochs, respectively. Consequently, entropy-based ambiguity reduction is emphasized at the early stage, when prediction confidence is less reliable, whereas confidence-weighted alignment gradually becomes dominant as training proceeds. This progressive transition avoids prematurely forcing one prediction to follow the other while ultimately promoting a more reliable consensus.

Combining the uncertain-agreement and informative-disagreement objectives, the complete CSDR objective is

$$\mathcal{L}_{\mathrm{CSDR}} = \mathcal{L}_{\mathrm{UA}} + \mathcal{L}_{\mathrm{ID}}. \tag{19}$$

3.3 TEACHER-PERTURBATION DISAGREEMENT PROBING

A single student–teacher comparison provides only an instantaneous view of predictive disagreement. To obtain a richer characterization, we introduce Teacher-Perturbation Disagreement Probing (TPDP), which probes how student–teacher disagreement responds to controlled variations in the teacher probability distribution.

Let $p^s(v)$ and $p^t(v)$ denote the student and teacher probability distributions at voxel v, respectively. TPDP generates K perturbed teacher predictions by adding voxel-wise Gaussian perturbations:

$$\tilde{p}_k^t(v) = \mathrm{Normalize}\left(\mathrm{Clip}\left(p^t(v) + \delta\xi_k(v)\right)\right), \qquad k = 1, \ldots, K, \tag{20}$$

where $\xi_k(v)$ denotes the Gaussian perturbation for the k-th probe and δ controls the perturbation strength. The clipping and normalization operations ensure that each perturbed prediction remains a valid probability distribution.

For each perturbed teacher prediction, we measure its voxel-wise disagreement with the student using KL divergence:

$$R_k(v) = \mathrm{KL}\left(\tilde{p}_k^t(v) \parallel p^s(v)\right), \tag{21}$$

where $R_k(v)$ represents the disagreement response to the k-th teacher perturbation. We summarize these responses by their mean and variance:

$$\mu_R(v) = \frac{1}{K}\sum_{k=1}^{K} R_k(v), \qquad \sigma_R^2(v) = \mathrm{Var}_k\left(R_k(v)\right). \tag{22}$$

The mean characterizes the average magnitude of disagreement across teacher perturbations, whereas the variance reflects its sensitivity to these perturbations. Accordingly, a large mean indicates that disagreement remains pronounced across teacher variations, while a large variance indicates that it changes sensitively with the teacher probability distribution.

We combine the two response statistics into a disagreement-probing score:

$$S(v) = \lambda_\mu \mu_R(v) + \lambda_\sigma \sigma_R^2(v), \tag{23}$$

where λ_μ and λ_σ control the contributions of the mean and variance terms, respectively.

Importantly, the perturbed teacher predictions are used only to probe disagreement rather than as supervisory targets. The original teacher prediction is retained as the target, while the probing score is detached from the computational graph and used to modulate the voxel-wise supervision:

$$\mathcal{L}_{\mathrm{TPDP}} = \frac{1}{|\Omega|}\sum_{v\in\Omega}\left(1 + S(v)\right)\mathrm{KL}\left(p^t(v) \parallel p^s(v)\right), \tag{24}$$

where Ω denotes the set of voxels in the input volume. In this way, TPDP places greater learning emphasis on discrepancies that remain pronounced or vary sensitively under teacher perturbations, while preserving the original teacher prediction as the supervisory reference.

3.4 TRAINING OBJECTIVE

Training consists of labeled-data pretraining followed by semi-supervised optimization. During pretraining, 3D CutMix is applied to labeled volumes, and the network is optimized using the sum of Dice and cross-entropy losses. The resulting parameters are used to initialize both the student and teacher networks.

During semi-supervised training, a labeled volume and an unlabeled volume are combined through 3D CutMix, with the mixed target consisting of ground-truth annotations in the labeled region and teacher predictions in the unlabeled region. Let y^m and $\hat{y}^m$ denote the mixed target and the corresponding student prediction, respectively. The mixed segmentation objective is

$$\mathcal{L}_{\mathrm{mix}} = \mathcal{L}_{\mathrm{Dice}}\left(\hat{y}^m, y^m\right) + \mathcal{L}_{\mathrm{CE}}\left(\hat{y}^m, y^m\right). \tag{25}$$

The overall semi-supervised objective is

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{mix}} + \lambda_{\mathrm{CSDR}}\mathcal{L}_{\mathrm{CSDR}} + \lambda_{\mathrm{TPDP}}\mathcal{L}_{\mathrm{TPDP}}, \tag{26}$$

where λ_{CSDR} and λ_{TPDP} control the contributions of CSDR and TPDP, respectively. The student is optimized by backpropagation, while the teacher is updated through EMA.

Table 1: Quantitative comparison on the LA dataset.

Method	LA (10% / 8 labeled data)				LA (20% / 16 labeled data)			
	Dice ↑	Jaccard ↑	ASD ↓	HD95 ↓	Dice ↑	Jaccard ↑	ASD ↓	HD95 ↓
UA-MT (MICCAI'19)	86.15	76.04	4.77	18.86	88.87	80.19	2.29	7.41
CPS (CVPR'21)	85.39	75.47	4.86	22.23	87.87	78.61	2.16	12.87
DTC (AAAI'21)	86.98	78.06	3.35	8.39	89.42	80.98	2.10	7.32
CPCL (JBHI'22)	85.96	78.03	3.74	9.05	88.32	81.12	2.02	8.05
URPC (MedIA'22)	86.07	78.15	3.93	9.26	88.46	81.16	2.23	8.21
BCP (CVPR'23)	89.38	82.13	2.76	8.27	90.12	82.13	2.55	7.69
CML (MM'24)	90.24	82.28	1.76	6.11	88.95	80.45	1.82	9.07
UniMatch (CVPR'23)	89.32	83.85	1.86	6.63	91.02	83.52	1.61	6.13
LeFeD (TMI'25)	90.12	81.78	1.74	7.02	91.39	84.17	1.47	6.15
UnCo (TMI'25)	90.37	82.54	1.74	6.11	90.91	83.40	1.61	5.36
PICK (IJCV'25)	88.58	79.69	2.32	9.74	90.85	83.37	1.74	5.70
CMFN (JBHI'26)	90.29	**84.40**	1.77	**5.52**	91.36	84.74	1.47	5.23
Ours	**90.81**	83.25	**1.58**	6.07	**91.85**	**84.99**	**1.42**	**5.00**

4 EXPERIMENTS

4.1 DATASETS AND EXPERIMENTAL SETUP

We evaluate the proposed method on three public 3D medical image segmentation datasets: **BraTS2019** Menze et al. (2014), **LA** Xiong et al. (2021), and **Pancreas-NIH** Roth et al. (2015). BraTS2019 contains 335 MRI scans, of which 250, 25, and 60 cases are used for training, validation, and testing, respectively. LA consists of 100 3D MRI scans for left-atrium segmentation, while Pancreas-NIH contains 82 3D CT scans for pancreas segmentation. For **LA**, 10% and 20% of the training data are used as labeled samples. The model is pretrained for $2{,}000$ iterations and subsequently trained for $15{,}000$ iterations in the semi-supervised stage, with an input patch size of $112 \times 112 \times 80$. For **BraTS2019**, 10% and 20% of the training data are used as labeled samples, with $5{,}000$ pretraining iterations followed by $30{,}000$ semi-supervised training iterations using an input patch size of $96 \times 96 \times 96$. For **Pancreas-NIH**, 10% and 20% of the training data are used as labeled samples, with $3{,}500$ pretraining iterations and $20{,}000$ semi-supervised training iterations using an input patch size of $96 \times 96 \times 96$.

For a fair comparison, the same segmentation backbone is used for our method and all compared methods. Specifically, 3D U-Net Çiçek et al. (2016) is adopted on BraTS2019, while V-Net Milletari et al. (2016) is used on LA and Pancreas-NIH. The proposed method is implemented in PyTorch and trained on an NVIDIA A100-SXM4-80GB GPU. Following previous studies Wu et al. (2022); Shi et al. (2021); Luo et al. (2022), the mask ratio of the 3D CutMix strategy is set to $\beta = 2/3$.

Segmentation performance is evaluated using the Dice similarity coefficient (Dice), Jaccard index (Jaccard), 95% Hausdorff distance (HD95), and average surface distance (ASD). Higher Dice and Jaccard values indicate better overlap, whereas lower HD95 and ASD values indicate better boundary agreement. All results are computed from the final predictions of the student network.

4.2 COMPARISON WITH STATE-OF-THE-ART METHODS

We compare the proposed method with representative state-of-the-art semi-supervised medical image segmentation methods, including UA-MT Yu et al. (2019), CPS Chen et al. (2021), DTC Luo et al. (2021), CPCL Xu et al. (2022), URPC Luo et al. (2022), BCP Bai et al. (2023), CML Wu et al. (2024), ICT Verma et al. (2022), CauSSL Miao et al. (2023), UniMatch Yang et al. (2023), LeFeD Zeng et al. (2025b), UnCo Zeng et al. (2025c), PICK Zeng et al. (2025a), CMFN Zhou et al. (2026), MCF Wang et al. (2023), EIC Huang et al. (2024), MLRP Su et al. (2024), and SKCDF Zhang et al. (2025c). Quantitative results on LA, BraTS2019, and Pancreas-NIH are reported in Tables 1–3. The best and second-best results in each column are highlighted in bold and underlined, respectively.

Quantitative results. The proposed method achieves the highest Dice score under all six labeled-data settings. Compared with the second-best methods, the Dice improvements are 0.44 and 0.46 percentage points on LA, 0.55 and 0.04 percentage points on BraTS2019, and 4.39 and 1.31 per-

Table 2: Quantitative comparison on the BraTS2019 dataset.

Method	BraTS2019 (10% / 25 labeled data)				BraTS2019 (20% / 50 labeled data)			
	Dice ↑	Jaccard ↑	ASD ↓	HD95 ↓	Dice ↑	Jaccard ↑	ASD ↓	HD95 ↓
UA-MT (MICCAI'19)	82.46	72.85	2.68	11.67	83.12	73.01	2.30	9.87
CPS (CVPR'21)	83.38	73.82	2.57	13.99	84.01	74.02	2.18	12.16
DTC (AAAI'21)	83.77	74.34	2.73	15.64	84.43	73.56	2.34	14.77
ICT (NN'22)	80.62	71.49	2.86	12.40	81.76	72.01	2.82	9.66
CPCL (JBHI'22)	82.59	72.97	7.06	18.33	83.48	74.08	2.08	9.53
URPC (MedIA'22)	82.03	71.03	4.87	17.36	82.93	72.57	4.19	15.93
BCP (CVPR'23)	83.15	73.14	3.57	12.81	84.17	74.37	3.24	11.69
CauSSL (ICCV'23)	80.57	69.88	3.08	13.15	81.09	71.01	3.76	11.90
UniMatch (CVPR'23)	84.89	76.12	2.62	9.55	86.09	76.80	2.47	8.51
UnCo (TMI'25)	85.09	75.64	**1.89**	**8.63**	85.16	75.04	**1.74**	8.41
CMFN (JBHI'26)	85.52	**76.82**	2.59	8.90	86.56	**77.77**	2.48	**7.72**
Ours	**86.07**	76.68	3.07	11.33	**86.60**	77.53	2.45	7.83

Table 3: Quantitative comparison on the Pancreas-NIH dataset.

Method	Pancreas-NIH (10% / 6 labeled data)				Pancreas-NIH (20% / 12 labeled data)			
	Dice ↑	Jaccard ↑	ASD ↓	HD95 ↓	Dice ↑	Jaccard ↑	ASD ↓	HD95 ↓
UA-MT (MICCAI'19)	66.44	52.02	3.03	17.04	76.10	62.62	2.43	10.87
DTC (AAAI'21)	66.58	51.79	4.16	15.46	76.27	62.82	2.20	8.70
MCF (CVPR'23)	67.71	53.83	2.34	17.17	75.00	61.27	3.27	11.59
CML (MM'24)	68.20	54.00	2.80	14.80	77.26	56.21	1.52	25.82
BCP (CVPR'23)	75.17	60.99	2.32	10.49	82.91	70.97	2.25	6.43
EIC (TMI'24)	–	–	–	–	81.17	68.68	1.46	6.17
MLRP (MIA'24)	75.93	62.12	1.54	9.07	81.53	69.35	1.33	6.81
SKCDF (CVPR'25)	70.71	56.48	1.37	19.84	78.30	65.27	1.38	9.34
UnCo (TMI'25)	78.53	65.20	**1.33**	7.36	81.93	69.77	1.36	5.34
Ours	**82.92**	**71.11**	1.53	**5.52**	**84.22**	**73.00**	**1.30**	**4.91**

centage points on Pancreas-NIH under the 10% and 20% settings, respectively. On LA, our method also achieves the best ASD under both settings and the best results across all four metrics with 20% labeled data. On Pancreas-NIH, it achieves the best Dice, Jaccard, and HD95 under both labeling settings, together with the best ASD at 20% labeled data. On BraTS2019, our method achieves the best Dice and the second-best Jaccard under both settings, while achieving the second-best HD95 with 20% labeled data. Overall, these results show that the proposed framework consistently improves region-overlap accuracy across different anatomical structures, imaging modalities, and annotation levels.

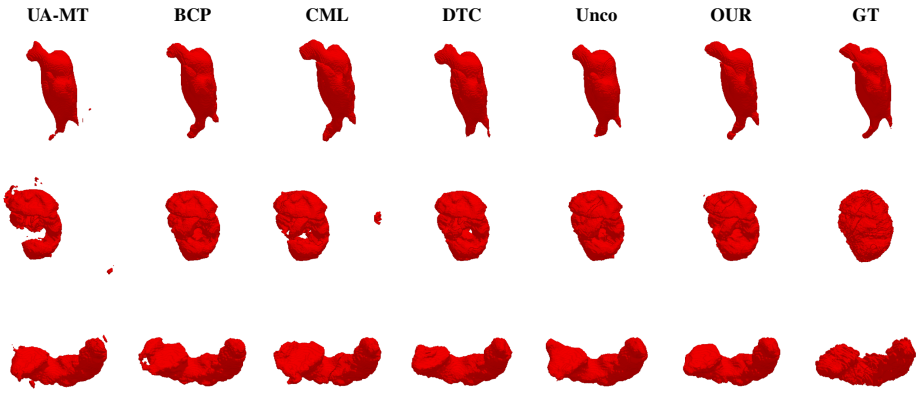


Figure 2: Qualitative comparison of different semi-supervised segmentation methods on LA, BraTS2019, and Pancreas-CT datasets. From top to bottom, the three rows correspond to LA, BraTS2019, and Pancreas-CT, respectively.

Qualitative results. As shown in Fig. 2, the proposed method produces more coherent segmentation results with fewer isolated false-positive regions and better preservation of the overall target morphology. Across the three datasets, the predicted shapes are generally more consistent with the corresponding ground truth than those of the competing methods.

4.3 Ablation Studies

We conduct ablation studies under the 10% labeled-data setting to evaluate the contributions of CSDR and TPDP and their internal components. Dice and Jaccard are used as evaluation metrics. The baseline denotes the same training framework without CSDR and TPDP, while the remaining training settings are kept unchanged. The best result in each column is highlighted in bold.

Overall contribution of CSDR and TPDP. As shown in Table 4, introducing either CSDR or TPDP consistently improves the baseline across all three datasets. Their combination yields further gains and achieves the best performance in every setting, indicating that state-adaptive supervision and disagreement probing provide complementary benefits. Compared with the baseline, the full model improves Dice by 2.05, 1.76, and 5.06 percentage points on LA, BraTS2019, and Pancreas-NIH, respectively.

Table 4: Ablation of CSDR and TPDP on LA, BraTS2019, and Pancreas-NIH under the 10% labeled-data setting.

Method	CSDR	TPDP	LA (10% / 8)		BraTS2019 (10% / 25)		Pancreas-NIH (10% / 6)	
			Dice ↑	Jaccard ↑	Dice ↑	Jaccard ↑	Dice ↑	Jaccard ↑
Baseline	–	–	88.76	80.21	84.31	74.96	77.86	64.12
+ CSDR	✓	–	89.84	81.87	85.18	76.08	80.21	67.02
+ TPDP	–	✓	89.57	81.46	85.02	75.84	80.47	67.34
Ours	✓	✓	**90.81**	**83.25**	**86.07**	**76.68**	**82.92**	**71.11**

Internal analysis of CSDR and TPDP. We further examine the internal components of the two mechanisms on LA and BraTS2019. For CSDR, Consensus Sharpening (CS) corresponds to sharpening uncertain agreement, while Disagreement Reconciliation (DR) handles informative disagreement. As shown in Table 5, both CS and DR individually improve the baseline, and combining them yields further gains on both datasets. This supports the use of differentiated supervision for the two predictive states. For TPDP, Teacher Perturbation (TP) probes disagreement under variations in the teacher probability distribution, while Disagreement Weighting (DW) uses the resulting disagreement information to modulate teacher-guided supervision. Table 6 shows that both components improve the baseline individually, while their combination consistently achieves the best performance, supporting the complementary roles of perturbation-based probing and response-guided weighting.

Table 5: Ablation of CSDR on LA and BraTS2019.

Method	CS	DR	LA (10% / 8)		BraTS2019 (10% / 25)	
			Dice ↑	Jaccard ↑	Dice ↑	Jaccard ↑
Baseline	–	–	88.76	80.21	84.31	74.96
+ CS	✓	–	89.31	81.42	84.79	75.57
+ DR	–	✓	89.46	81.58	84.92	75.73
+ CSDR	✓	✓	**89.84**	**81.87**	**85.18**	**76.08**

Table 6: Ablation of TPDP on LA and BraTS2019.

Method	TP	DW	LA (10% / 8)		BraTS2019 (10% / 25)	
			Dice ↑	Jaccard ↑	Dice ↑	Jaccard ↑
Baseline	–	–	88.76	80.21	84.31	74.96
+ TP	✓	–	89.31	81.17	84.72	75.43
+ DW	–	✓	89.46	81.34	84.88	75.58
+ TPDP	✓	✓	**89.57**	**81.46**	**85.02**	**75.84**

5 Conclusion

In this work, we presented a reflective teacher–student framework for semi-supervised 3D medical image segmentation that provides differentiated supervision for uncertain agreement and informative disagreement. CSDR sharpens tentative consensus and progressively reconciles disagreement, while TPDP probes disagreement responses to teacher-probability perturbations to modulate teacher-guided learning. Experiments on three public datasets demonstrate consistent improvements under limited-label settings, supporting the effectiveness of the proposed framework.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with writing refinement, literature retrieval and discovery, and drafting portions of the manuscript. We have not used generative AI tools for conducting research experiments, generating synthetic datasets, proving mathematical claims, or other research-related purposes not specified above.

Specifically, generative AI tools were used to improve the clarity and readability of the manuscript, assist in organizing and refining certain sections of the paper, and support the discovery of relevant literature. All AI-assisted outputs were carefully reviewed, verified, and revised by the authors. Literature suggestions provided by AI tools were manually checked against the original publications, and AI-assisted writing was evaluated and edited to ensure accuracy, originality, and consistency with the authors' intended contributions. The authors take full responsibility for the final content of this work, including all text, claims, experimental results, and conclusions.

REFERENCES

Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10076–10095, 2024.

Yunhao Bai, Duowen Chen, Qingli Li, Wei Shen, and Yan Wang. Bidirectional copy-paste for semi-supervised medical image segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11514–11524, 2023.

Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. *Medical image analysis*, 87:102792, 2023.

Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2613–2622. IEEE, 2021.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, volume 9901 of *Lecture Notes in Computer Science*, pp. 424–432. Springer, 2016. doi: 10.1007/978-3-319-46723-8_49.

Guiyuhan Deng, Hao Sun, and Wei Xie. Correlation-based switching mean teacher for semi-supervised medical image segmentation. *Neurocomputing*, 633:129818, 2025.

Rongjun Ge, Chong Wang, Yuxin Liu, Chunqiang Lu, Cong Xia, Yehui Jiang, Fangyi Xu, Yinsu Zhu, Daoqiang Zhang, Chengyu Liu, et al. Human gaze-based dual teacher guidance learning for semi-supervised medical image segmentation. *Neural Networks*, 192:107865, 2025.

Along He, Tao Li, Juncheng Yan, Kai Wang, and Huazhu Fu. Bilateral supervision network for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 43(5): 1715–1726, 2023.

Kaiwen Huang, Tao Zhou, Huazhu Fu, Yizhe Zhang, Yi Zhou, and Xiao-Jun Wu. Uncertainty-aware cross-training for semi-supervised medical image segmentation. *IEEE Transactions on Image Processing*, 34:5543–5556, 2025.

Wei Huang, Lei Zhang, Zizhou Wang, and Lituan Wang. Exploring inherent consistency for semi-supervised anatomical structure segmentation in medical imaging. *IEEE Transactions on Medical Imaging*, 43(11):3731–3741, 2024.

Rushi Jiao, Yichi Zhang, Le Ding, Bingsen Xue, Jicong Zhang, Rong Cai, and Cheng Jin. Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. *Computers in Biology and Medicine*, 169:107840, 2024.

Wei Li, Ruifeng Bian, Wenyi Zhao, Weijin Xu, and Huihua Yang. Diversity matters: Cross-head mutual mean-teaching for semi-supervised medical image segmentation. *Medical Image Analysis*, 97:103302, 2024.

Zihang Liu, Haoran Zhang, and Chunhui Zhao. Prototype-oriented contrastive learning for semi-supervised medical image segmentation. *Biomedical Signal Processing and Control*, 88:105571, 2024.

Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 8801–8809, 2021.

Xiangde Luo, Guotai Wang, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Dimitris N Metaxas, and Shaoting Zhang. Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *Medical Image Analysis*, 80:102517, 2022.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34 (10):1993–2024, 2014.

Juzheng Miao, Cheng Chen, Furui Liu, Hao Wei, and Pheng-Ann Heng. Caussl: Causality-inspired semi-supervised learning for medical image segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21369–21380. IEEE, 2023.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571. IEEE, 2016. doi: 10.1109/3DV.2016.79.

S Niyas, SJ Pawan, M Anand Kumar, and Jeny Rajan. Medical image segmentation with 3d convolutional neural networks: A survey. *Neurocomputing*, 493:397–413, 2022.

Zhen Peng, Shengwei Tian, Long Yu, Dezhi Zhang, Weidong Wu, and Shaofeng Zhou. Semi-supervised medical image classification with adaptive threshold pseudo-labeling and unreliable sample contrastive loss. *Biomedical Signal Processing and Control*, 79:104142, 2023.

Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 556–564. Springer, 2015.

Yinghuan Shi, Jian Zhang, Tong Ling, Jiwen Lu, Yefeng Zheng, Qian Yu, Lei Qi, and Yang Gao. Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation. *IEEE transactions on medical imaging*, 41(3):608–620, 2021.

Jiawei Su, Zhiming Luo, Sheng Lian, Dazhen Lin, and Shaozi Li. Mutual learning with reliable pseudo label for semi-supervised medical image segmentation. *Medical Image Analysis*, 94: 103111, 2024.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Vikas Verma, Kenji Kawaguchi, Alex Lamb, Juho Kannala, Arno Solin, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. *Neural Networks*, 145:90–106, 2022.

Liansheng Wang, Jiacheng Wang, Lei Zhu, Huazhu Fu, Ping Li, Gary Cheng, Zhipeng Feng, Shuo Li, and Pheng-Ann Heng. Dual multiscale mean teacher network for semi-supervised infection segmentation in chest ct volume for covid-19. *IEEE Transactions on Cybernetics*, 53(10):6363–6375, 2022.

Yongchao Wang, Bin Xiao, Xiuli Bi, Weisheng Li, and Xinbo Gao. Mcf: Mutual correction framework for semi-supervised medical image segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15651–15660. IEEE, 2023.

Fuping Wu and Xiahai Zhuang. Minimizing estimated risks on unlabeled data: A new formulation for semi-supervised medical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6021–6036, 2022.

Han Wu, Chong Wang, and Zhiming Cui. Dual cross-image semantic consistency with self-aware pseudo labeling for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 45(1):177–189, 2026.

Song Wu, Xiaoyu Wei, Xinyue Chen, Yazhou Ren, Jing He, and Xiaorong Pu. Cross-view mutual learning for semi-supervised medical image segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9253–9261, 2024.

Yicheng Wu, Zhonghua Wu, Qianyi Wu, Zongyuan Ge, and Jianfei Cai. Exploring smoothness and class-separation for semi-supervised medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 34–43. Springer, 2022.

Zhaohan Xiong, Qing Xia, Zhiqiang Hu, Ning Huang, Cheng Bian, Yefeng Zheng, Sulaiman Vesal, Nishant Ravikumar, Andreas Maier, Xin Yang, et al. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical image analysis*, 67:101832, 2021.

Zhe Xu, Yixin Wang, Donghuan Lu, Lequan Yu, Jiangpeng Yan, Jie Luo, Kai Ma, Yefeng Zheng, and Raymond Kai-yu Tong. 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.

Zhe Xu, Yixin Wang, Donghuan Lu, Xiangde Luo, Jiangpeng Yan, Yefeng Zheng, and Raymond Kai-yu Tong. Ambiguity-selective consistency regularization for mean-teacher semi-supervised medical image segmentation. *Medical Image Analysis*, 88:102880, 2023.

Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7236–7246. IEEE, 2023.

Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. 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.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6023–6032, 2019.

Qingjie Zeng, Zilin Lu, Yutong Xie, and Yong Xia. Pick: Predict and mask for semi-supervised medical image segmentation. *International Journal of Computer Vision*, 133(6):3296–3311, 2025a.

Qingjie Zeng, Yutong Xie, Zilin Lu, Mengkang Lu, Jingfeng Zhang, and Yong Xia. Consistency-guided differential decoding for enhancing semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(1):44–56, 2025b.

Xiang Zeng, Shengwu Xiong, Jinming Xu, Guangxing Du, and Yi Rong. Uncertainty co-estimator for improving semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(9):3870–3881, 2025c.

Huaikun Zhang, Pei Ma, Jizhao Liu, Jing Lian, and Yide Ma. Prototype-augmented mean teacher for robust semi-supervised medical image segmentation. *Pattern Recognition*, 166:111722, 2025a.

Shuo Zhang, Jiaojiao Zhang, Biao Tian, Thomas Lukasiewicz, and Zhenghua Xu. Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 83:102656, 2023a.

Yichi Zhang, Rushi Jiao, Qingcheng Liao, Dongyang Li, and Jicong Zhang. Uncertainty-guided mutual consistency learning for semi-supervised medical image segmentation. *Artificial Intelligence in Medicine*, 138:102476, 2023b.

Yichi Zhang, Bohao Lv, Le Xue, Wenbo Zhang, Yuchen Liu, Yu Fu, Yuan Cheng, and Yuan Qi. Semisam+: Rethinking semi-supervised medical image segmentation in the era of foundation models. *Medical image analysis*, 106:103733, 2025b.

Zheng Zhang, Guanchun Yin, Bo Zhang, Wu Liu, Xiuzhuang Zhou, and Wendong Wang. A semantic knowledge complementarity based decoupling framework for semi-supervised class-imbalanced medical image segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25940–25949. IEEE, 2025c.

Zhen Zhao, Zicheng Wang, Longyue Wang, Dian Yu, Yixuan Yuan, and Luping Zhou. Alternate diverse teaching for semi-supervised medical image segmentation. In *Computer Vision – ECCV 2024*, volume 15063 of *Lecture Notes in Computer Science*, pp. 227–243. Springer, 2025.

Xinghu Zhou, Guanghan Wang, Yuanzhi Cheng, Zixuan Wang, Xin Wang, Guohua Wang, and Shinichi Tamura. Joint learning of confidence fusion, semantic alignment and group-guided reliability: A novel semi-supervised learning framework for 3d medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 30(3):2487–2499, 2026.

A Appendix

A.1 More Ablation study

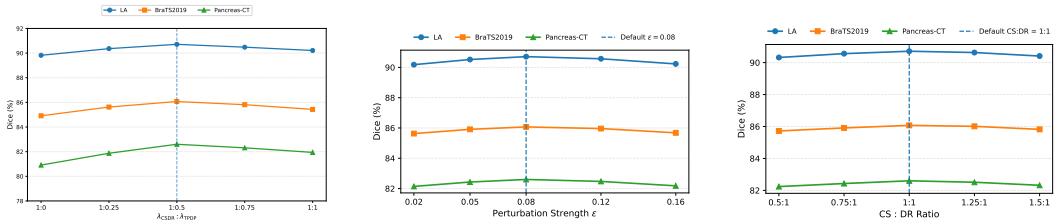
Loss Weight Sensitivity Analysis: To further investigate the robustness of the proposed framework, we conduct additional ablation studies on the relative contributions of the CSDR and TPDP mechanisms. Specifically, we fix $\lambda_{\mathrm{CSDR}} = 1$ and vary λ_{TPDP} under the 10% labeled-data setting. Experiments are conducted on the LA, BraTS2019, and Pancreas-CT datasets to evaluate whether the performance is sensitive to the balance between the two proposed components. The corresponding results are presented in Fig. 3a.

Perturbation Strength Sensitivity Analysis: To further investigate the sensitivity of the TPDP mechanism to the perturbation strength, we conduct an additional hyperparameter analysis under the 10% labeled-data setting. Specifically, the perturbation magnitude ϵ is varied among $\{0.02, 0.05, 0.08, 0.12, 0.16\}$, while keeping all other experimental configurations unchanged. Experiments are conducted on the LA, BraTS2019, and Pancreas-CT datasets, and the results are reported in Fig. 3b.

CS-DR Ratio Sensitivity Analysis: To further analyze the effect of the two state-adaptive supervision strategies in CSDR, we conduct an additional ablation study on the relative balance between Consensus Sharpening (CS) and Disagreement Reconciliation (DR). Specifically, the CS-to-DR ratio is varied under the 10% labeled-data setting, while the remaining experimental settings are kept unchanged. The results are shown in Fig. 3c.

A.2 More Qualitative Results

To further examine the segmentation quality across different cases, we provide additional three-dimensional qualitative results on the LA and Pancreas-CT datasets in Figs. 4 and 5. We do not include additional three-dimensional visualizations for BraTS2019 because the current experimental results were obtained without post-processing, which may introduce fragmented or isolated regions in the predictions and consequently affect the readability of the surface-based visualization. Therefore, the additional visualizations are focused on the LA and Pancreas-CT datasets, where the three-dimensional structures can be presented more clearly.



(a) CSDR–TPDP weight sensitivity analysis.

(b) TPDP perturbation magnitude sensitivity analysis.

(c) CS–DR balance sensitivity analysis.

Figure 3: Sensitivity analyses under the 10% labeled-data setting: (a) relative weights of CSDR and TPDP, (b) perturbation magnitude of TPDP, and (c) CS–DR balance.

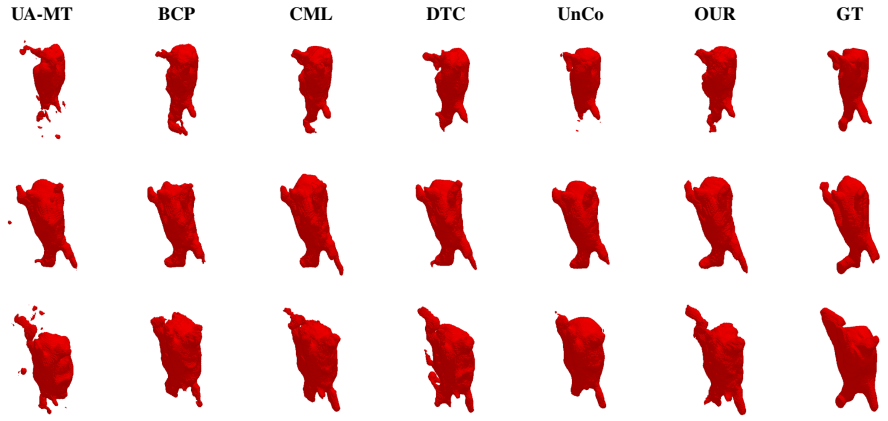


Figure 4: Additional qualitative comparison on the LA dataset. From left to right, the columns correspond to UA-MT, BCP, CML, DTC, UnCo, our method, and GT, respectively. Three additional cases are presented to provide a broader qualitative assessment of the segmentation results.

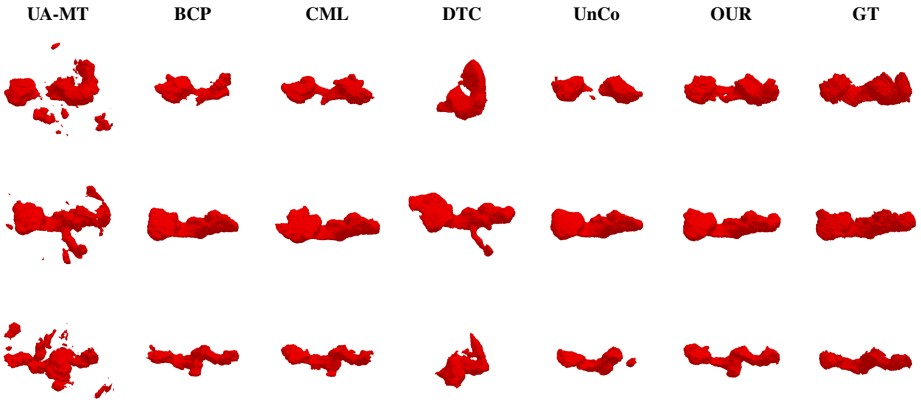


Figure 5: Additional qualitative comparison on the Pancreas-CT dataset. Three-dimensional visualization results of different semi-supervised segmentation methods are presented for three additional cases. From left to right, the columns correspond to UA-MT, BCP, CML, DTC, UnCo, our method, and GT, respectively.

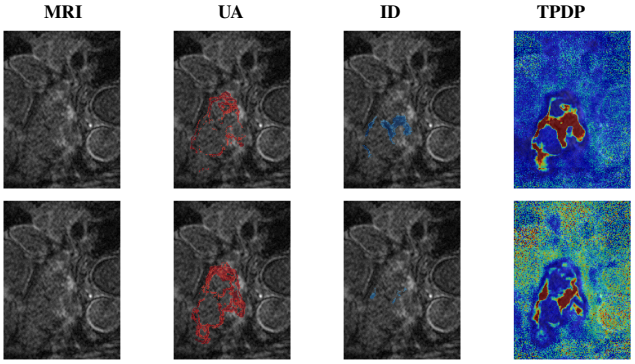


Figure 6: Stage-wise visualization of the proposed supervision mechanisms during teacher–student training. The first row shows the early training stage, while the second row shows the late training stage. From left to right, the columns correspond to the MRI input, uncertain agreement regions (UA), informative disagreement regions (ID), and the TPDP response, respectively.

A.3 Stage-wise Analysis of the Proposed Losses

To further analyze the behavior of CSDR and TPDP during training, we visualize the uncertain agreement, informative disagreement, and TPDP responses at the early and late stages in Fig. 6. In the early stage, broader UAR and IDR regions are observed, indicating unstable teacher–student predictions and substantial uncertainty in unlabeled regions. As training progresses, these regions gradually become concentrated around challenging anatomical structures and ambiguous boundaries. CSDR progressively sharpens uncertain consensus regions by reducing residual ambiguity, while TPDP explores the response of disagreement under teacher perturbations and emphasizes informative discrepancies. These observations demonstrate the complementary roles of the two mechanisms in adapting supervision throughout teacher–student training.