CHCL-Seg: A Competitive Hierarchical Contrastive Learning Framework for Semi-Supervised 3D Medical Image Segmentation

Anonymous authors
Paper under double-blind review

Abstract

3D medical image segmentation plays an important role in intelligent clinical applications. Although existing deep learning methods have achieved substantial progress in this field, they are limited by the high cost of 3D manual annotation. Semi-supervised learning, which combines a small amount of labeled data with a large amount of unlabeled data for training, is an effective approach to alleviate label scarcity. However, existing semi-supervised methods for 3D medical image segmentation still face the following problems: confirmation bias is difficult to effectively mitigate, target boundary segmentation accuracy is insufficient, and cross-dataset semantic representation is difficult. To address these challenges, this paper proposes a competitive hierarchical contrastive learning framework for semi-supervised 3D medical image segmentation, termed CHCL-Seg. We first construct a dual-branch network with shared parameters, and introduce an uncertainty-aware competitive loss (UnCL) to balance noise filtering and boundary information utilization, thereby mitigating confirmation bias. We then propose a hierarchical contrastive learning mechanism, where at the high level, the Dense-OT module performs class prototype alignment to alleviate class imbalance, and at the low level, the V-DACL module mines hard boundary samples to enhance boundary feature discrimination. Extensive experiments on three public datasets, LA, BraTS-2019, and ACDC demonstrate that the proposed method comprehensively outperforms existing state-of-the-art semi-supervised segmentation methods. Our code will be released upon acceptance.

1 Introduction

3D medical image segmentation is the basis of intelligent clinical applications such as tumor radiotherapy target delineation and minimally invasive surgical navigation Isensee et al. (2024). Although deep learning methods have achieved substantial progress in this field, they usually rely on large-scale, high-quality voxel-level annotations, which limits their applicability in real clinical scenarios where annotation cost is high and expert resources are limited Cheplygina et al. (2019); Li et al. (2023). Semi-supervised learning leverages a small amount of labeled data and a large amount of unlabeled data for training, and is an effective approach to alleviate label scarcity Su et al. (2024); Zeng et al. (2025).

The teacher–student framework is a widely adopted paradigm in semi-supervised 3D medical image segmentation Yu et al. (2019). Its core idea is to use a teacher model to generate pseudo-labels for unlabeled data and guide the student model through consistency regularization. However, as shown in Fig. 1, the performance of this paradigm heavily depends on the quality of pseudo-labels. When the teacher model produces incorrect predictions, erroneous supervision signals may be continuously learned by the student model and further reinforced through the exponential moving average (EMA) update of the teacher model, leading to confirmation bias. To mitigate this issue, existing methods typically improve pseudo-label quality through confidence filtering Han et al. (2025), dual-network competition Hu et al. (2025), or mask-based filtering Cheng et al. (2024). Although these methods

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.

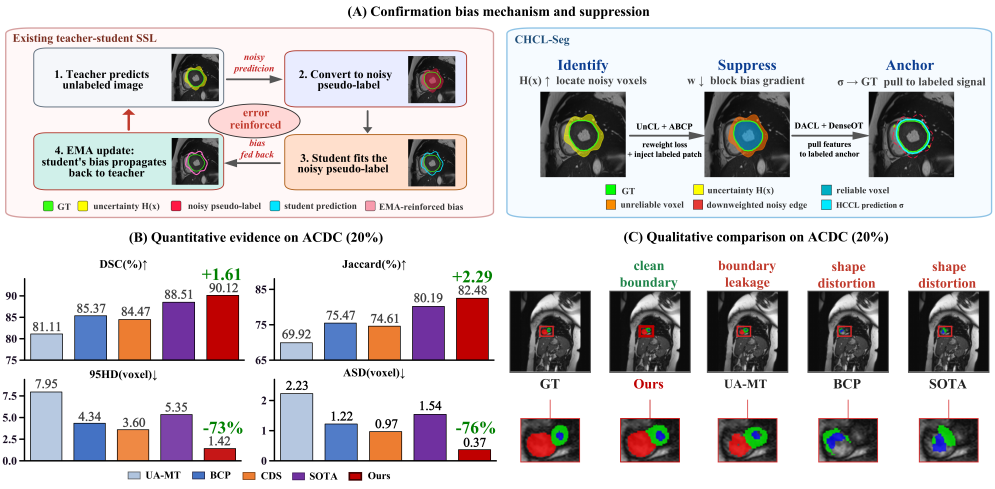

Figure 1: (A) Left: existing teacher–student semi-supervised segmentation methods. Right: the proposed CHCL-Seg. (B) Quantitative evaluation on the ACDC dataset with 20% labeled data, compared with the strongest baseline (MC-Net): Dice increased by 1.61% and Jaccard increased by 2.29%, while HD95 decreased by 73.5% and ASD decreased by 76%. (C) Visualization on the ACDC dataset with 20% labeled data: the segmentation results of CHCL-Seg closely match the ground truth.

can reduce the impact of low-confidence pseudo-labels to some extent, they mainly operate at the prediction level and cannot fully handle the complex uncertainty in semi-supervised 3D medical image segmentation Azad et al. (2024). On the one hand, boundary regions inherently exhibit high prediction uncertainty, and simple hard filtering strategies tend to misclassify these informative hard voxels as noise and discard them, weakening the model's ability to learn boundary regions Li et al. (2024); Peiris et al. (2023). On the other hand, there may exist differences in imaging conditions, anatomical structures, and lesion morphology between labeled and unlabeled data, and relying solely on prediction-level pseudo-label optimization makes it difficult to obtain stable cross-dataset semantic representations Huang et al. (2025b); Qiu et al. (2023).

To address the above issues, this paper proposes a competitive hierarchical contrastive learning framework for semi-supervised 3D medical image segmentation, termed CHCL-Seg. Specifically, we first design a dual-branch network with shared parameters. Based on an uncertainty-aware loss (UnCL), a competitive mechanism dynamically scales the supervision weight of each voxel, balancing noise filtering and boundary information utilization to mitigate confirmation bias. On this basis, we further propose a hierarchical contrastive learning mechanism: at the high level, a dense optimal transport-based prototype alignment module (Dense-OT) is designed to alleviate learning bias caused by class imbalance and provide global anchors across labeled and unlabeled data; at the low level, a voxel-wise density-aware contrastive learning module (V-DACL) is introduced, which mines boundary-confusing hard samples based on local density in the feature space and focuses optimization gradients on boundary and artifact regions, achieving intra-class compactness and inter-class separability. The main contributions of this paper are as follows:

- We propose a competitive hierarchical contrastive learning framework for semi-supervised 3D medical image segmentation, termed CHCL-Seg, to mitigate confirmation bias, data distribution discrepancy, and insufficient boundary segmentation accuracy.
- We propose an uncertainty-aware competitive loss (UnCL), which dynamically adjusts supervision weights based on voxel-wise entropy, adapts to boundary prediction uncertainty in 3D medical images, and balances noise filtering and boundary information utilization.
- We propose a hierarchical contrastive learning mechanism, where the high level Dense-OT module aligns feature distributions across data, and the low level V-

DACL module accurately mines hard boundary samples, improving model generalization and segmentation accuracy.

- We conduct extensive experiments on three public 3D medical image segmentation datasets, LA, BraTS-2019, and ACDC, and the results show that the proposed method consistently outperforms existing state-of-the-art methods.

2 Related Work

2.1 3D Medical Image Semi-supervised Segmentation

Pseudo-label generation and quality optimization are the main approaches to alleviate label scarcity in semi-supervised 3D medical image segmentation. Early studies adopt consistency regularization–based teacher–student frameworks, where a teacher model updated by exponential moving average generates pseudo-labels for unlabeled data Yu et al. (2019); Bar et al. (2019); Huang et al. (2021). Subsequent works improve pseudo-label stability through uncertainty-aware weighting Huang et al. (2026); Dong et al. (2025); Oh et al. (2024); Assefa et al. (2025), 3D medical-specific augmentation Bai et al. (2023); Liu et al. (2025); Li et al. (2025a), and multi-scale consistency constraints Tian et al. (2022); Guo et al. (2025). However, feature coupling in the teacher–student framework leads to the accumulation of erroneous pseudo-labels during training. To address this issue, some studies introduce competitive mechanisms for pseudo-label optimization Hu et al. (2025), selecting reliable samples through multi-branch predictions and forming mainstream paradigms such as cross-supervision Zhou et al. (2024a); Kumari & Singh (2025) and multi-task collaboration Zhou et al. (2024b). Nevertheless, these methods mostly rely on hard-threshold filtering, which cannot adapt to the sparse foreground Yang et al. (2025) and high boundary uncertainty in 3D medical images Liu et al. (2024b), making it difficult to balance noise suppression and boundary information preservation and thus exacerbating confirmation bias in semi-supervised training. To this end, we construct a dual-branch competitive mechanism with shared parameters and propose an uncertainty-aware competitive loss (UnCL), which addresses this issue through voxel-wise dynamic weight assignment.

2.2 Contrastive Learning for 3D Medical Image Segmentation

In semi-supervised 3D medical image segmentation, distribution shift between labeled and unlabeled data and feature confusion between foreground and background voxels are the main factors leading to insufficient model generalization and suboptimal boundary segmentation accuracy Adiga et al. (2024); Luo et al. (2021). Contrastive learning is an effective approach to improve feature representation, and existing studies can be broadly divided into two categories based on their optimization objectives: one focuses on global feature distribution alignment, which reduces cross-data distribution discrepancies through self-supervised pretraining Zhao et al. (2025); Miao et al. (2023); Wang & Ma (2023), prototype matching Ma et al. (2024); Yuan et al. (2024), and optimal transport Qi et al. (2024); the other focuses on voxel-level discriminative feature learning You et al. (2022); Xu et al. (2025); Goncharov et al. (2023); Gao et al. (2026); You et al. (2023), which enhances intra-class compactness and inter-class separability via voxel-wise contrastive loss and emphasizes hard sample mining. However, these methods have notable limitations: global alignment may disrupt the spatial continuity of 3D medical images, while voxel-level contrastive learning lacks precise boundary delineation, and the two are not jointly optimized. Therefore, this paper designs a hierarchical contrastive learning mechanism that combines high level global distribution alignment with low level boundary hard sample mining, achieving improved segmentation performance.

3 Method

3.1 Problem Definition

Given a 3D medical image segmentation dataset $\mathcal{D}$, it consists of n labeled samples $\mathcal{D}^l = \{(X_i^l, Y_i^l)\}_{i=1}^n$ and m unlabeled samples $\mathcal{D}^u = \{X_j^u\}_{j=1}^m$, where $n \ll m$. Here, $X \in \mathbb{R}^{D\times H\times W}$

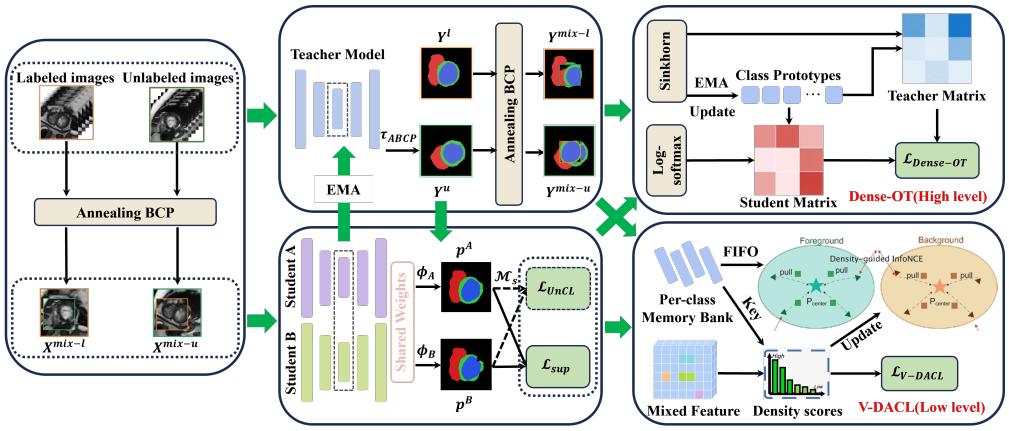

Figure 2: Overview of the proposed semi-supervised 3D medical image segmentation framework CHCL-Seg: labeled and unlabeled data pairs are bi-directionally mixed via ABCP, and pseudo-labels for unlabeled data are generated by the teacher network and combined with ground-truth annotations to supervise the student model. The mixed data pairs are fed into a dual-branch student network, which is trained with the UnCL loss, while the teacher network is updated via EMA. At the feature level, we further design a hierarchical contrastive learning scheme, which improves segmentation accuracy by combining high level class prototype alignment and low level boundary feature aggregation.

denotes a 3D volume, and $Y \in \{0, 1, \ldots, K-1\}^{D \times H \times W}$ is the voxel-wise annotation with K classes. The goal is to build a competition-based hierarchical contrastive learning framework for semi-supervised 3D medical image segmentation, leveraging limited labeled data and abundant unlabeled data to improve segmentation performance. The overall framework is shown in Fig. 2.

To mitigate confirmation bias, we design a parameter-shared competitive dual-student model. The student network f_θ produces predictions using two independent Dropout operations ϕ_A and ϕ_B:

$$p^k(X) = f_\theta(X; \phi_k), \quad k \in \{A, B\}. \tag{1}$$

The teacher network f_ξ is updated by the EMA of the student parameters. Dropout is disabled during inference, and low-confidence voxels are filtered using a confidence threshold $\tau_{ABCP} = 0.8$. The pseudo-labels are defined as: $Y^u = \text{argmax}\,(f_\xi(X^u; \tau_{ABCP}))$.

3.2 Annealing Bidirectional Copy-Paste (ABCP)

For input images, we adopt a linearly annealing bidirectional copy-paste ($ABCP$) augmentation to reduce the distribution gap between labeled and unlabeled data. Given a random binary mask $M \in \{0, 1\}^{D \times H \times W}$, the bidirectional mixing is defined as:

$$ABCP(A, B, M) = A \odot M + B \odot (1 - M). \tag{2}$$

To reduce boundary noise, we use the following linear annealing schedule, where t is the iteration:

$$\mathcal{P}_{ABCP}(t) = \begin{cases} 1.0 & t < t_{start}, \\ 1 - \frac{t - t_{start}}{t_{end} - t_{start}} & t_{start} \le t < t_{end}, \\ 0.0 & t \ge t_{end}. \end{cases} \tag{3}$$

For the student network f_θ, we construct mixed inputs as:

$$X^{mix-l} = ABCP(X^l, X^u, M), X^{mix-u} = ABCP(X^u, X^l, M). \tag{4}$$

We also construct mixed labels using pseudo-labels Y^u and ground-truth labels Y^l:

$$Y^{mix-l} = ABCP(Y^l, Y^u, M), Y^{mix-u} = ABCP(Y^u, Y^l, M). \tag{5}$$

Based on the mask M, different loss weights are assigned to different regions. The mixed loss for each student branch k is defined as:

$$\mathcal{L}^k_{mix}(X, Y^l, Y^u, M) = w_l \cdot \mathcal{L}_{seg}(p^k(X), Y^l; M) + w_u \cdot \mathcal{L}_{seg}(p^k(X), Y^u; 1 - M), \quad (6)$$

where the weight for labeled regions is $w_l = 1.0$, and for pseudo-labeled regions is $w_u = 0.5$. The segmentation loss $\mathcal{L}_{seg}$ is the average of cross-entropy and Dice loss.

The final supervised loss is:

$$\mathcal{L}_{sup} = \sum_{k\in\{A,B\}} \left[\mathcal{L}^k_{mix}(X^{mix-l}, Y^l, Y^u, M) + \mathcal{L}^k_{mix}(X^{mix-u}, Y^u, Y^l, M)\right]. \quad (7)$$

3.3 Uncertainty-aware Competitive Loss (UnCL)

To improve the student network, we compare the voxel-wise prediction confidence of the two student branches and construct a competition mask $\mathcal{M}_s$. The more reliable branch is selected to form the final student prediction P^s, ensuring that each voxel learns pseudo-labels only from the higher-confidence branch:

$$\mathcal{M}_s(A, B) = \mathbb{1}\left(\max_k p^i_k(A) > \max_k p^i_k(B)\right), P^s = P^A \odot \mathcal{M}_s(A, B) + P^B \odot \mathcal{M}_s(B, A). \quad (8)$$

Beyond selecting the learning branch, it is also important to control the learning strength. Voxels near boundaries usually have lower confidence and are more likely to be ignored. To address this, we propose an uncertainty-aware competitive loss (UnCL), which quantifies uncertainty using normalized voxel-wise prediction entropy: $H^{(i)} = -\frac{1}{\ln K}\sum_k p^i_k \ln\left(p^i_k + \epsilon\right)$, where $\epsilon = 10^{-6}$ and K is the number of classes.

The final UnCL loss is defined as:

$$\mathcal{L}_{UnCL} = \sum_{i\in\mathcal{M}_s} \frac{CE\left(P^s_{(i)}, Y^u_{(i)}\right)}{\exp\left(\beta H^{(i)}_s\right) + \exp\left(\beta H^{(i)}_t\right)} + \beta\left(H^{(i)}_s + H^{(i)}_t\right), \quad (9)$$

where $H^{(i)}_s$ and $H^{(i)}_t$ denote the normalized prediction entropy of the student and teacher at voxel i, respectively. The temperature β decays over training iterations.

3.4 Hierarchical Contrastive Learning (HCL)

To enhance intra-class compactness and inter-class separability in the feature space, we introduce a hierarchical contrastive learning scheme on top of the ABCP data-level alignment. It consists of two complementary modules: a high level global prototype alignment module (Dense-OT) and a low level boundary feature optimization module (V-DACL).

3.4.1 High level dense optimal transport-based prototype alignment (Dense-OT)

Dense-OT performs global class-balanced prototype alignment via Sinkhorn-Knopp optimal transport Cuturi (2013), alleviating class imbalance. The class prototypes $C = \{c_k\}_{k=1}^K$ are updated with EMA, using two sources: ground-truth aggregation from labeled data and high-confidence sample aggregation from unlabeled data.

The teacher matrix between teacher projected tokens z_t and prototypes is built with stop-gradient:

$$\mathcal{M}_{teacher}[n, k] = \text{Sinkhorn}(\text{CosSim}(z^{(n)}_t, c_k)/\tau_{dot}, \pi), \quad (10)$$

where the class prior $\pi \in \mathbb{R}^K$ is estimated from teacher predictions with truncation and EMA smoothing. The student matrix is computed as:

$$\mathcal{M}_{student}[n, k] = \log(\text{softmax}(\text{CosSim}(z^{(n)}_s, c_k)/\tau_{dot})), \quad (11)$$

where z_s denotes student projected tokens, τ_{dot} follows a linear warm-up schedule.

The Dense-OT loss aligns student features to the teacher optimal transport soft targets:

$$\mathcal{L}_{Dense-OT} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathcal{M}_{teacher}[n,k]\mathcal{M}_{student}[n,k]. \tag{12}$$

3.4.2 Low level voxel-wise density-aware contrastive learning (V-DACL)

V-DACL focuses on optimizing low-density boundary voxels. For each class, we sample an embedding set $\mathcal{E}_k$ and combine it with a memory bank $\mathcal{G}_k$ to form $\mathcal{P}_k$. The density is computed using multi-scale kNN:

$$\rho_j = \frac{1}{|\mathcal{K}|}\sum_{k_s\in\mathcal{K}}\frac{1}{k_s}\sum_{m\in\mathrm{TopK}_{k_s}(\mathcal{P}_k,e_j)}\mathrm{CosSim}(f_\theta(e_j),f_\theta(e_m)), \tag{13}$$

where $\mathcal{P}_k = \mathcal{E}_k \cup \mathcal{G}_k$, and $\mathcal{K}$ is the set of neighborhood sizes.

For each valid class $k \in \mathcal{K}_{valid}$, which satisfies both minimum sample size and memory capacity, we select N_q samples with the lowest density as anchors $\{q_n\}$. We further select $N_p^+/2$ highest-density samples from the current batch and $N_p^+/2$ highest-density samples from the cross-batch memory bank as positive samples $\{m_i\}$, and N_p^- samples from other classes as negative samples $\{k_j^-\}$.

The class center and effective prototype are computed as:

$$p_{center} = \mathrm{normalize}\left(\frac{1}{N_p^+}\sum_{i=1}^{N_p^+}m_i\right),\ p_{eff} = \mathrm{normalize}\left(\sum_{i=1}^{N_p^+}w_i\cdot m_i\right), \tag{14}$$

where the weight w_i enhances the contribution of highly similar intra-class samples:

$$w_i = \frac{\exp\left(m_i\cdot p_{center}^\top\right)}{\sum_{j=1}^{N_p^+}\exp\left(m_j\cdot p_{center}^\top\right)}. \tag{15}$$

The InfoNCE loss for class k is calculated as:

$$\mathcal{L}_{info}^{(k)} = -\frac{1}{N_q}\sum_{n=1}^{N_q}\log\frac{\exp\left(q_n\cdot p_{eff}/\tau_{dacl}\right)}{\exp\left(q_n\cdot p_{eff}/\tau_{dacl}\right)+\sum_{j=1}^{N_p^-}\exp\left(q_n\cdot k_j^-/\tau_{dacl}\right)}, \tag{16}$$

where $\tau_{dacl} = 0.4$ is the temperature parameter.

The overall V-DACL loss is:

$$\mathcal{L}_{V-DACL} = \frac{1}{|\mathcal{K}_{valid}|}\sum_{k\in\mathcal{K}_{valid}}\mathcal{L}_{info}^{(k)}. \tag{17}$$

3.5 Overall Loss

The overall loss consists of four components: the supervised loss $\mathcal{L}_{sup}$, the uncertainty-aware competitive loss $\mathcal{L}_{UnCL}$, the Dense-OT global prototype alignment loss $\mathcal{L}_{Dense-OT}$, and the V-DACL voxel-wise density-aware contrastive loss $\mathcal{L}_{V-DACL}$:

$$\mathcal{L}_{overall} = \mathcal{L}_{sup} + \mathcal{L}_{UnCL} + \lambda_{dot}\cdot w_{dot}(t)\cdot\mathcal{L}_{Dense-OT} + \lambda_{dacl}\cdot w_{dacl}(t)\cdot\mathcal{L}_{V-DACL}, \tag{18}$$

where $\lambda_{dot} = 0.05$ and $\lambda_{dacl} = 0.03$ are weighting coefficients, and $w_{dot}(t)$ and $w_{dacl}(t)$ are warm-up scheduling functions. The detailed scheduling strategy is provided in Appendix A.1.

4 Experiments

To validate the effectiveness of the proposed method, we conducted experiments on three widely used semi-supervised medical image segmentation datasets: LA Chen et al. (2018), BraTS-2019 Bakas (2020) and ACDC Bernard et al. (2018). We conducted experiments at two annotation ratios of 10% and 20%. The backbone networks followed the CDS settings: 3D V-Net was used for LA and BraTS-2019, and 2D U-Net was used for ACDC. Detailed information regarding the datasets, training configurations, evaluation metrics, implementation details, and additional experimental results is provided in the Appendix B.

Table 1: Comparison with SOTA semi-supervised segmentation methods on the LA dataset. The best results are highlighted in bold.

	10% Scans				20% Scans			
Method	Dice↑	Jaccard↑	HD95↓	ASD↓	Dice↑	Jaccard↑	HD95↓	ASD↓
UA-MT (MICCAI 2019)	86.28	76.11	18.71	4.63	88.17	79.21	11.24	3.07
SASSNet (MICCAI 2020)	85.22	75.09	11.18	2.89	89.17	80.69	8.57	2.87
DTC (AAAI 2021)	81.89	71.23	15.81	3.80	89.43	81.01	7.76	1.90
BCP (CVPR 2023)	89.09	80.49	7.49	1.95	90.16	82.25	6.93	1.94
CDS (EAAI 2025)	87.21	77.56	11.02	2.97	90.33	82.46	6.16	1.91
UnCo (TMI 2025)	90.37	82.54	6.11	1.74	90.91	83.40	5.36	1.61
SBFC (ESWA 2026)	90.23	82.30	7.93	1.66	91.43	84.29	5.51	1.55
BCSI (AAAI 2026)	91.07	83.67	5.57	1.53	91.27	84.08	5.67	1.58
Ours	91.34	84.19	5.72	1.46	92.08	85.41	5.03	1.36

4.1 Comparison with State-of-the-Art Methods

LA Dataset Results. Table 1 compares UA-MT Yu et al. (2019), SASSNet Li et al. (2020), DTC Luo et al. (2021), BCP Bai et al. (2023), CDS Hu et al. (2025), UnCo Chen et al. (2025b), SBFC Jia et al. (2026) and BCSI Huang et al. (2026) on the left-atrium benchmark. At 20% labels, CHCL-Seg reaches 92.08 Dice and 85.41 Jaccard (+0.65 / +1.12 over SBFC), with HD95 / ASD down to 5.03 / 1.36. The 10% setting stays competitive. Additional baselines are in Appendix B.6.

Table 2: Comparison with SOTA semi-supervised segmentation methods on the BraTS-2019 dataset. The best results are highlighted in bold.

	10% Scans				20% Scans			
Method	Dice↑	Jaccard↑	HD95↓	ASD↓	Dice↑	Jaccard↑	HD95↓	ASD↓
DTC (AAAI 2021)	82.74	72.74	11.76	3.24	83.80	73.24	9.55	2.97
BCP (CVPR 2023)	84.35	75.01	10.95	2.60	86.47	77.41	7.88	1.44
AUSS (MedIA 2024)	81.99	71.92	11.47	2.44	83.21	73.55	11.39	2.03
CDS (EAAI 2025)	84.03	74.25	9.67	2.60	86.12	77.58	7.87	1.83
TDMVI (MICCAI 2025)	85.27	75.23	8.77	1.35	86.69	77.83	7.65	1.24
BCSI (AAAI 2026)	86.17	77.10	8.43	1.37	86.86	77.80	7.62	1.26
Ours	86.34	77.21	8.57	1.74	88.49	80.20	7.49	1.32

BraTS-2019 Dataset Results. Tumors here vary more in shape and have weaker boundaries. Table 2 compares DTC Luo et al. (2021), BCP Bai et al. (2023), AUSS Adiga et al. (2024), CDS Hu et al. (2025), TDMVI Huang et al. (2025a) and BCSI Huang et al. (2026). At 20% labels we obtain 88.49 Dice and 80.20 Jaccard (+1.63 Dice over BCSI, +2.37 Jaccard over TDMVI), with HD95 at 7.49. Additional baselines are in Appendix B.6.

ACDC Dataset Results. Multi-class cardiac segmentation stresses boundary quality. Table 3 reports SOTA results at both 10% and 20% against UA-MT Yu et al. (2019), SASSNet Li et al. (2020), DTC Luo et al. (2021), MC-Net Wu et al. (2022a), BCP Bai et al. (2023), DiffRect Liu et al. (2024a), CDS Hu et al. (2025) and IIIC Wu et al. (2026), with a large drop in HD95 / ASD. Fig. 3 shows smoother, GT-aligned contours. Additional baselines are in Appendix B.6.

Table 3: Comparison with SOTA semi-supervised segmentation methods on the ACDC dataset. The best results are highlighted in bold.

	10% Scans				20% Scans			
Method	Dice↑	Jaccard↑	HD95↓	ASD↓	Dice↑	Jaccard↑	HD95↓	ASD↓
UA-MT (MICCAI 2019)	78.65	66.97	6.12	1.71	81.11	69.92	7.95	2.23
SASSNet (MICCAI 2020)	84.50	74.34	5.42	1.86	86.55	77.90	4.47	1.10
DTC (AAAI 2021)	84.29	73.92	12.81	4.01	87.33	78.52	3.73	1.12
MC-Net (MedIA 2022)	86.44	77.04	5.50	1.84	88.51	80.19	5.35	1.54
BCP (CVPR 2023)	82.88	72.35	5.35	1.42	85.37	75.47	4.34	1.22
DiffRect (MICCAI 2024)	84.47	74.31	5.37	1.53	84.31	74.19	5.28	1.57
CDS (EAAI 2025)	80.58	69.60	5.23	1.58	84.47	74.61	3.60	0.97
IIIC (TNNLS 2026)	84.50	74.30	5.37	1.53	84.89	74.85	3.38	0.92
Ours	89.65	81.74	1.47	0.41	90.12	82.48	1.42	0.37

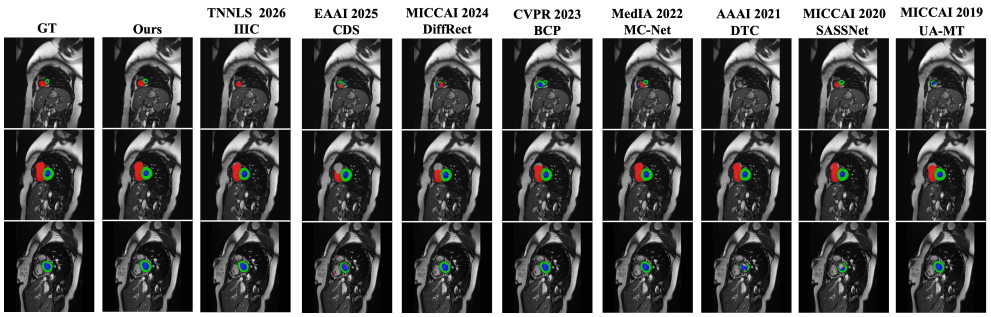

Figure 3: Visualization results of cardiac structure segmentation on the ACDC dataset with 20% labeled data. More visualization results in Appendix Fig. 8.

4.2 Ablation Studies

We further conduct ablation studies to analyze the contribution of each module. Most ablation studies are performed under the 20% labeled setting unless otherwise specified.

Table 4: Ablation study of core modules on the LA dataset (20% labeled). The best results are highlighted in bold.

Dataset	ABCP	UnCL	HCL	Dice↑	Jaccard↑	HD95↓	ASD↓
LA	–	–	–	89.53	81.23	13.03	3.70
LA	✓	–	–	90.82	83.33	6.84	2.23
LA	–	✓	–	90.49	82.74	6.19	2.03
LA	–	–	✓	89.96	81.94	7.40	1.60
LA	✓	✓	–	91.76	84.88	5.28	1.44
LA	✓	–	✓	91.69	84.75	5.34	1.54
LA	–	✓	✓	89.21	80.72	8.49	1.75
LA	✓	✓	✓	92.08	85.41	5.03	1.36

Table 5: Ablation study of V-DACL design on the BraTS-2019 dataset (20% labeled). The best results are highlighted in bold.

Method	Dice↑	Jaccard↑	HD95↓	ASD↓
Var 1: V-DACL w/o Bank	87.07	77.88	10.28	2.78
Var 2: w/o V-DACL	88.10	79.52	7.93	2.23
Var 3: Instance-level MAP	88.08	79.34	9.05	2.63
Var 4: Voxel-level Random	88.27	79.69	6.96	1.96
Var 5: V-DACL	88.49	80.20	7.49	1.32

Core Modules Ablation. As shown in Table 4, the full model that integrates all three components achieves the best performance, improving Dice by 2.55 compared with the baseline. From other combinations, we observe that ABCP provides the most direct performance gain, as it facilitates semantic knowledge transfer. UnCL improves region-based metrics in most combinations, except when combined with HCL alone (in the absence of ABCP). HCL is most effective when combined with ABCP-based data perturbation: without ABCP, the combination of UnCL and HCL alone may even slightly degrade performance, suggesting that data-level alignment is a prerequisite for feature-level contrastive learning.

V-DACL Internal Ablation. In addition to module combinations, we analyze the design of V-DACL module. Table 5 presents results of different V-DACL variants on BraTS-2019. The instance-level MAP contrastive learning (Var 3) does not bring improvement and performs similarly to the variant without V-DACL (Var 2). Both random voxel-level sampling (Var 4) and density-aware sampling (Var 5) substantially reduce HD and ASD, confirming that voxel-level sampling itself effectively improves boundary segmentation. Building on this,

Var 5 further achieves the highest Dice and Jaccard and the lowest ASD, showing that density-aware sampling preserves the boundary improvement of voxel-level supervision while delivering larger overall performance gains. In contrast, removing the memory bank (Var 1) leads to a clear performance drop, even falling below Var 2 (no V-DACL); this indicates that a cross-batch stable sample queue is crucial for voxel-wise contrastive learning in small-batch 3D medical image training, and without it the noisy negative samples within a single batch hurt feature learning more than they help.

Table 6: Effect of HCL under different labeling ratios on the BraTS-2019 dataset. Bar chart in Appendix Fig. 5.

Ratio	w/o HCL (Dice / Jaccard / HD95 / ASD)	w/ HCL (Dice / Jaccard / HD95 / ASD)	ΔDice	ΔJaccard	ΔHD95	ΔASD
5%	85.27 / 75.41 / 10.23 / 2.94	86.46 / 77.09 / 7.59 / 2.15	+1.19	+1.68	-2.64	-0.79
10%	85.82 / 76.23 / 10.07 / 2.75	86.34 / 77.21 / 8.57 / 1.74	+0.52	+0.98	-1.50	-1.01
20%	88.00 / 79.52 / 7.46 / 1.24	88.49 / 80.20 / 7.49 / 1.32	+0.49	+0.68	+0.03	+0.08

Impact of HCL Under Different Labeling Ratios. We further analyze the impact of HCL under different labeling ratios. As shown in Table 6, HCL consistently improves overlap metrics, with more significant gains in low-label ratio. These observations validate the motivation of the HCL design, showing that voxel-level feature constraints provide complementary supervision and are particularly beneficial in low-data scenarios.

Sensitivity Analysis. Scanning λ_{dacl} and λ_{dot} in Equation (18) on BraTS-2019 (20% labels) over a 1000× range yields stable Dice with no collapse (Appendix Table 8 and Appendix Fig. 4), so fine-grained retuning is unnecessary in practice.

5 Conclusion

This paper addresses three key challenges in semi-supervised 3D medical image segmentation: confirmation bias, insufficient boundary accuracy, and difficulty in semantic alignment. We propose a competitive hierarchical contrastive learning framework, termed CHCL-Seg. The framework employs UnCL to dynamically optimize pseudo-labels, effectively mitigating confirmation bias during semi-supervised training. Meanwhile, a hierarchical contrastive learning mechanism is introduced, which combines high level Dense-OT class prototype alignment with low level V-DACL hard boundary sample mining, jointly addressing semantic misalignment and boundary inaccuracy. Extensive experiments on three public datasets, LA, BraTS-2019, and ACDC, demonstrate that CHCL-Seg consistently outperforms existing state-of-the-art semi-supervised segmentation methods and shows strong generalization capability. This work provides a new solution for 3D medical image segmentation under low-annotation settings, and future work will further explore its adaptability and deployment potential on multi-center clinical data.

References

Sukesh Adiga, Jose Dolz, and Herve Lombaert. Anatomically-aware uncertainty for semi-supervised image segmentation. Medical Image Analysis, 91:103011, 2024.

Maregu Assefa, Muzammal Naseer, Iyyakutti Iyappan Ganapathi, Syed Sadaf Ali, Mohamed L Seghier, and Naoufel Werghi. Dycon: Dynamic uncertainty-aware consistency and contrastive learning for semi-supervised medical image segmentation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 30850–30860, 2025.

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 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11514–11524, 2023.

Spyridon Bakas. BraTS MICCAI brain tumor dataset. IEEE Dataport, 2020. URL https://dx.doi.org/10.21227/hdtd-5j88.

Andreas Bar, Fabian Huger, Peter Schlicht, and Tim Fingscheidt. On the robustness of redundant teacher-student frameworks for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0, 2019.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohé, Xavier Pennec, Maxime Sermesant, Fabian Isensee, Paul Jäger, Klaus H. Maier-Hein, Peter M. Full, Ivo Wolf, Sandy Engelhardt, Christian F. Baumgartner, Lisa M. Koch, Jelmer M. Wolterink, Ivana Išgum, Yeonggul Jang, Yoonmi Hong, Jay Patravali, Shubham Jain, Olivier Humbert, and Pierre-Marc Jodoin. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Transactions on Medical Imaging, 37(11):2514–2525, 2018. doi: 10.1109/TMI.2018.2837502.

Chen Chen, Wenjia Bai, and Daniel Rueckert. Multi-task learning for left atrial segmentation on ge-mri. In International workshop on statistical atlases and computational models of the heart, pp. 292–301. Springer, 2018.

Jiahe Chen, Jiahe Ying, Shen Wang, and Jianwei Zheng. Decoupled competitive framework for semi-supervised medical image segmentation. arXiv preprint arXiv:2505.24667, 2025a.

Xi Chen, Lyuyang Tong, Huangxuan Zhao, and Bo Du. Uncertainty-guided adaptive correction for semi-supervised medical image segmentation. IEEE Transactions on Image Processing, 34:7975–7988, 2025b.

Ziming Cheng, Shidong Wang, Tong Xin, and Haofeng Zhang. Frequency-aware adaptive filtering network for few-shot medical image segmentation. In 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1890–1897. IEEE, 2024.

Veronika Cheplygina, Marleen De Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical image analysis, 54:280–296, 2019.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems, 26, 2013.

Min Dong, Ating Yang, Zhenhang Wang, Dezhen Li, Jing Yang, and Rongchang Zhao. Uncertainty-aware consistency learning for semi-supervised medical image segmentation. Knowledge-Based Systems, 309:112890, 2025.

Shujian Gao, Yuan Wang, Jiacheng Yang, Mengwen Ye, Weifan Liu, and Zekuan Yu. Tricsart: Semi-supervised medical image segmentation with triple-level contrastive learning and selective active re-training. IEEE Journal of Biomedical and Health Informatics, 2026.

Mikhail Goncharov, Vera Soboleva, Anvar Kurmukov, Maxim Pisov, and Mikhail Belyaev. vox2vec: A framework for self-supervised contrastive learning of voxel-level representations in medical images. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 605–614. Springer, 2023.

Xiurui Guo, Kai Sun, and Yuanjie Zheng. Multi-scale consistency adversarial learning for semi-supervised 3d medical image segmentation. Biomedical Signal Processing and Control, 103:107475, 2025.

Kai Han, Shuhui Wang, Jun Chen, Chengxuan Qian, Chongwen Lyu, Siqi Ma, Chengjian Qiu, Victor S Sheng, Qingming Huang, and Zhe Liu. Region uncertainty estimation for medical image segmentation with noisy labels. IEEE Transactions on Medical Imaging, 2025.

Wenlong Hang, Beijing Wang, Shuang Liang, Qingfeng Zhang, Qiang Wu, Yukun Jin, Qiong Wang, and Jing Qin. Conflict-aware semi-supervised mutual learning for medical image segmentation. Expert Systems with Applications, pp. 131544, 2026.

Gang Hu, Feng Zhao, and Essam H Houssein. Competitive dual-students using bi-level contrastive learning for semi-supervised medical image segmentation. Engineering Applications of Artificial Intelligence, 144:110082, 2025.

Kaiwen Huang, Yi Zhou, Huazhu Fu, Yizhe Zhang, Chen Gong, and Tao Zhou. Text-driven multiplanar visual interaction for semi-supervised medical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 604–613. Springer, 2025a.

Kaiwen Huang, Yizhe Zhang, Yi Zhou, Tianyang Xu, and Tao Zhou. Bidirectional channel-selective semantic interaction for semi-supervised medical segmentation. arXiv preprint arXiv:2601.05855, 2026.

Senlong Huang, Yongxin Ge, Dongfang Liu, Mingjian Hong, Junhan Zhao, and Alexander C Loui. Rethinking copy-paste for consistency learning in medical image segmentation. IEEE Transactions on Image Processing, 2025b.

Tengteng Huang, Yifan Sun, Xun Wang, Haotian Yao, and Chi Zhang. Spatial ensemble: a novel model smoothing mechanism for student-teacher framework. Advances in Neural Information Processing Systems, 34:15957–15968, 2021.

Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 488–498. Springer, 2024.

Xibin Jia, Wang Zhang, Luo Wang, Chuanxu Yang, Xunjie Yin, Hao Jia, Yiming Zheng, Zhenghan Yang, Dawei Yang, Min Hong, et al. Semi-supervised boundary-aware medical image segmentation via symmetric boundary-foreground collaboration. Expert Systems with Applications, 295:128764, 2026.

Suruchi Kumari and Pravendra Singh. Leveraging fixed and dynamic pseudo-labels in cross-supervision framework for semi-supervised medical image segmentation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 3491–3501, 2025.

Congcong Li, Jinshuo Zhang, Dongmei Niu, Xiuyang Zhao, Bo Yang, and Caiming Zhang. Boundary-aware uncertainty suppression for semi-supervised medical image segmentation. IEEE Transactions on Artificial Intelligence, 5(8):4074–4086, 2024.

Shuailin Li, Chuyu Zhang, and Xuming He. Shape-aware semi-supervised 3d semantic segmentation for medical images. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 552–561. Springer, 2020.

Shumeng Li, Lei Qi, Qian Yu, Jing Huo, Yinghuan Shi, and Yang Gao. Stitching, fine-tuning, and re-training: A sam-enabled framework for semi-supervised 3d medical image segmentation. IEEE Transactions on Medical Imaging, 44(10):3909–3923, 2025a.

Xiaokang Li, Menghua Xia, Jing Jiao, Shichong Zhou, Cai Chang, Yuanyuan Wang, and Yi Guo. Hal-ia: A hybrid active learning framework using interactive annotation for medical image segmentation. Medical Image Analysis, 88:102862, 2023.

Zihan Li, Dandan Shan, Yunxiang Li, Paul E Kinahan, and Qingqi Hong. Scale-aware adaptive supervised network with limited medical annotations. Pattern Recognition, pp. 112743, 2025b.

Xinyu Liu, Wuyang Li, and Yixuan Yuan. Diffrect: Latent diffusion label rectification for semi-supervised medical image segmentation. In International conference on medical image computing and computer-assisted intervention, pp. 56–66. Springer, 2024a.

Xinyu Liu, Zhen Chen, Wuyang Li, Chenxin Li, and Yixuan Yuan. Harnessing lightweight transformer with contextual synergic enhancement for efficient 3d medical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025.

Yuyuan Liu, Yu Tian, Chong Wang, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Translation consistent semi-supervised segmentation for 3d medical images. IEEE Transactions on Medical Imaging, 44(2):952–968, 2024b.

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.

Qinghe Ma, Jian Zhang, Lei Qi, Qian Yu, Yinghuan Shi, and Yang Gao. Constructing and exploring intermediate domains in mixed domain semi-supervised medical image segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11642–11651, 2024.

Juzheng Miao, Si-Ping Zhou, Guang-Quan Zhou, Kai-Ni Wang, Meng Yang, Shoujun Zhou, and Yang Chen. Sc-ssl: Self-correcting collaborative and contrastive co-training model for semi-supervised medical image segmentation. IEEE Transactions on Medical Imaging, 43(4):1347–1364, 2023.

Seok-Hwan Oh, Guil Jung, Sang-Yun Kim, Myeong-Gee Kim, Young-Min Kim, Hyeon-Jik Lee, Hyuk-Sool Kwon, and Hyeon-Min Bae. Uncertainty-aware meta-weighted optimization framework for domain-generalized medical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 775–785. Springer, 2024.

Himashi Peiris, Munawar Hayat, Zhaolin Chen, Gary Egan, and Mehrtash Harandi. Uncertainty-guided dual-views for semi-supervised volumetric medical image segmentation. Nature Machine Intelligence, 5(7):724–738, 2023.

Jizong Peng, Ping Wang, Christian Desrosiers, and Marco Pedersoli. Self-paced contrastive learning for semi-supervised medical image segmentation with meta-labels. In Advances in Neural Information Processing Systems, volume 34, pp. 16686–16699, 2021.

Wenbo Qi, Jiafei Wu, and Shing-Chow Chan. Gradient-aware for class-imbalanced semi-supervised medical image segmentation. In European Conference on Computer Vision, pp. 473–490. Springer, 2024.

Zhongxi Qiu, Yan Hu, Xiaoshan Chen, Dan Zeng, Qingyong Hu, and Jiang Liu. Rethinking dual-stream super-resolution semantic learning in medical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(1):451–464, 2023.

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.

Mu Tian, Qinzhu Yang, and Yi Gao. Multi-scale multi-task distillation for incremental 3d medical image segmentation. In European Conference on Computer Vision, pp. 369–384. Springer, 2022.

Mengzhu Wang, Houcheng Su, Jiao Li, Chuan ji Li, Nan Yin, Li Shen, and Jingcai Guo. Graphcl: Graph-based clustering for semi-supervised medical image segmentation. In International Conference on Machine Learning, 2025a. URL https://api.semanticscholar.org/CorpusID:283566988.

Tao Wang, Xinlin Zhang, Yuanbin Chen, Yuanbo Zhou, Longxuan Zhao, Tao Tan, and Tong Tong. Synergy-guided regional supervision of pseudo labels for semi-supervised medical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 530–540. Springer, 2025b.

Yongchao Wang, Bin Xiao, Xiuli Bi, Weisheng Li, and Xinbo Gao. Mcf: Mutual correction framework for semi-supervised medical image segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 15651–15660, 2023.

Ziyang Wang and Congying Ma. Dual-contrastive dual-consistency dual-transformer: A semi-supervised approach to medical image segmentation. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 870–879, 2023.

Shuai Wu, Ruyi Liu, Hang Wei, Linrunjia Liu, Jie Wen, and Qiguang Miao. Boosting semi-supervised medical image segmentation through inter-instance information complementarity. IEEE Transactions on Neural Networks and Learning Systems, 2026.

Yicheng Wu, Zongyuan Ge, Donghao Zhang, Minfeng Xu, Lei Zhang, Yong Xia, and Jianfei Cai. Mutual consistency learning for semi-supervised medical image segmentation. Medical image analysis, 81:102530, 2022a.

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, 2022b.

Xianrun Xu, Baoyao Yang, Wanyun Li, Jingsong Lin, and Yufei Xu. Simple but effective: Sub-volume contrastive learning for class-imbalanced semi-supervised 3d medical image segmentation. In Proceedings of the 33rd ACM International Conference on Multimedia, pp. 4884–4893, 2025.

Ke Yan, Qing Cai, Fan Zhang, Ziyan Cao, and Zhi Liu. Sgtc: Semantic-guided triplet co-training for sparsely annotated semi-supervised medical image segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pp. 9112–9120, 2025.

Yang Yang, Jiaxin Zhuang, Guoying Sun, Ruixuan Wang, and Jingyong Su. Boundary-guided contrastive learning for semi-supervised medical image segmentation. IEEE Transactions on Medical Imaging, 2025.

Chenyu You, Yuan Zhou, Ruihan Zhao, Lawrence Staib, and James S Duncan. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. IEEE Transactions on Medical Imaging, 41(9):2228–2237, 2022.

Chenyu You, Weicheng Dai, Yifei Min, Fenglin Liu, David Clifton, S Kevin Zhou, Lawrence Staib, and James Duncan. Rethinking semi-supervised medical image segmentation: A variance-reduction perspective. Advances in neural information processing systems, 36: 9984–10021, 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.

Yuchen Yuan, Xi Wang, Xikai Yang, and Pheng-Ann Heng. Effective semi-supervised medical image segmentation with probabilistic representations and prototype learning. IEEE Transactions on Medical Imaging, 44(3):1181–1193, 2024.

Qingjie Zeng, Yutong Xie, Zilin Lu, Mengkang Lu, Yicheng Wu, and Yong Xia. Segment together: A versatile paradigm for semi-supervised medical image segmentation. IEEE Transactions on Medical Imaging, 2025.

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 Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 25940–25949, 2025.

Haifeng Zhao, Haiyang Li, Lei-Lei Ma, and Dengdi Sun. Segment anything model meets semi-supervised medical image segmentation: A novel perspective. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), Advances in Neural Information Processing Systems, volume 38, pp. 170128–170148. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/f88306d7c33a024c255ed6aa5e1211fd-Paper-Conference.pdf.

Quan Zhou, Yewei Feng, Zhiwen Huang, Mingyue Ding, and Xuming Zhang. Specific instance and cross-prompt-based robust 3-d semi-supervised medical image segmentation. IEEE Transactions on Instrumentation and Measurement, 73:1–14, 2024a.

Quan Zhou, Bin Yu, Feng Xiao, Mingyue Ding, Zhiwei Wang, and Xuming Zhang. Robust semi-supervised 3d medical image segmentation with diverse joint-task learning and decoupled inter-student learning. IEEE Transactions on Medical Imaging, 43(6):2317–2331, 2024b.

Reproducibility Statement

We provide the information needed to reproduce the main results in this paper. Dataset splits, preprocessing, licenses, and augmentation follow Appendix B.1. Backbone choice, optimizer, learning-rate schedule, EMA, two-phase training, per-dataset iteration counts, ABCP/Dense-OT/V-DACL hyperparameters, hardware, wall-clock time, and random seeds are listed in Appendix B.2. Evaluation metrics are defined in Appendix B.3. Sensitivity of the two contrastive weights is reported in Appendix Table 8 and Appendix Fig. 4. Code will be released upon acceptance.

AI Use Statement

In this work, we used generative AI tools to polish the writing of the manuscript (grammar, wording, and readability). We also used them to help debug implementation code. All scientific contributions, including the problem formulation, method design, experimental design, and interpretation of results, were made by the authors. We did not use generative AI tools to generate synthetic datasets, prove mathematical claims, propose the research idea, design the method or experiments, or draft the scientific content of the paper. We reviewed all AI-assisted text and code, and we take responsibility for the final content of this work.

A Technical Appendices

A.1 Loss Weight Scheduling Strategy

A.1.1 V-DACL weight scheduling $w_{dacl}(t)$

V-DACL adopts a five-stage weight schedule:

$$
\begin{aligned}
&\text{(i)} \quad t < t_w^s: \quad w_{dacl}(t) = 0;\\
&\text{(ii)} \quad t_w^s \le t < t_w^e: \quad w_{dacl}(t) = \exp\left(-5\left(1-\tilde{t}_w\right)^2\right),\\
&\qquad \text{where } \tilde{t}_w = \frac{t - t_w^s}{t_w^e - t_w^s};\\
&\text{(iii)} \quad t_w^e \le t \le t_a^s: \quad w_{dacl}(t) = 1;\\
&\text{(iv)} \quad t_a^s < t < t_a^e: \quad w_{dacl}(t) = \max\left(0,\ 1-\tilde{t}_a\right),\\
&\qquad \text{where } \tilde{t}_a = \frac{t - t_a^s}{t_a^e - t_a^s};\\
&\text{(v)} \quad t \ge t_a^e: \quad w_{dacl}(t) = 0.
\end{aligned}
$$

The motivation is as follows. After ABCP annealing ends, the data level no longer provides cross-distribution mixing signals. Therefore, the V-DACL weight is linearly decayed to prevent the feature-level contrastive loss from dominating the gradient direction at the end of training and interfering with the stabilized decision boundary.

A.1.2 Dense-OT weight scheduling $w_{dot}(t)$

Dense-OT adopts a linear warm-up schedule, different from the Gaussian warm-up used in V-DACL. It consists of three stages:

(i) $t < t_w^s: \quad w_{dot}(t) = 0;$
The prototypes are not well initialized, and the Sinkhorn assignment is unstable. Thus, the OT loss is completely disabled.

(ii) $t_w^s \le t < t_w^e: \quad w_{dot}(t) = \dfrac{t - t_w^s}{t_w^e - t_w^s};$
The weight increases linearly from 0 to 1, enabling a smooth transition to the full OT constraint. A linear schedule is used instead of a Gaussian one because the gradient signal of Dense-OT comes from Sinkhorn soft assignment, which is relatively stable after prototype initialization. Therefore, it does not require a convex Gaussian curve to suppress early noise.

(iii) $t \ge t_w^e: \quad w_{dot}(t) = 1.$

Note: Unlike V-DACL, Dense-OT does not decay after ABCP annealing. This is because Dense-OT operates on original unlabeled images (rather than BCP-mixed images), and its spatial correspondence is not affected by ABCP annealing. Therefore, no annealing is required.

B Supplementary Material

B.1 Datasets and Preprocessing

We evaluate the proposed method on three widely used 3D medical image segmentation benchmarks.

LA Chen et al. (2018). The Left Atrium dataset contains 100 gadolinium-enhanced MR scans with expert-annotated left-atrium masks. Following the standard split adopted by BCP Bai et al. (2023) and CDS Hu et al. (2025), we use 80 scans for training and 20 for testing. Labeled subsets of 8 and 16 scans are used for the 10% and 20% labeled settings, respectively. Each volume is z-score normalized to zero mean and unit variance, and cropped to a tight bounding box around the left atrium with a small random margin in each dimension to retain surrounding context. During training, sub-volumes of size $112\times112\times80$ are randomly cropped from this region.

BraTS-2019 Bakas (2020). The dataset consists of 335 multi-modal MRI scans (HGG and LGG) with whole-tumor annotations. Following the standard split used in prior semi-supervised work Bai et al. (2023); Hu et al. (2025), we use 250 scans for training, 25 for validation, and 60 for testing. Labeled subsets of 25 and 50 scans are used for the 10% and 20% labeled settings, respectively. Each scan is intensity-normalized within the brain region and cropped to a tight bounding box around the brain. During training, sub-volumes of size $96\times96\times96$ are randomly sampled.

ACDC Bernard et al. (2018). The Automated Cardiac Diagnosis Challenge dataset contains short-axis cine-MRI scans from 100 patients, with three foreground classes: right ventricle, left-ventricle myocardium, and left ventricle. We follow the patient-level split commonly used in semi-supervised cardiac segmentation Bai et al. (2023); Liu et al. (2024a), with 70 patients for training, 10 for validation, and 20 for testing. Labeled subsets of 7 and 14 patients are used for the 10% and 20% labeled settings, respectively (selection is performed at patient level, so all slices belonging to a labeled patient are considered labeled). Slices are pre-processed following the pipeline of Peng et al. Peng et al. (2021) to a common in-plane size of 384×384; during training we randomly crop 256×256 patches for 2D U-Net optimization, and at inference each slice is resized to 256×256 via cubic interpolation, with predictions then mapped back to the original resolution for per-volume metric computation.

Data augmentation. For all datasets we apply random scaling, random rotation and flipping, random cropping, elastic deformation, additive Gaussian noise, and Gaussian blur during training, following the augmentation pipeline used in CDS Hu et al. (2025).

Dataset licenses and terms of use. All three datasets are publicly released benchmarks distributed for academic and non-commercial research only. The LA dataset Chen et al. (2018) is provided by the MICCAI 2018 Atrial Segmentation Challenge organizers under the challenge's data use agreement. The BraTS-2019 dataset Bakas (2020) is released by the Center for Biomedical Image Computing and Analytics (CBICA) under the BraTS challenge data use agreement, with all images previously de-identified by the providers. The ACDC dataset Bernard et al. (2018) is released by CREATIS for non-commercial research use following the MICCAI 2017 challenge protocol. We use all three datasets strictly for academic research, in compliance with their respective terms of use, and do not redistribute the raw imaging data.

B.2 Implementation Details

Backbone networks. Following the CDS protocol, we use 3D V-Net for LA and BraTS-2019, and 2D U-Net for ACDC. Both architectures are kept identical to the corresponding CDS baseline so that all reported gains are attributable to the proposed CHCL-Seg components rather than backbone differences.

Optimization. All models are trained with SGD (momentum 0.9, weight decay 1×10^{-4}). The base learning rate is 0.01 with linear warm-up over the first 1,000 iterations and cosine decay down to 5% of the base value; the teacher network is updated by an exponential moving average (EMA) of the student parameters with decay coefficient 0.99. Following the BCP

two-phase protocol, training comprises a labeled-only pre-training phase (BCP supervised loss only) and a semi-supervised self-training phase in which ABCP, UnCL, Dense-OT, and V-DACL are jointly optimized. The two phases use independent iteration counters. Per-dataset settings are summarized in Table 7.

Table 7: Per-dataset training configuration. Each mini-batch is split evenly into a labeled stream and an unlabeled stream by a two-stream sampler; in the pre-training phase only the labeled stream is used.

Dataset	Backbone	Pre-train iter	Self-train iter	Batch (L:U)
LA	3D V-Net	2,000	10,000	8 (4:4)
BraTS-2019	3D V-Net	2,000	12,000	8 (4:4)
ACDC	2D U-Net	15,000	60,000	24 (12:12)

ABCP scheduling. The bidirectional copy-paste mixing probability is annealed linearly from 1.0 to 0 between iterations 3,000 and 5,000.

Dense-OT hyperparameters. The class-prototype dimension is 256, Sinkhorn–Knopp regularization $\epsilon = 0.05$ with 3 iterations, prototype EMA momentum 0.99, and contrastive temperature $\tau_{dot} = 0.1$. The Dense-OT loss weight is set to $\lambda_{dot} = 0.05$ (the best value selected by the sensitivity analysis in Sec. 4.2) and follows a linear warm-up between iterations 1,000 and 2,000.

V-DACL hyperparameters. V-DACL operates on the smallest spatial element of each backbone: a 3D voxel for LA and BraTS-2019 (3D V-Net) and a 2D pixel for ACDC (2D U-Net); for notational consistency we use the term voxel throughout. On ACDC we adopt the original instance-level Masked Average Pooling formulation of DACL (one prototype per sample-class), since the 2D U-Net bottleneck has very few pixels per class per slice; on LA and BraTS-2019 we adopt the proposed voxel-level extension of V-DACL. The projected feature dimension is 256, the memory bank stores up to 2,000 features per class, the multi-scale density neighborhoods are $\{10, 20, 50\}$, and the contrastive temperature is $\tau_{dacl} = 0.4$. At each iteration we randomly sample 32 voxels per (sample, class) from foreground regions whose teacher confidence exceeds 0.8 to form the per-iteration feature pool. Within each class we then select $N_q = 256$ low-density anchors, $N_p^+ = 512$ high-density positives (half from the current batch and half from the memory bank), and $N_p^- = 512$ negatives drawn from other classes. The V-DACL weight $\lambda_{dacl} = 0.03$ (the best value selected by the sensitivity analysis in Sec. 4.2) follows a Gaussian warm-up between iterations 1,000 and 2,000, after which it linearly decays in lock-step with the end of ABCP annealing (see Appendix A.1).

Hardware and training time. All experiments are conducted on a single NVIDIA RTX 4090 GPU (24 GB memory) using PyTorch 2.8.0 with CUDA 12.8. A complete training run takes approximately 14 hours on LA, 9 hours on BraTS-2019, and 12 hours on ACDC. Random seeds are fixed (seed = 1337 for the main training pipeline; the ACDC patient-level split additionally uses seed = 42 following the BCP/SSL4MIS convention) and cudnn.deterministic is enabled to ensure reproducibility.

B.3 Evaluation Metrics

We report four standard segmentation metrics. Let A denote the predicted foreground voxel set and B the ground-truth foreground voxel set, and let ∂A, ∂B denote their respective surface voxels.

Dice Similarity Coefficient (DSC). Measures volumetric overlap; higher is better.

$$\mathrm{DSC}(A, B) = \frac{2\,|A \cap B|}{|A| + |B|}. \tag{19}$$

Jaccard Coefficient (JC). Equivalent to Intersection-over-Union; higher is better.

$$\mathrm{JC}(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{20}$$

95% Hausdorff Distance (HD95). The 95-th percentile of all bidirectional surface-point distances; lower is better. It is more robust to a small number of outlier voxels than the maximum Hausdorff distance.

$$\mathrm{HD95}(A,B) = \mathrm{P}_{95}(\{\, d(a,\partial B) : a \in \partial A \,\} \cup \{\, d(b,\partial A) : b \in \partial B \,\}), \tag{21}$$

where $d(\cdot,\cdot)$ denotes the Euclidean distance and $\mathrm{P}_{95}(\cdot)$ the 95-th percentile.

Average Surface Distance (ASD). The mean of bidirectional surface-point distances; lower is better. It reflects the overall surface quality across the entire boundary rather than extreme outliers.

$$\mathrm{ASD}(A,B) = \frac{1}{|\partial A| + |\partial B|} \left(\sum_{a \in \partial A} d(a,\partial B) + \sum_{b \in \partial B} d(b,\partial A) \right). \tag{22}$$

B.4 Sensitivity Analysis of Loss Weights in Equation (18)

Table 8: Sensitivity analysis of loss weights on the BraTS-2019 dataset. The model shows strong robustness across a 1000× search range, with low variance and no performance collapse.

Scanned weight	Fixed weight	Search range	Dice range	Dice mean ± std	Default reference
λ_{dacl}	$\lambda_{dot} = 0.05$	0.001–1.0	85.40–88.49	86.16 ± 0.79	0.03 / 88.49
λ_{dot}	$\lambda_{dacl} = 0.03$	0.001–1.0	85.44–88.49	86.39 ± 0.72	0.05 / 88.49

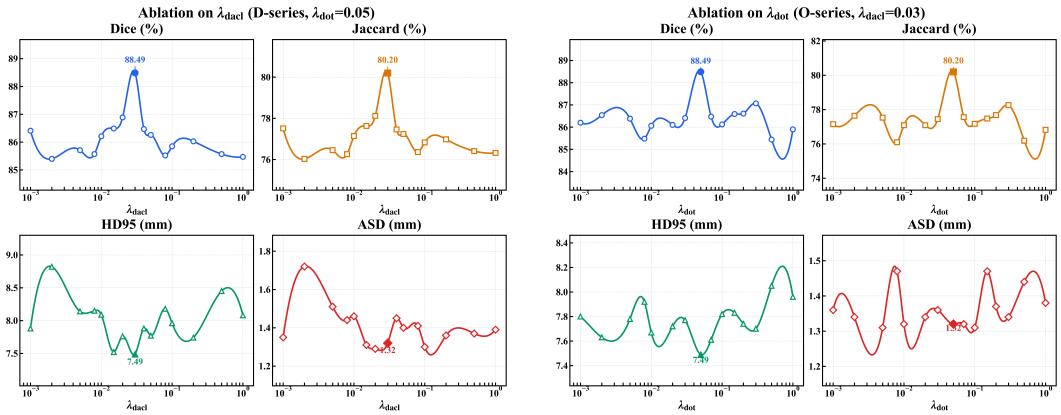

Figure 4: Dice curves over the logarithmic search range on BraTS-2019 (20% labeled). Left: sensitivity to λ_{dacl}. Right: sensitivity to λ_{dot}. Performance is stable across three orders of magnitude.

B.5 Influence of Feature-Level Contrastive Learning under Different Annotation Ratios

We further examine the effect of HCL under different labeling ratios to validate the motivation of the HCL design. The goal is to supplement supervision with voxel-level feature constraints, enabling more stable representations when labeled data is limited.

Low-label ratio (5% and 10%): HCL provides clear gains. Dice improves by +1.19% and +0.52%, and Jaccard improves by +1.68% and +0.98%. At the same time, HD95 decreases by 2.64 voxels and 1.50 voxels, and ASD is also reduced. This shows that contrastive constraints improve segmentation quality and boundary stability when supervision is limited.

Mid-label ratio (20%): HCL mainly improves region overlap metrics. Dice and Jaccard increase by +0.49% and +0.68%, while boundary metrics change only slightly (HD95 +0.03, ASD +0.08). This suggests that when labels are sufficient, boundaries are already well constrained, and HCL mainly enhances overlap quality.

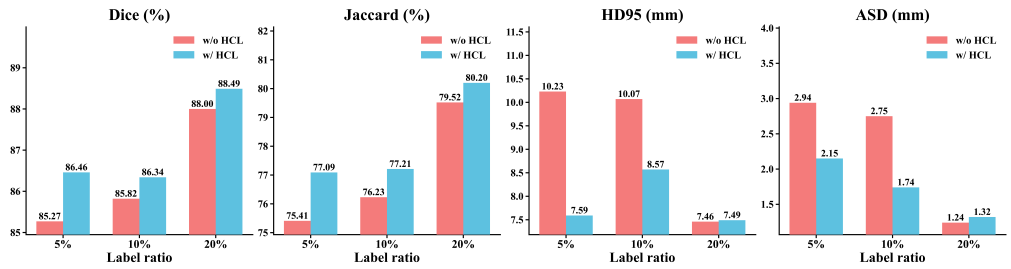

Figure 5: Effect of HCL under different labeling ratios on BraTS-2019. From left to right: Dice, Jaccard, HD95, and ASD. Each group compares results without and with HCL, with values annotated on top of the bars. At 5% and 10% labeling ratios, HCL improves all metrics consistently, including reductions of 2.64 voxels and 1.50 voxels in HD95. At 20%, the gains mainly appear in overlap metrics (Dice +0.49, Jaccard +0.68), consistent with Table 6.

Design motivation: Overall, these results support the motivation of HCL. Voxel-level feature constraints provide additional supervision and lead to more stable representations under limited labels. The fewer the labels, the larger the gain from HCL, which is consistent with the intuition that feature regularization is more important in low-data scenario.

B.6 Full Comparison Results

Due to space limitations in the main text, only a subset of methods is reported. Here we provide full comparisons with all baselines across datasets. The best results are shown in bold.

Table 9: Comparison with state-of-the-art semi-supervised segmentation methods on the LA dataset with 20% labels.

Method	Dice ↑	Jaccard ↑	HD95 ↓	ASD ↓
UA-MT (MICCAI 2019) Yu et al. (2019)	88.17	79.21	11.24	3.07
SASSNet (MICCAI 2020) Li et al. (2020)	89.17	80.69	8.57	2.87
DTC (AAAI 2021) Luo et al. (2021)	89.43	81.01	7.76	1.90
BCP (CVPR 2023) Bai et al. (2023)	90.16	82.25	6.93	1.94
CDS (EAAI 2025) Hu et al. (2025)	90.33	82.46	6.16	1.91
VerSemi (TMI 2025) Zeng et al. (2025)	91.29	84.08	5.62	1.86
TraCoCo (TMI 2025) Liu et al. (2024b)	89.41	80.95	8.30	2.01
SGTC (AAAI 2025) Yan et al. (2025)	87.30	77.70	16.28	4.74
BoCLIS (TMI 2025) Yang et al. (2025)	90.91	83.29	6.08	1.64
GraphCL (ICML 2025) Wang et al. (2025a)	90.43	82.74	6.29	1.73
SKCDF (CVPR 2025) Zhang et al. (2025)	90.65	82.98	6.70	1.61
UnCo (TMI 2025) Chen et al. (2025b)	90.91	83.40	5.36	1.61
SBFC (ESWA 2026) Jia et al. (2026)	91.43	84.29	5.51	1.55
BCSI (AAAI 2026) Huang et al. (2026)	91.27	84.08	5.67	1.58
Ours	92.08	85.41	5.03	1.36

Table 10: Comparison with state-of-the-art methods on the LA dataset with 10% labels.

Method	Dice ↑	Jaccard ↑	HD95 ↓	ASD ↓
UA-MT (MICCAI 2019) Yu et al. (2019)	86.28	76.11	18.71	4.63
SASSNet (MICCAI 2020) Li et al. (2020)	85.22	75.09	11.18	2.89
DTC (AAAI 2021) Luo et al. (2021)	81.89	71.23	15.81	3.80
BCP (CVPR 2023) Bai et al. (2023)	89.09	80.49	7.49	1.95
DCF (ECAI 2024) Chen et al. (2025a)	89.94	81.78	6.38	1.80
CDS (EAAI 2025) Hu et al. (2025)	87.21	77.56	11.02	2.97
TraCoCo (TMI 2025) Liu et al. (2024b)	89.29	80.82	6.92	2.28
BoCLIS (TMI 2025) Yang et al. (2025)	90.16	82.50	6.91	1.92
VerSemi (TMI 2025) Zeng et al. (2025)	89.93	81.27	8.62	2.38
SGTC (AAAI 2025) Yan et al. (2025)	84.70	73.80	17.44	4.26
GraphCL (ICML 2025) Wang et al. (2025a)	90.24	82.31	6.42	1.71
SAMM (NeurIPS 2025) Zhao et al. (2025)	90.47	82.71	6.41	1.97
ST (TMI 2025) Zeng et al. (2025)	89.83	81.27	8.62	2.38
UnCo (TMI 2025) Chen et al. (2025b)	90.37	82.54	6.11	1.74
SKCDF (CVPR 2025) Zhang et al. (2025)	89.42	80.46	8.18	2.38
SBFC (ESWA 2026) Jia et al. (2026)	90.23	82.30	7.93	1.66
BCSI (AAAI 2026) Huang et al. (2026)	91.07	83.67	5.57	1.53
Ours	91.34	84.19	5.72	1.46

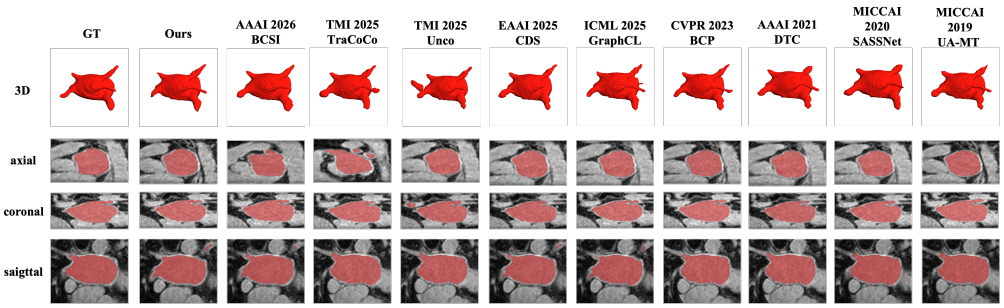

Figure 6: Visualization results of LA dataset with 20% labeled data.

Table 11: Comparison on BraTS-2019 with 20% labels.

Method	Dice ↑	Jaccard ↑	HD95 ↓	ASD ↓
UA-MT (MICCAI 2019) Yu et al. (2019)	85.32	75.93	8.68	1.98
SASSNet (MICCAI 2020) Li et al. (2020)	85.64	76.33	9.17	2.04
DTC (AAAI 2021) Luo et al. (2021)	83.80	73.24	9.55	2.97
BCP (CVPR 2023) Bai et al. (2023)	86.47	77.41	7.88	1.44
AUSS (MedIA 2024) Adiga et al. (2024)	83.21	73.55	11.39	2.03
MiDSS (CVPR 2024) Ma et al. (2024)	83.74	73.93	11.06	2.29
GALoss (ECCV 2024) Qi et al. (2024)	83.49	73.72	10.61	2.18
CDS (EAAI 2025) Hu et al. (2025)	86.12	77.58	7.87	1.83
SAMM (NeurIPS 2025) Zhao et al. (2025)	86.46	77.21	7.64	2.21
TDMVI (MICCAI 2025) Huang et al. (2025a)	86.69	77.83	7.65	1.24
SASNet (PR 2025) Li et al. (2025b)	85.84	76.79	7.52	1.62
SGRS (MICCAI 2025) Wang et al. (2025b)	85.91	76.59	8.69	2.36
SKCDF (CVPR 2025) Zhang et al. (2025)	84.40	74.78	9.31	1.68
BCSI (AAAI 2026) Huang et al. (2026)	86.86	77.80	7.62	1.26
Ours	88.49	80.20	7.49	1.32

Table 12: Comparison on BraTS-2019 with 10% labels.

Method	Dice ↑	Jaccard ↑	HD95 ↓	ASD ↓
DTC (AAAI 2021) Luo et al. (2021)	82.74	72.74	11.76	3.24
URPC (MedIA 2022) Luo et al. (2022)	84.16	74.29	11.01	2.63
SS-Net (MICCAI 2022) Wu et al. (2022b)	82.00	71.82	10.68	1.82
BCP (CVPR 2023) Bai et al. (2023)	84.35	75.01	10.95	2.60
MCF (CVPR 2023) Wang et al. (2023)	79.28	69.08	12.43	3.58
AUSS (MedIA 2024) Adiga et al. (2024)	81.99	71.92	11.47	2.44
MiDSS (CVPR 2024) Ma et al. (2024)	81.91	71.03	11.38	2.37
GALoss (ECCV 2024) Qi et al. (2024)	82.13	72.11	11.25	2.39
CDS (EAAI 2025) Hu et al. (2025)	84.03	74.25	9.67	2.60
SGRS (MICCAI 2025) Wang et al. (2025b)	85.67	76.05	9.33	2.76
TDMVI (MICCAI 2025) Huang et al. (2025a)	85.27	75.23	8.77	1.35
SASNet (PR 2025) Li et al. (2025b)	82.84	73.00	10.91	2.31
IIIC (TNNLS 2026) Wu et al. (2026)	85.49	76.21	9.34	1.92
SKCDF (CVPR 2025) Zhang et al. (2025)	83.00	73.08	10.28	1.78
CSE-Light-UNETR (TPAMI 2026) Liu et al. (2025)	79.73	68.76	11.65	2.25
BCSI (AAAI 2026) Huang et al. (2026)	86.17	77.10	8.43	1.37
Ours	86.34	77.21	8.57	1.74

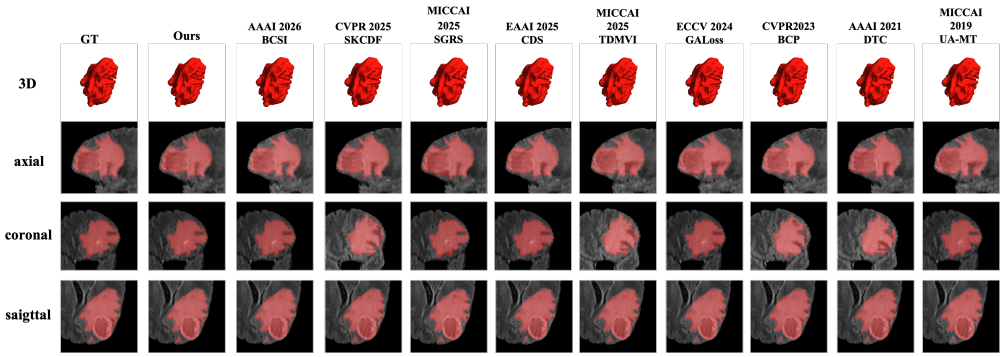

Figure 7: Visualization results of BraTS-2019 dataset with 20% labeled data.

Table 13: Ablation study of core modules on the BraTS-2019 dataset. The best results are highlighted in bold.

Dataset	ABCP	UnCL	HCL	Dice↑	Jaccard↑	HD95↓	ASD↓
BraTS-2019	–	–	–	85.05	75.37	10.48	2.81
BraTS-2019	✓	–	–	87.09	78.34	8.17	1.36
BraTS-2019	–	✓	–	86.03	76.69	8.12	1.37
BraTS-2019	–	–	✓	85.81	76.54	8.57	1.33
BraTS-2019	✓	✓	–	88.00	79.52	7.46	1.24
BraTS-2019	✓	–	✓	86.97	78.16	7.61	1.33
BraTS-2019	–	✓	✓	86.26	77.16	8.17	1.23
BraTS-2019	✓	✓	✓	88.49	80.20	7.49	1.32

Table 14: Comparison on ACDC with 20% labels.

Method	Dice ↑	Jaccard ↑	HD95 ↓	ASD ↓
UA-MT (MICCAI 2019) Yu et al. (2019)	81.11	69.92	7.95	2.23
SASSNet (MICCAI 2020) Li et al. (2020)	86.55	77.90	4.47	1.10
DTC (AAAI 2021) Luo et al. (2021)	87.33	78.52	3.73	1.12
URPC (MedIA 2022) Luo et al. (2022)	86.42	77.61	6.77	1.50
MC-Net (MedIA 2022) Wu et al. (2022a)	88.51	80.19	5.35	1.54
BCP (CVPR 2023) Bai et al. (2023)	85.37	75.47	4.34	1.22
MCF (CVPR 2023) Wang et al. (2023)	82.58	71.92	5.64	1.57
DiffRect (MICCAI 2024) Liu et al. (2024a)	84.31	74.19	5.28	1.57
CDS (EAAI 2025) Hu et al. (2025)	84.47	74.61	3.60	0.97
SBFC (ESWA 2026) Jia et al. (2026)	88.38	79.25	3.31	1.23
CSSML (ESWA 2026) Hang et al. (2026)	80.59	69.10	6.40	1.71
IIIC (TNNLS 2026) Wu et al. (2026)	84.89	74.85	3.38	0.92
Ours	90.12	82.48	1.42	0.37

Table 15: Comparison on ACDC with 10% labels.

Method	Dice ↑	Jaccard ↑	HD95 ↓	ASD ↓
UA-MT (MICCAI 2019) Yu et al. (2019)	78.65	66.97	6.12	1.71
SASSNet (MICCAI 2020) Li et al. (2020)	84.50	74.34	5.42	1.86
DTC (AAAI 2021) Luo et al. (2021)	84.29	73.92	12.81	4.01
URPC (MedIA 2022) Luo et al. (2022)	83.10	72.41	4.84	1.53
MC-Net (MedIA 2022) Wu et al. (2022a)	86.44	77.04	5.50	1.84
BCP (CVPR 2023) Bai et al. (2023)	82.88	72.35	5.35	1.42
DiffRect (MICCAI 2024) Liu et al. (2024a)	84.47	74.31	5.37	1.53
CDS (EAAI 2025) Hu et al. (2025)	80.58	69.60	5.23	1.58
SBFC (ESWA 2026) Jia et al. (2026)	79.43	66.48	5.38	1.27
IIIC (TNNLS 2026) Wu et al. (2026)	84.50	74.30	5.37	1.53
Ours	89.65	81.74	1.47	0.41

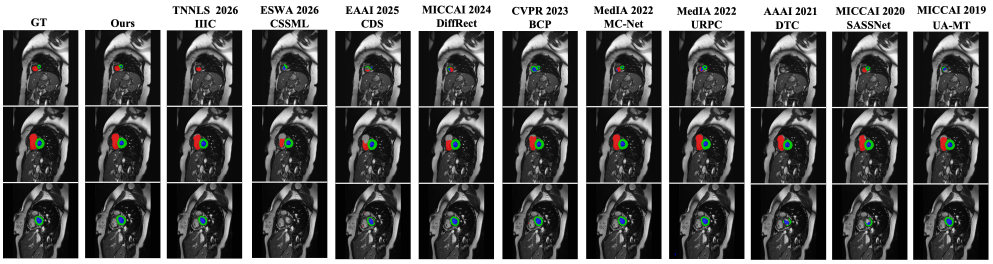

Figure 8: Visualization results of ACDC dataset with 20% labeled data.

B.7 Limitations

While CHCL-Seg achieves consistent improvements across three benchmarks, two practical limitations remain.

Hyperparameter sensitivity. Although the sensitivity analysis (Table 8) shows the model is robust to the loss weights λ_{dacl} and λ_{dot} within a 1000× range, the warm-up schedules and the V-DACL temperature are dataset-dependent and may need re-tuning when transferring to a new modality with substantially different feature statistics.

Boundary metrics under sufficient labels. As reported in the HCL ratio analysis, the gain of HCL on boundary metrics (HD95, ASD) becomes marginal at higher label ratios (e.g., 20%) where boundaries are already well constrained. The benefit of HCL is most pronounced under low-label regimes (5%, 10%).

B.8 Broader Impacts

Positive impacts. Semi-supervised medical image segmentation reduces the need for dense voxel-level annotations, which are expensive and time-consuming to acquire and typically require expert radiologists. By exploiting unlabeled scans more effectively, CHCL-Seg can lower the annotation burden for clinical research and accelerate the development of computer-aided diagnosis tools, particularly in resource-constrained settings where expert annotation is scarce.

Potential risks. As with any learning-based segmentation system, CHCL-Seg may produce inaccurate predictions on out-of-distribution data (e.g., scans from unseen scanners, populations, or pathologies). Pseudo-label-based self-training can amplify systematic biases present in the labeled subset, leading to degraded performance on under-represented patient groups or rare anatomical variants. Direct deployment without expert oversight could therefore lead to erroneous clinical decisions.

Recommended use and mitigation. The proposed method is intended as an assistive tool to support, rather than replace, clinical experts. We recommend (i) deploying the model only after careful validation on the target population and imaging protocol; (ii) reporting uncertainty estimates alongside predicted masks so that low-confidence regions are flagged

for human review; (iii) periodically auditing model behavior on demographically diverse subsets to monitor fairness; and (iv) maintaining expert-in-the-loop verification before any clinically actionable decision is made.