# HCT-NET: COORDINATED HETEROGENEOUS CO-TRAINING FOR SEMI-SUPERVISED 3D MEDICAL IMAGE SEGMENTATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Heterogeneous learners can provide complementary training signals for semi-supervised segmentation, but their relative quality may change throughout training. We study this setting as a coordination problem involving three decisions: which learner provides supervision, when cross-branch learning is enabled, and how strongly each unlabeled sample contributes. We present HCT-Net, a heterogeneous Mamba–Transformer co-training framework in which smoothed labeled-batch performance serves as an online routing proxy. A relative-performance gate controls cross-branch activation, while training ramp-up and confidence coverage determine the strength of pseudo-label supervision. Prediction-level and feature-level transfer follow the same directed route with detached teacher targets. We evaluate HCT-Net on ACDC and LA under 10% and 15% labeled settings. At 15% labeling, HCT-Net achieves Dice scores of 83.01% and 85.00%, respectively, improving over symmetric cross-pseudo supervision using the same Mamba–Transformer pair by 1.09 and 2.20 percentage points. Across the evaluated backbone pairings, the largest observed CPS-to-HCT Dice improvements occur for the heterogeneous Mamba–Transformer pair on both datasets at 15% labeling. Epoch-wise diagnostics further characterize differences in branch diversity and error behavior across HCT configurations.

## 1 INTRODUCTION

Semi-supervised 3D medical image segmentation uses unlabeled images to reduce dependence on dense voxel-wise annotations. Teacher–student methods construct training targets from model predictions, while cross-pseudo supervision (CPS) allows two learners to exchange hard pseudo-labels (Tarvainen & Valpola, 2017; Yu et al., 2019; Chen et al., 2021). The usefulness of these targets depends on their quality: incorrect predictions can reinforce confirmation bias during subsequent updates (Arazo et al., 2020).

Architectural heterogeneity provides different representations for collaborative learning. CNN–Transformer cross teaching and Mamba-based semi-supervised methods explore this direction (Luo et al., 2022; Ma & Wang, 2024; Li et al., 2026). However, architectural differences do not ensure equally reliable training targets at every stage. Heterogeneous branches may evolve at different rates, whereas symmetric CPS maintains both teaching directions without adapting them to relative branch performance. A temporarily weaker branch may therefore provide erroneous targets to its peer.

Performance-based role assignment and adaptive teaching have already been explored in semi-supervised learning. DiCo, for example, dynamically assigns teacher–student roles according to supervised performance on labeled data (Xu et al., 2025). Other approaches regulate pseudo-supervision through uncertainty, teacher diversity, or role exchange (Zhao et al., 2025; Huang et al., 2025; Wu et al., 2025; Lei et al., 2025). These studies suggest that adaptive collaboration is useful, but leave open how several control decisions should be coordinated when two heterogeneous learners evolve at different rates.

We therefore formulate heterogeneous co-training around three coupled questions: *which branch should provide supervision, when should cross-branch learning be active, and how strongly should*



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.

*each unlabeled sample contribute?* HCT-Net addresses these decisions through a common directed policy. An exponential moving average of labeled mini-batch foreground Dice serves as a branch-level routing proxy. Following supervised warm-up, a relative-performance gate controls whether cross-branch supervision is active. A training ramp-up and confidence coverage regulate the magnitude of pseudo-label supervision. The same selected route is also used for projected-feature transfer, so prediction-level and representation-level supervision remain directionally consistent. Teacher targets are detached from cross-branch objectives, while both branches continue to receive ground-truth supervision.

Our contributions are:

- We formulate asynchronous heterogeneous co-training as a coordinated supervision-control problem over routing, activation, and unlabeled-data weighting rather than treating these decisions independently.
- We instantiate this formulation in a Mamba–Transformer framework where pseudo-label supervision and projected-feature transfer follow the same dynamically selected teacher–student route with student-only cross-branch gradients.
- We evaluate the complete strategy on ACDC and LA under 10% and 15% labeled settings, including matched comparisons against symmetric CPS with the same heterogeneous backbone pair and homogeneous backbone controls. At 15% labeling, HCT-Net improves Dice over heterogeneous CPS by 1.09 and 2.20 percentage points on ACDC and LA, respectively, while controlled homogeneous-backbone experiments and training dynamics further characterize the behavior of the complete HCT strategy across backbone configurations.

## 2 Related Work

**Semi-supervised medical image segmentation.** Consistency learning and pseudo-labeling provide common approaches to exploiting unlabeled images. Representative methods include Mean Teacher, uncertainty-aware consistency, geometric regularization, and cross-pseudo supervision (Tarvainen & Valpola, 2017; Yu et al., 2019; Li et al., 2020; Luo et al., 2021; Chen et al., 2021). GuidedNet further uses labeled information to guide unlabeled segmentation (Zhao et al., 2024). HCT-Net studies the regulation of supervision exchange between heterogeneous learners.

**Heterogeneous collaborative learning.** CNN–Transformer cross teaching exchanges predictions between architecturally different networks (Luo et al., 2022). Semi-Mamba-UNet and DCMamba explore Mamba-based collaborative segmentation (Ma & Wang, 2024; Li et al., 2026). These studies motivate architectural diversity as a source of different learning signals. HCT-Net adopts a Mamba–Transformer pair and coordinates the exchange of both prediction and projected-feature targets. Its matched CPS comparison evaluates the complete training strategy while keeping the Mamba–Transformer backbone pair fixed.

**Adaptive supervision and closest collaborative methods.** Confidence, uncertainty, teacher diversity, and role exchange have been used to regulate pseudo-supervision (Sohn et al., 2020; Zhao et al., 2025; Huang et al., 2025; Wu et al., 2025; Lei et al., 2025). DiCo is particularly relevant because it dynamically assigns teacher–student roles according to supervised performance on labeled data (Xu et al., 2025). Semi-Mamba-UNet and DCMamba further demonstrate collaborative learning with Mamba-based segmentation models (Ma & Wang, 2024; Li et al., 2026).

The distinguishing design of HCT-Net is the joint coordination of these decisions within a single directed policy for asynchronously evolving heterogeneous learners. Unlike unconditional bidirectional cross-teaching, the supervision route is determined online from smoothed labeled performance, while a relative-performance gate determines whether cross-branch learning is activated and confidence coverage determines how strongly each unlabeled sample contributes. Prediction-level supervision and projected-feature transfer then follow the same selected direction. This formulation couples supervision direction, activation, and magnitude rather than treating them as separate training heuristics.

# 3 METHOD

## 3.1 PROBLEM FORMULATION AND FRAMEWORK OVERVIEW

We consider a labeled training set and an unlabeled training set,

$$\mathcal{D}_l = \{(\mathbf{x}_i^l, \mathbf{y}_i^l)\}_{i=1}^{N_l}, \qquad \mathcal{D}_u = \{\mathbf{x}_j^u\}_{j=1}^{N_u}, \tag{1}$$

where $N_l \ll N_u$. Each input is a 3D image $\mathbf{x} \in \mathbb{R}^{H\times W\times D}$, and $\mathbf{y} \in \{0, \dots, C-1\}^{H\times W\times D}$ is its voxel-wise annotation.

HCT-Net comprises two segmentation networks with separate parameters: a Mamba-based branch $A$ and a Transformer-based branch $B$. For $k \in \{A, B\}$ and $q \in \{l, u\}$, the branch produces

$$\mathbf{F}_k^q = E_k(\mathbf{x}^q), \qquad \mathbf{p}_k^q = \mathrm{softmax}(D_k(\mathbf{F}_k^q)), \tag{2}$$

where $E_k$ and $D_k$ denote the encoder and decoder, respectively. We write $f_k(\mathbf{x})$ for the resulting class-probability map.

Both networks learn from labeled data throughout training. For unlabeled data, a controller selects one teacher-to-student route for each mini-batch, gates cross-branch learning, and weights the pseudo-label loss of each unlabeled sample. Teacher and student are temporary roles, denoted by $T$ and $S$, rather than fixed network identities.

## 3.2 SUPERVISED LEARNING AND WARM-UP

Each branch is trained with the supervised objective

$$\begin{aligned}\mathcal{L}_{\mathrm{sup}}^k = \mathcal{L}_{\mathrm{Dice}}(\mathbf{p}_k^l, \mathbf{y}^l) + \mathcal{L}_{\mathrm{CE}}(\mathbf{p}_k^l, \mathbf{y}^l) \\ + \lambda_{\mathrm{hd}}\mathcal{L}_{\mathrm{HD}}(\mathbf{p}_k^l, \mathbf{y}^l),\end{aligned} \tag{3}$$

where $\mathcal{L}_{\mathrm{Dice}}$ and $\mathcal{L}_{\mathrm{CE}}$ are the Dice and cross-entropy losses, and $\mathcal{L}_{\mathrm{HD}}$ is a Hausdorff-distance surrogate for boundary supervision (Karimi & Salcudean, 2020). The two branch losses are summed:

$$\mathcal{L}_{\mathrm{sup}} = \mathcal{L}_{\mathrm{sup}}^A + \mathcal{L}_{\mathrm{sup}}^B. \tag{4}$$

During the first $T_{\mathrm{warm}}$ iterations, only $\mathcal{L}_{\mathrm{sup}}$ is used for optimization. This provides an initial supervised training period before predictions and features are exchanged between the branches.

## 3.3 ONLINE RELIABILITY ESTIMATION AND TEACHER ROUTING

At iteration $t$, we evaluate each branch on the labeled training mini-batch. Its mean foreground Dice is

$$d_k^{(t)} = \frac{1}{B_l}\sum_{i=1}^{B_l} \mathrm{DSC}_{\mathrm{fg}}\left(\arg\max_c \mathbf{p}_{k,i,c}^l, \mathbf{y}_i^l\right), \tag{5}$$

where $B_l$ is the labeled batch size and $\mathrm{DSC}_{\mathrm{fg}}$ averages Dice over foreground classes for multi-class segmentation. These labeled training scores serve as an online proxy for relative branch reliability.

We smooth mini-batch fluctuations using an exponential moving average:

$$R_k^{(t)} = \beta R_k^{(t-1)} + (1-\beta)d_k^{(t)}, \qquad k \in \{A, B\}, \tag{6}$$

where $\beta \in [0, 1)$ and the initial score is set to the first available labeled mini-batch Dice. The moving average is applied to performance scores, not to network parameters.

After updating the scores, the teacher is selected as

$$T^{(t)} = \arg\max_{k\in\{A,B\}} R_k^{(t)}, \tag{7}$$

and the other branch becomes $S^{(t)}$. The selected route is shared by all unlabeled samples in that mini-batch and can change as the score ranking changes. We omit the iteration superscript on $T$ and $S$ where the context is clear.

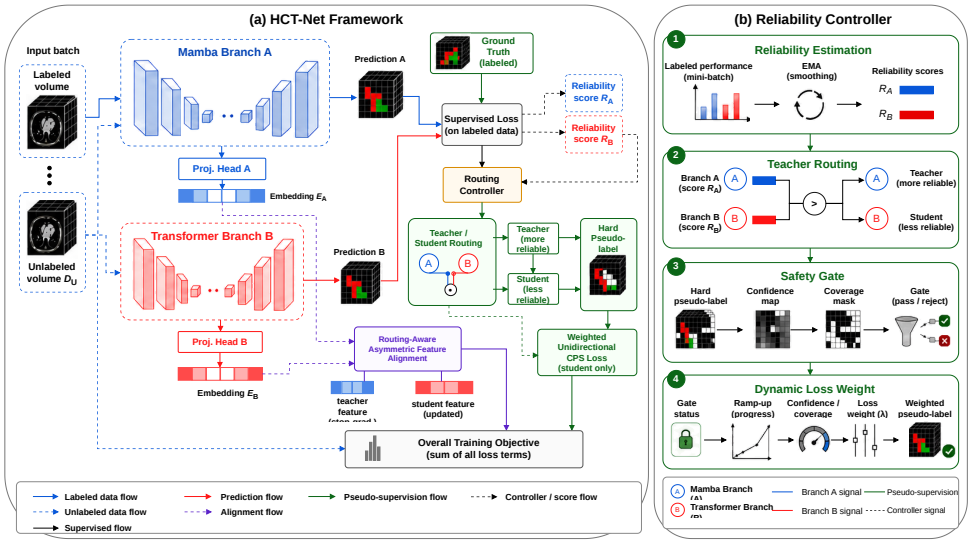



Figure 1: Overview of HCT-Net. (a) Mamba- and Transformer-based branches receive ground-truth supervision, while the selected student additionally learns from the teacher's detached pseudo-labels and projected features. (b) The reliability controller combines online reliability estimation, teacher routing, safety gating, and dynamic loss weighting. The safety gate uses warm-up status and labeled mini-batch performance, whereas confidence coverage contributes to sample-level pseudo-supervision weighting. Panel (b) schematically summarizes these control functions; their computational dependencies and the meanings of "coverage mask" and "weighted pseudo-label" are specified in Section 3.4.

### 3.4 Performance-Gated Asymmetric Pseudo-Supervision

In addition to comparing the branches, we compare the selected teacher's current performance with its smoothed score:

$$g^{(t)} = \mathbf{1}[t > T_{\text{warm}}]\,\mathbf{1}[d_T^{(t)} > \alpha R_T^{(t)}], \tag{8}$$

Here, $\mathbf{1}[\cdot]$ is the indicator function and $\alpha$ sets a relative-performance threshold. Cross-branch learning is enabled only after warm-up and when the selected teacher's current labeled Dice exceeds a fraction of its smoothed score. The "safety gate" in Figure 1 denotes a labeled-performance check rather than a voxel-wise confidence filter or a guarantee of pseudo-label correctness.

The selected teacher's prediction for an unlabeled sample is detached:

$$\overline{\mathbf{p}}_{T,i}^{u} = \text{sg}(\mathbf{p}_{T,i}^{u}), \tag{9}$$

where sg denotes stop-gradient. The hard target is

$$\widehat{\mathbf{y}}_{T,i}^{u}(v) = \arg\max_{c} \overline{\mathbf{p}}_{T,i}^{u}(v,c), \tag{10}$$

with $v \in \Omega$ indexing voxels. We define confidence coverage as

$$\rho_i^{(t)} = \frac{1}{|\Omega|} \sum_{v \in \Omega} \mathbf{1}\left[\max_{c} \overline{\mathbf{p}}_{T,i}^{u}(v,c) > \tau\right], \tag{11}$$

where $\tau$ is a confidence threshold. Both the hard pseudo-label and the confidence map are derived from the teacher's class-probability map: the most probable class defines the hard target, and its probability defines the voxel-wise confidence. The "coverage mask" in Figure 1 denotes the binary threshold indicator in equation 11. Its spatial average yields $\rho_i^{(t)}$, the fraction of voxels exceeding the confidence threshold. This scalar adjusts the contribution of the corresponding unlabeled training sample; the binary mask is not used to exclude individual voxels from the pseudo-supervision loss.

The weight assigned to unlabeled sample $i$ is

$$w_i^{(t)} = w_{\max} g^{(t)} r(t) \rho_i^{(t)}, \tag{12}$$

where $w_{\max}$ is the maximum pseudo-supervision weight and $r(t) \in [0, 1]$ is a nondecreasing training ramp-up coefficient. The gate determines whether pseudo-supervision is enabled, the ramp-up coefficient controls its scheduled scale, and confidence coverage adjusts each sample's contribution. The loss weight illustrated in panel (b) of Figure 1 corresponds to $w_i^{(t)}$. The term "weighted pseudo-label" is shorthand for pseudo-supervision with this loss weight: the hard target retains its categorical values, while its associated cross-entropy loss is multiplied by $w_i^{(t)}$. The resulting loss is

$$\mathcal{L}_{\text{cps}}^{(t)} = \frac{1}{B_u} \sum_{i=1}^{B_u} w_i^{(t)} \mathcal{L}_{\text{CE}}\big(\mathbf{p}_{S,i}^u, \widehat{\mathbf{y}}_{T,i}^u\big)\,, \tag{13}$$

where $B_u$ is the unlabeled batch size. Cross-entropy is evaluated over the full voxel domain; low-confidence voxels are not individually excluded. The batch reduction divides by $B_u$, not by the sum of sample weights. For each routed mini-batch, this loss supplies gradients only to the selected student.

### 3.5 Routing-Aware Asymmetric Feature Alignment

To transfer intermediate representations along the same route, we map features from the two branches into a common latent space:

$$\mathbf{Z}_k^u = h_k(\mathbf{F}_k^u), \qquad k \in \{A, B\}, \tag{14}$$

where $h_A$ and $h_B$ are lightweight projection heads and their outputs have matching dimensions. The feature loss is

$$\mathcal{L}_{\text{feat}}^{(t)} = \frac{g^{(t)}}{B_u} \sum_{i=1}^{B_u} \frac{\big\|\mathbf{Z}_{S,i}^u - \text{sg}(\mathbf{Z}_{T,i}^u)\big\|_2^2}{|\mathbf{Z}_{S,i}^u|}, \tag{15}$$

where $|\mathbf{Z}_{S,i}^u|$ is the number of elements in the projected tensor. Alignment therefore minimizes the element-wise mean squared difference in the projected space.

Detaching the teacher target blocks this loss from updating the teacher branch or its projection head; gradients propagate through the student's projected features. Unlike the pseudo-label loss, the feature loss is gated by $g^{(t)}$ but is not multiplied by $r(t)$ or $\rho_i^{(t)}$.

### 3.6 Overall Objective and Optimization

The complete objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{cps}} + \lambda_{\text{feat}} \mathcal{L}_{\text{feat}}, \tag{16}$$

where $\lambda_{\text{feat}}$ independently controls feature alignment.

At each iteration, we compute labeled predictions, update the reliability scores, select the teacher–student route, and evaluate the gate. When $g^{(t)} = 0$, both cross-branch losses vanish and training uses only ground-truth supervision. When $g^{(t)} = 1$, the student additionally learns from the teacher's detached pseudo-labels and projected features.

Student-only cross-branch supervision does not freeze the teacher: both segmentation branches continue to receive their supervised gradients. The teacher–student assignment controls only the direction of the cross-branch losses for the current mini-batch.

### 3.7 Inference

The training-time reliability controller and projection heads are disabled during inference. For each dataset and annotation setting, the prediction rule is selected using validation data and then fixed before test evaluation. Depending on the validated configuration, predictions are obtained either from a selected branch or from a fixed convex combination of the two branch probabilities,

$$\mathbf{p}_{\text{ens}} = \gamma \mathbf{p}_A + (1 - \gamma)\mathbf{p}_B, \tag{17}$$

where $\gamma$ is determined on the validation split. Test-time augmentation and connected-component post-processing, when used, are also fixed before test evaluation. All inference choices, including the fusion coefficient, TTA, and post-processing configuration, are selected using validation data and fixed before evaluation on the test set.

# 4 EXPERIMENTS

## 4.1 EXPERIMENTAL SETUP

**Datasets and annotation settings.** We evaluate HCT-Net on the Automated Cardiac Diagnosis Challenge (ACDC) dataset (Bernard et al., 2018) and the 2018 Left Atrium Segmentation Challenge (LA) dataset (Xiong et al., 2021). ACDC involves multi-class segmentation of the right ventricle, myocardium, and left ventricle, whereas LA involves binary left-atrium segmentation. We consider 10% and 15% labeled training settings. Within each setting, the labeled case list is fixed, and the remaining training volumes are treated as unlabeled.

**Evaluation metrics.** We report Dice, Jaccard, the 95th-percentile Hausdorff distance (HD95), and average surface distance (ASD). Higher Dice and Jaccard indicate better overlap, whereas lower HD95 and ASD indicate smaller surface distances. HD95 and ASD are computed on the preprocessed 3D volumes and are reported in voxel units. Tables 1 and 2 report the mean and standard deviation across test volumes under the same evaluation protocol.

**Comparison methods and protocol.** We compare with UA-MT (Yu et al., 2019), SASS-Net (Li et al., 2020), DTC (Luo et al., 2021), CNN–Transformer cross teaching (Luo et al., 2022), GuidedNet (Zhao et al., 2024), and Semi-Mamba-UNet (Ma & Wang, 2024). The comparison uses a 3D adaptation of Semi-Mamba-UNet. The methods share fixed dataset partitions and labeled case lists. Preprocessing, training budgets, sliding-window inference, and metric computation follow a common protocol wherever applicable.

**Implementation details.** HCT-Net is implemented in PyTorch with SegMamba and SwinUNETR branches and a crop size of $32 \times 160 \times 160$. During training, each network input is a sampled 3D crop. Routing scores and confidence coverage are computed on the corresponding training crops, whereas evaluation metrics are computed on reconstructed 3D volumes. For ACDC, we use $(\beta, \alpha, \tau, w_{\max}) = (0.9, 0.75, 0.80, 2.0)$; for LA, the corresponding values are $(0.9, 0.85, 0.80, 2.5)$. We set $\lambda_{\mathrm{hd}} = \lambda_{\mathrm{feat}} = 0.1$ for both datasets. All model-selection and inference choices are fixed using training and validation data before test evaluation.

## 4.2 COMPARISON WITH EXISTING METHODS

**Results on ACDC.** At 15% labeling, HCT-Net achieves the best mean result across all four metrics in Table 1, including a Dice of $0.8301$, 1.93 percentage points above DTC. At 10%, HCT-Net obtains $0.7773$ Dice versus $0.7770$ for CNN–Transformer, while DTC remains stronger on Jaccard and the surface-distance metrics.

**Results on LA.** At 15% labeling, HCT-Net achieves the best mean result across all four metrics in Table 2, with $0.8500$ Dice, 3.68 percentage points above GuidedNet. At 10%, HCT-Net obtains the best Dice ($0.8121$), Jaccard ($0.7071$), and ASD ($3.9953$), while CNN–Transformer achieves the lowest HD95.

## 4.3 QUALITATIVE RESULTS

Figure 2 provides representative 3D segmentation results on ACDC and LA. On ACDC, HCT-Net produces more continuous ventricular structures and contours that more closely follow the reference annotation than several comparison methods. On LA, HCT-Net better preserves the global atrial structure while reducing conspicuous detached regions. These visual observations are consistent with the overlap and surface-distance improvements reported in Tables 1 and 2.

## 4.4 COMPARISON OF TRAINING STRATEGIES AND ARCHITECTURES

Table 3 provides a controlled comparison of training strategy and backbone pairing at 15% labeling. The most direct comparison keeps the Mamba–Transformer pair fixed and replaces symmetric CPS with the complete HCT policy. Under this matched architecture, Dice increases from $0.8192$ to $0.8301$ on ACDC and from $0.8280$ to $0.8500$ on LA, together with improvements in the reported

Table 1: Quantitative comparison on ACDC under 15% and 10% labeled settings. Values are reported as mean ± standard deviation over test volumes. Best mean values in each setting are shown in bold.

| | 15% labeled | | | | 10% labeled | | | |
|---|---|---|---|---|---|---|---|---|
| Method | Dice ↑ | HD95 ↓ | Jaccard ↑ | ASD ↓ | Dice ↑ | HD95 ↓ | Jaccard ↑ | ASD ↓ |
| UA-MT | 0.7927±0.1631 | 5.1075±9.2402 | 0.6821±0.1686 | 1.8757±6.7752 | 0.7382±0.2128 | 6.5750±9.4699 | 0.6219±0.2104 | 2.1413±5.4748 |
| SASS-Net | 0.7986±0.0819 | 9.2019±6.6260 | 0.6797±0.1014 | 3.0173±2.5415 | 0.7718±0.1255 | 10.9763±11.0913 | 0.6507±0.1383 | 3.5316±4.4849 |
| DTC | 0.8108±0.1202 | 4.5300±10.0400 | 0.7035±0.1257 | 1.9900±9.8800 | 0.7693±0.1424 | **5.7617±8.6549** | **0.6578±0.1418** | **1.8607±6.5991** |
| CNN-Trans | 0.7924±0.1392 | 7.0916±10.8945 | 0.6753±0.1686 | 2.2607±3.3179 | 0.7770±0.1426 | 7.5160±11.4062 | 0.6552±0.1730 | 2.5173±4.4520 |
| GuidedNet | 0.7923±0.1119 | 7.2581±9.8012 | 0.6743±0.1329 | 2.5667±3.8108 | 0.7531±0.1498 | 8.4123±11.9029 | 0.6298±0.1578 | 3.3473±6.0477 |
| Semi-Mamba-UNet | 0.7227±0.1841 | 9.3932±9.5698 | 0.6061±0.2065 | 3.1220±3.9304 | 0.7093±0.1981 | 10.6618±13.5619 | 0.5943±0.2140 | 3.8362±6.8962 |
| HCT-Net | **0.8301±0.0997** | **4.1835±4.8556** | **0.7270±0.1194** | **1.1185±1.4227** | **0.7773±0.1176** | 10.2542±9.7962 | 0.6570±0.1328 | 3.2788±3.8393 |

Table 2: Quantitative comparison on LA under 15% and 10% labeled settings. Values are reported as mean ± standard deviation over test volumes. Best mean values in each setting are shown in bold.

| | 15% labeled | | | | 10% labeled | | | |
|---|---|---|---|---|---|---|---|---|
| Method | Dice ↑ | HD95 ↓ | Jaccard ↑ | ASD ↓ | Dice ↑ | HD95 ↓ | Jaccard ↑ | ASD ↓ |
| UA-MT | 0.7619±0.2057 | 16.2575±12.8599 | 0.6507±0.2160 | 5.9253±8.0079 | 0.7613±0.1758 | 23.3994±14.1921 | 0.6413±0.1911 | 8.4134±6.7821 |
| SASS-Net | 0.7926±0.1888 | 19.6709±16.0040 | 0.6868±0.1986 | 7.2765±8.6740 | 0.7868±0.1841 | 17.9423±13.5855 | 0.6757±0.1830 | 5.5955±7.9648 |
| DTC | 0.8051±0.1433 | 16.9000±11.7300 | 0.6931±0.1627 | 5.5700±4.9700 | 0.7946±0.1464 | 17.9800±13.3200 | 0.6792±0.1666 | 6.3100±5.7500 |
| CNN-Trans | 0.8091±0.1376 | 15.1377±12.2418 | 0.6979±0.1615 | 5.3497±6.0017 | 0.8008±0.1458 | **14.9045±11.5884** | 0.6872±0.1618 | 4.9608±5.5334 |
| GuidedNet | 0.8132±0.1364 | 16.2302±12.3818 | 0.7022±0.1500 | 4.7802±6.0876 | 0.7973±0.1350 | 18.2130±12.9553 | 0.6793±0.1481 | 5.1231±6.2088 |
| Semi-Mamba-UNet | 0.8088±0.1514 | 18.1559±15.0069 | 0.7002±0.1732 | 6.1589±6.2951 | 0.7935±0.1393 | 16.2677±11.9292 | 0.6750±0.1542 | 5.5799±4.9903 |
| HCT-Net | **0.8500±0.1529** | **14.2280±12.9333** | **0.7610±0.1668** | **3.3450±4.7699** | **0.8121±0.1631** | 16.5935±12.4186 | **0.7071±0.1733** | **3.9953±4.4322** |

overlap and surface metrics. This matched comparison keeps the heterogeneous backbone pair fixed and compares the complete HCT and symmetric-CPS training configurations under the same architectural setting.

On ACDC, homogeneous Mamba performs worse with HCT than with CPS ($0.7871$ versus $0.8219$), while homogeneous Swin improves. Both homogeneous pairs improve in Dice on LA, although homogeneous Mamba has a lower Jaccard with HCT. These results show that the outcome depends on the backbone configuration. Across the evaluated pairings, the largest observed CPS-to-HCT Dice improvements occur for the heterogeneous Mamba–Transformer pair on both datasets.

### 4.5 Descriptive Training Dynamics

Figure 3 reports epoch-wise diagnostics from matched 15% labeled control runs. On ACDC, Hetero-HCT maintains larger foreground prediction disagreement than the homogeneous controls while exhibiting lower shared foreground error overlap over much of training. On LA, Hetero-HCT likewise maintains substantially larger foreground disagreement; its pseudo-label Dice-error trajectory remains below Homo-Swin and close to Homo-Mamba.

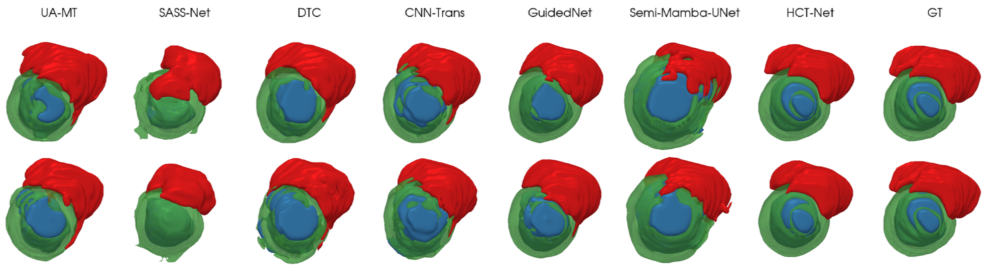
Foreground disagreement is defined as the fraction of disagreeing branch predictions within the union of voxels predicted as foreground by either branch. For ACDC, shared foreground error overlap is the intersection-over-union of the two branch error sets restricted to ground-truth foreground voxels. For LA, disagreement is evaluated on the validation split, whereas pseudo-label Dice error is computed as $1 - \mathrm{Dice}(\hat{\mathbf{y}}_T^u, \mathbf{y}^u)$ over active pseudo-supervision mini-batches. The reference masks used for these diagnostics do not enter the controller, loss, or parameter updates.

Together with Table 3, these trajectories characterize the training behavior of the complete heterogeneous HCT configuration across optimization.

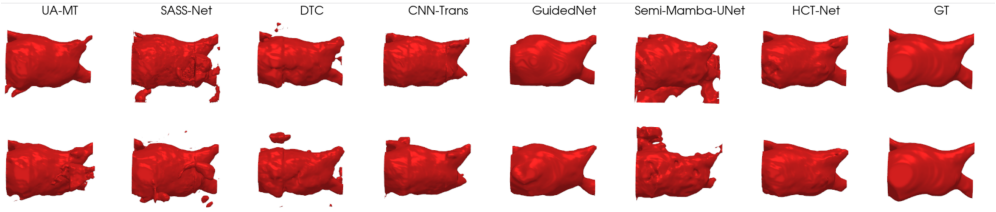
### 4.6 Parameter Count Analysis

The heterogeneous design introduces additional training-time capacity relative to a single branch. A SwinUNETR branch contains approximately $14.44$M parameters and a SegMamba branch approximately $40.95$M parameters, giving HCT-Net approximately $55.40$M parameters during dual-branch training. For comparison, the Homo-Swin and Homo-Mamba controls contain approximately $28.89$M and $81.91$M parameters, respectively.

The reported parameter counts include the projection heads used during training. Since the larger Homo-Mamba control does not outperform HCT-Net on either dataset, parameter count alone does not explain the observed performance differences.




(a) ACDC




(b) LA

Figure 2: Representative qualitative comparisons on ACDC and LA. (a) ACDC multi-class segmentation of the right ventricle, myocardium, and left ventricle. (b) LA segmentation. GT denotes the reference annotation. HCT-Net shows improved structural continuity and fewer conspicuous segmentation artifacts in the displayed cases, consistent with the quantitative results in Tables 1 and 2.

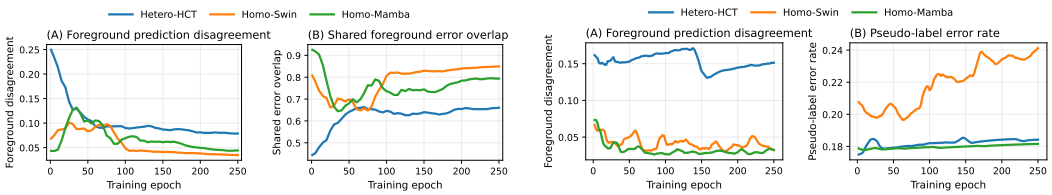
Table 3: Comparison of training strategies and backbone pairs on ACDC and LA with 15% labeled data. Mean values are reported. CPS denotes symmetric cross-pseudo supervision, and HCT denotes the complete proposed training strategy. Best values for each dataset are shown in bold.

| Dataset | Variant | Dice ↑ | HD95 ↓ | Jaccard ↑ | ASD ↓ |
|---|---|---|---|---|---|
| ACDC | HCT w/ Homo-Mamba | 0.7871 | 5.1471 | 0.6682 | 1.4423 |
| | HCT w/ Homo-Swin | 0.8057 | 4.6817 | 0.6895 | 1.1837 |
| | CPS w/ Homo-Mamba | 0.8219 | 5.2125 | 0.7175 | 1.3696 |
| | CPS w/ Homo-Swin | 0.7976 | 5.5067 | 0.6849 | 1.5555 |
| | CPS w/ Mamba-Transformer | 0.8192 | 4.9802 | 0.7102 | 1.3091 |
| | HCT-Net | **0.8301** | **4.1835** | **0.7270** | **1.1185** |
| LA | HCT w/ Homo-Mamba | 0.8289 | 15.9300 | 0.7294 | 4.1000 |
| | HCT w/ Homo-Swin | 0.8086 | 15.8200 | 0.7011 | 4.3500 |
| | CPS w/ Homo-Mamba | 0.8277 | 17.2000 | 0.7322 | 4.2900 |
| | CPS w/ Homo-Swin | 0.8013 | 19.0100 | 0.6924 | 4.9400 |
| | CPS w/ Mamba-Transformer | 0.8280 | 16.9236 | 0.7293 | 4.8118 |
| | HCT-Net | **0.8500** | **14.2280** | **0.7610** | **3.3450** |

### 4.7 Discussion

The experiments consistently indicate that architectural diversity and supervision coordination should be considered jointly. When the Mamba–Transformer architecture is held fixed, replacing symmetric CPS with HCT improves all four reported metrics at the 15% labeled setting on both ACDC and LA. The homogeneous controls further show that simply increasing model capacity or applying the same controller to identical backbones does not reproduce the heterogeneous result. The training trajectories provide complementary evidence that HCT-Net maintains larger inter-branch diversity and, on ACDC, substantially lower shared foreground error overlap.

The 10% ACDC setting also illustrates that improvements need not be uniform across every metric: its Dice is competitive, whereas DTC obtains stronger surface-distance results. This suggests that the usefulness of cross-branch coordination depends on the quality and complementarity of the available supervision signals.




(a) ACDC (b) LA

Figure 3: Training dynamics under the 15% labeled setting. (a) ACDC: foreground prediction disagreement and shared foreground error overlap. (b) LA: foreground prediction disagreement and pseudo-label Dice error, computed as one minus pseudo-label Dice. Homo-Swin and Homo-Mamba denote homogeneous-backbone HCT controls. Reference labels used for the error diagnostics are used only for analysis and do not enter model optimization.

The labeled-batch Dice is used as a relative routing proxy rather than as a calibrated estimate of pseudo-label correctness. The current evaluation focuses on two 3D cardiac MRI benchmarks; extending HCT-Net to broader anatomical and modality settings is a natural direction for future work.

## 5 Conclusion

We presented HCT-Net, a coordinated heterogeneous co-training framework for semi-supervised 3D medical image segmentation. The framework jointly regulates which branch provides supervision, when cross-branch learning is activated, and how strongly each unlabeled sample contributes, while prediction- and feature-level transfer follow the same directed route.

With the Mamba–Transformer backbone pair held fixed, HCT-Net improves Dice over symmetric CPS by 1.09 percentage points on ACDC and 2.20 points on LA at the 15% labeled setting, together with consistent gains in the reported overlap and surface metrics. Across the evaluated backbone pairings, the heterogeneous Mamba–Transformer configuration exhibits the largest observed CPS-to-HCT Dice improvements at 15% labeling. Together with the training diagnostics, these results support the complete coordinated HCT configuration as an effective strategy for exploiting complementary heterogeneous learners in label-scarce 3D segmentation.

### AI use statement

In this work, generative AI tools were used selectively to assist with research ideation, literature retrieval and discovery, drafting limited portions of the manuscript, and language polishing. The overall research methodology, experimental design, implementation, data analysis, and scientific conclusions were developed and validated by the authors. All AI-assisted content was critically reviewed, revised where necessary, and independently verified against the relevant experiments and original sources. The authors take full responsibility for the final content of this paper.

### Reproducibility statement

Sections 3 and 4 specify the method formulation, datasets, annotation settings, evaluation metrics, implementation details, and inference protocol. All model-selection and inference choices are fixed using training and validation data, with the test set reserved for final evaluation. Dataset partitions, preprocessing, and model configurations are fixed within each annotation setting.

## References

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, July 2020. doi: 10.1109/IJCNN48605.2020.9207304. URL https://ieeexplore.ieee.org/document/9207304. ISSN: 2161-4407.

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 Rohe, Xavier Pennec, Maxime Sermesant, Fabian Isensee, Paul Jager, Klaus H. Maier-Hein, Peter M. Full, Ivo Wolf, Sandy Engelhardt, Christian F. Baumgartner, Lisa M. Koch, Jelmer M. Wolterink, Ivana Isgum, 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, November 2018. ISSN 1558-254X. doi: 10.1109/TMI.2018.2837502.

Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-Supervised Semantic Segmentation With Cross Pseudo Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2613–2622, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Semi-Supervised_Semantic_Segmentation_With_Cross_Pseudo_Supervision_CVPR_2021_paper.html.

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. ISSN 1941-0042. doi: 10.1109/TIP.2025.3599783.

Davood Karimi and Septimiu E. Salcudean. Reducing the Hausdorff Distance in Medical Image Segmentation With Convolutional Neural Networks. *IEEE transactions on medical imaging*, 39 (2):499–513, February 2020. ISSN 1558-254X. doi: 10.1109/TMI.2019.2930068.

Tao Lei, Ziyao Yang, Xingwu Wang, Yi Wang, Xuan Wang, Feiman Sun, and Asoke K. Nandi. Adaptive learning of high-value regions for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21450–21459, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Lei_Adaptive_Learning_of_High-Value_Regions_for_Semi-Supervised_Medical_Image_Segmentation_ICCV_2025_paper.html.

Shuailin Li, Chuyu Zhang, and Xuming He. Shape-Aware Semi-supervised 3D Semantic Segmentation for Medical Images. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pp. 552–561, Cham, 2020. Springer International Publishing. ISBN 978-3-030-59710-8. doi: 10.1007/978-3-030-59710-8_54.

Shumeng Li, Jian Zhang, Lei Qi, Luping Zhou, Yinghuan Shi, and Yang Gao. Diversity-Enhanced Collaborative Mamba for Semi-Supervised Medical Image Segmentation. *IEEE transactions on medical imaging*, 45(2):500–514, February 2026. ISSN 1558-254X. doi: 10.1109/TMI.2025.3601450.

Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised Medical Image Segmentation through Dual-task Consistency. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):8801–8809, May 2021. ISSN 2374-3468. doi: 10.1609/aaai.v35i10.17066. URL https://ojs.aaai.org/index.php/AAAI/article/view/17066.

Xiangde Luo, Minhao Hu, Tao Song, Guotai Wang, and Shaoting Zhang. Semi-Supervised Medical Image Segmentation via Cross Teaching between CNN and Transformer. In *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*, pp. 820–833. PMLR, December 2022. URL https://proceedings.mlr.press/v172/luo22b.html.

Chao Ma and Ziyang Wang. Semi-Mamba-UNet: Pixel-level contrastive and cross-supervised visual Mamba-based UNet for semi-supervised medical image segmentation. *Knowledge-Based Systems*, 300:112203, September 2024. ISSN 0950-7051. doi: 10.1016/j.knosys.2024.112203. URL https://www.sciencedirect.com/science/article/pii/S0950705124008372.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin Dogus Cubuk, Alexey Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying semi-supervised learning

with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33, pp. 596–608, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/06964dce9addb1c5cb5d6e3d9838f733-Abstract.html.

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*, 30:1195–1204, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/68053af2923e00204c3ca7c6a3150cf7-Abstract.html.

Yonghuang Wu, Guoqing Wu, Jixian Lin, Yuanyuan Wang, and Jinhua Yu. Role exchange-based self-training semi-supervision framework for complex medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 36(5):8372–8386, 2025. doi: 10.1109/TNNLS.2024.3432877. URL https://ieeexplore.ieee.org/document/10620642/.

Zhaohan Xiong, Qing Xia, Zhiqiang Hu, Ning Huang, Cheng Bian, Yefeng Zheng, Sulaiman Vesal, Nishant Ravikumar, Andreas Maier, Xin Yang, Pheng-Ann Heng, Dong Ni, Caizi Li, Qianqian Tong, Weixin Si, Elodie Puybareau, Younes Khoudli, Thierry Géraud, Chen Chen, Wenjia Bai, Daniel Rueckert, Lingchao Xu, Xiahai Zhuang, Xinzhe Luo, Shuman Jia, Maxime Sermesant, Yashu Liu, Kuanquan Wang, Davide Borra, Alessandro Masci, Cristiana Corsi, Coen de Vente, Mitko Veta, Rashed Karim, Chandrakanth Jayachandran Preetha, Sandy Engelhardt, Menyun Qiao, Yuanyuan Wang, Qian Tao, Marta Nuñez-Garcia, Oscar Camara, Nicolo Savioli, Pablo Lamata, and Jichao Zhao. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical Image Analysis*, 67: 101832, January 2021. ISSN 1361-8423. doi: 10.1016/j.media.2020.101832.

Jiao Xu, Xin Chen, and Lihe Zhang. Learning dynamic collaborative network for semi-supervised 3D vessel segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10445–10454, 2025.

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 Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pp. 605–613, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32245-8. doi: 10.1007/978-3-030-32245-8_67.

Haochen Zhao, Hui Meng, Deqian Yang, Xiaozheng Xie, Xiaoze Wu, Qingfeng Li, and Jianwei Niu. GuidedNet: Semi-Supervised Multi-Organ Segmentation via Labeled Data Guide Unlabeled Data. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 886–895, New York, NY, USA, October 2024. Association for Computing Machinery. ISBN 979-8-4007-0686-8. doi: 10.1145/3664647.3681526. URL https://dl.acm.org/doi/10.1145/3664647.3681526.

Zhen Zhao, Zicheng Wang, Longyue Wang, Dian Yu, Yixuan Yuan, and Luping Zhou. Alternate Diverse Teaching for Semi-supervised Medical Image Segmentation. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 227–243, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72652-1. doi: 10.1007/978-3-031-72652-1_14.