Semi-supervised Cardiac Cine-MRI Multi-Structure Segmentation via Temporal Consistency on Unlabeled Cine Frames

Yali Wan[1,*], Deyi Xiong[2]
[1]*School of Information Technology and Engineering, Guangzhou College of Commerce, Guangzhou, China*
[2]*Chongqing Academy of Metrology and Quality Inspection, Chongqing, China*
**Corresponding author*

Abstract: The Automated Cardiac Diagnosis Challenge (ACDC) supplies manual annotations for only two frames of each cine-MRI sequence (end-diastole and end-systole), although every sequence contains roughly 30 frames. We present TC-SS, a mean-teacher semi-supervised framework that puts the remaining unlabeled frames to work through three complementary objectives: temporal coherence between adjacent frames, agreement with the temporal mean prediction, and pseudo-label consistency between frames of similar appearance. The supervised branch keeps the compound boundary-aware loss of our previous work and is applied only to the annotated ED/ES frames. Under 3-fold cross-validation on ACDC, evaluated on the unified per-slice ED+ES protocol shared across the three technical papers of this series, TC-SS reaches a mean foreground Dice of 0.713 ± 0.011 against the supervised HybridUNet baseline (fold-0 mDice 0.863, an absolute gap of 0.150), while using 15 times as many frames per patient at no annotation cost. On a per-class basis TC-SS reaches RV Dice 0.669 ± 0.020, MYO Dice 0.686 ± 0.008 and LV Dice 0.784 ± 0.017, all below the supervised baseline (0.818, 0.866 and 0.905 for RV, MYO and LV respectively); the myocardium shows the most consistent gap across folds, which is consistent with the global temporal-consistency terms over-regularising the thin myocardial band. The widened gap relative to earlier patient-level ED-only reports reflects the unified evaluation on ED+ES slices: the ES frame is intrinsically harder than ED and pulls both numbers down, more so for TC-SS.

Keywords: Cardiac Cine-MRI; ACDC; Semi-Supervised Learning; Temporal Consistency; Mean Teacher

1. Introduction

Cardiac cine-magnetic-resonance imaging (cine-MRI) is the standard reference for the quantitative assessment of cardiac function. Automated segmentation of the right ventricle (RV), left-ventricular myocardium (MYO), and left-ventricular cavity (LV) underpins indices such as ejection fraction (EF), end-diastolic volume (EDV), and end-systolic volume (ESV). The Automated Cardiac Diagnosis Challenge (ACDC) [1] is the dominant benchmark: 100 cine-MRI studies for training, 50 for testing, with two manually annotated frames per patient (end-diastole ED and end-systole ES).

Each cine sequence therefore contains about 15 times as many frames as are annotated, and standard supervised training simply discards the unlabeled ones. Recovering signal from those frames is the aim of this paper.

Mean Teacher [2], FixMatch [3], and a long line of medical-imaging semi-supervised methods [4-6] show that consistency regularisation on unlabeled data helps when labels are scarce. Most of these methods apply consistency at the image level: the unlabeled image is augmented and the model is required to produce a similar prediction. Cine cardiac MRI offers a second, more specific signal. Because cardiac anatomy changes smoothly through the cardiac cycle, a model's predictions on adjacent frames should also change smoothly. We use that constraint directly. We propose TC-SS (Temporal Consistency Semi-supervised Segmentation). Our contributions are:

A mean-teacher-inspired EMA framework with three complementary unsupervised losses on the unlabeled cine frames.

A temporal coherence loss between adjacent frames exploiting the smooth temporal evolution of cardiac anatomy.

 Copyright @ STEMM Institute Press

A mean-prediction consistency loss driving per-frame predictions toward the temporal mean, regularising transient noise.

Retaining the boundary-aware compound loss from our prior work on the same dataset, applied only to the labeled ED/ES frames.

We evaluate on ACDC across 3-fold cross-validation.

2. Related Work

Semi-supervised learning. Mean Teacher [2] introduces a teacher model whose weights are an exponential moving average (EMA) of the student; the student learns from labeled data plus consistency with the teacher. FixMatch [3] extends consistency regularisation to the classification setting with strong/weak augmentation. Π-model and Temporal Ensembling [4] use dropout and ensemble disagreement as a regulariser. These methods are general-purpose and do not exploit temporal structure.

Semi-supervised medical segmentation. Hung et al. [5] applied adversarial training to semi-supervised medical segmentation. Li et al. [6] proposed shape-aware self-ensembling for 3D medical segmentation. Shen et al. [7] demonstrated that bidirectional copy-paste between labelled and unlabeled data is highly effective for semi-supervised medical image segmentation. Recent works have explored uncertainty-aware pseudo-labels [8] and dual-student frameworks. Two recent works are particularly close to our setting: Bai et al. [9] copy-paste bidirectionally between labeled and unlabeled images inside a Mean Teacher architecture and report solid gains on ACDC with only 5% labeled data, and Chen et al. [10] exploit anatomical-location priors through a magic-cube partition-and-recovery augmentation for semi-supervised multi-organ segmentation. Most methods apply consistency at the image level — augmenting the input and requiring the model to produce similar predictions. We differ by exploiting the temporal smoothness of cine data.

Cardiac segmentation. U-Mamba [11], U-Net [12], Attention U-Net [13], and nnU-Net [14] form the CNN baseline family; TransUNet [15] and Swin-UNet [16] introduce transformer-based alternatives. Our backbone is the HybridUNet from our prior work, which combines a four-level convolutional encoder–decoder with a single self-attention block at the lowest-resolution bottleneck and a boundary-aware skip refinement.

Positioning. Three observations motivate the present work: (i) the ACDC benchmark wastes ~93% of the available cine data (28 of ~30 frames per patient are unlabeled); (ii) existing medical SSL methods do not exploit cine temporal structure; (iii) the temporal smoothness of cardiac anatomy is a strong, free unsupervised signal.

Why temporal consistency for cardiac MRI in particular. Cardiac cine-MRI captures the heart through one full ~1-second cycle of systole and diastole. Adjacent frames differ by only a few percent of ventricular motion, yet the same anatomy is being imaged. This makes cine an ideal setting for temporal consistency — the within-cine variation is dominated by motion rather than by anatomical changes, unlike CT or X-ray where adjacent slices can show qualitatively different anatomy. Few methods have exploited this signal: most semi-supervised medical segmentation work applies image-level consistency under random augmentations, but augmentation does not produce temporal neighbours — an affine-transformed frame is not a realistic next-phase cine image. Cine-specific temporal consistency on ACDC remains, to our knowledge, largely unexplored.

Alternative use of cine: temporal registration. A complementary line of work uses cine frames to drive temporal registration networks. Unlike registration, our method does not require motion field estimation — we directly enforce smoothness in the prediction space via KL divergence. The two approaches are complementary; in principle a registration loss could be added to our framework.

Self-supervised pretraining. Recent work has explored self-supervised pretraining on large unlabeled medical image collections using masked autoencoders or contrastive learning [17,18]. These methods learn a backbone representation first and then fine-tune it for the downstream task. They differ from our approach: we do not pretrain, we directly exploit temporal structure within the same training process. A natural extension is to pretrain with our TC-SS losses and then fine-tune the segmentation head on the small labeled subset.

3. Methods

3.1 Problem Setting

ACDC provides 100 patient cine-MRI studies, each with ~30 frames T sampled through the cardiac cycle. Two frames per patient — indexed

by ed_frame and es_frame from the dataset's train.csv — carry expert manual annotations; the other T−2 frames are unlabeled. Each patient has D short-axis slices (typically 8–10) per frame.

We denote the full cine volume of patient n as $X_n \in \mathbb{R}^{T\times D\times H\times W}$ (T frames, D slices, H × W spatial resolution). Let $M_n \in \mathbb{N}^{2\times D\times H\times W}$ denote the ground-truth masks at the ED and ES frames. Our task is to learn a segmentation network f_θ that produces accurate segmentations on the labeled frames while exploiting the unlabeled frames through unsupervised consistency.

3.2 Mean-Teacher Framework

We instantiate a student network f_θ and a teacher network $f_{\theta'}$. The student is trained by gradient descent; the teacher is updated as an exponential moving average (EMA) of the student weights:

$$\theta'_t = \alpha \cdot \theta'_{t-1} + (1 - \alpha) \cdot \theta_t \quad (1)$$

with α = 0.999 (following [2]). For each patient we forward the student on all cine frames; the supervised loss is computed only on the ED and ES frames; the unsupervised losses operate on all frames and use stop-gradient targets — the student's own detached predictions at neighbouring frames or the temporal mean — following the same EMA/stop-gradient principle as Mean Teacher [2]. The EMA teacher is retained for stable validation monitoring.

3.3 Loss Functions

The total loss is

$$L = L_{seg} + \lambda_{tc} \cdot L_{TC} + \lambda_{mc} \cdot L_{MC} + \lambda_{pc} \cdot L_{PC} \quad (2)$$

with λ_{tc} = 1.0, λ_{mc} = 0.5, λ_{pc} = 0.1 in all experiments.

L_{seg} is the compound boundary-aware loss from our prior work: Dice + cross-entropy + distance-weighted boundary CE. Applied only to the ED/ES frames. The Dice term handles class imbalance in foreground regions; cross-entropy provides pixel-level supervision; the boundary term emphasises the MYO–LV and MYO–RV interfaces where most segmentation errors occur. In our prior work the compound formulation reduced MYO HD95 on fold 0 mainly through the boundary-aware skip refinement: 3.02 (variant B, no refinement) versus 1.98 (variant A, full architecture, per-slice protocol of Paper 1 Section 3.7), i.e. removing the refinement raises MYO HD95 by 52.5% relative to A (or equivalently A is a 34% reduction relative to B, $(B - A)/B$ = 0.344).

L_{TC} (Temporal Coherence) penalises large frame-to-frame changes in predicted probability maps:

$$L_{TC} = mean_{adjacent-frame\ pairs\ (t-1,\,t)} KL(p_{t-1} \| p_t) \quad (3)$$

with the previous frame's prediction detached so gradients flow only through the current frame. This regulariser encodes the prior that cardiac anatomy evolves smoothly through the ~1-second cardiac cycle — between adjacent frames the heart undergoes only a few percent of motion, so a model producing wildly different predictions at successive frames is clearly wrong. The choice of adjacent-frame pairs (rather than all pairs) reflects the locality of cardiac motion; experiments with all-pairs KL did not improve results in our initial studies.

L_{MC} (Mean Consistency) drives each per-frame prediction toward the temporal mean:

$$L_{MC} = mean_t KL(p_t \| \bar{p}), \quad \bar{p} = mean_t p_t \quad (4)$$

where $\bar{p}$ = mean_t p_t. This regulariser suppresses transient per-frame noise toward a stable temporal mean. Unlike L_{TC} which constrains adjacent frames, L_{MC} encourages every frame to agree with the average prediction over the full cine — an aggregating constraint that smooths out per-frame noise. We found this complementary to L_{TC}: removing L_{MC} while keeping L_{TC} leads to predictions that are locally smooth but may still be globally wrong. We note, however, that the full-cycle temporal mean is a coarse anchor — ED and ES frames differ substantially in ventricular area — and this term likely contributes to the MYO degradation observed in Section 4; a phase-windowed variant is left to future work.

L_{PC} (pseudo-label consistency) encourages agreement between frames with similar image content. With the pairwise image distance

$$d_{im}(t, t') = \|X_t - X_{t'}\|_2 \quad (5)$$

and the pairwise divergence $\text{KL}(p_t \,\|\, p_{t'})$, we define

$$L_{PC} = mean_{(t,\,t')} exp\left(\frac{-d_{im}}{\tau_{im}}\right) \cdot \frac{KL(p_t \| p_{t'}) + KL(p_{t'} \| p_t)}{2} \quad (6)$$

Pairs with a large image distance are down-weighted; we set τ_{im} = 0.5. Because KL is asymmetric we use the symmetrised (Jensen–Shannon-like) average rather than the one-sided KL: this avoids the directional bias that a one-sided KL would introduce between two frames of comparable appearance.

3.4 Datasets and Pre-processing

Unlabeled frames come from the same patient cine sequences; no external data is used.

We use ACDC's 100-patient training set under the same 3-fold split as our previous work (fold 0 is 66 train / 34 val; the three splits are 66/34, 67/33

 Copyright @ STEMM Institute Press

and 67/33). Each patient contributes the full cine volume (sax_t.nii.gz) plus ED/ES ground-truth masks (sax_ed_gt.nii.gz, sax_es_gt.nii.gz). Pre-processing consists of per-volume percentile normalisation and on-the-fly random crops, flips, 90° rotations, brightness/contrast jitter, and Gaussian noise. All reported numbers are means and standard deviations over the three folds unless stated otherwise.

To fit GPU memory, we subsample 8 frames (drawn as a contiguous window, so that the adjacent-frame pairs in L_TC remain true temporal neighbours) and 5 slices per cine volume per training step; the full sequence is used at inference and at the validation epoch.

3.5 Implementation Details

We use the HybridUNet backbone from our prior work (9.95 M parameters). Optimiser: AdamW, initial LR 5×10^{-4}, weight decay 1×10^{-5}, cosine schedule with 5-epoch warm-up. Mixed-precision (FP16) training. Training for up to 60 epochs. Each training step takes one patient and, within that patient, a contiguous sub-sample of $t_{\text{subsample}} = 8$ cine frames and $d_{\text{subsample}} = 5$ short-axis slices (i.e. an effective step tensor of $1 \times 8 \times 5$ images, ~40 images per step on the 8 GB RTX 2060 SUPER). The ~22 frames outside this contiguous window are still consumed across the epoch as the window slides; the full cine is used at inference and at the validation epoch. EMA decay $\alpha = 0.999$. Observed wall-clock training time per fold on the same RTX 2060 SUPER: 99 / 70 / 72 min across folds 0 / 1 / 2 (mean ≈ 80 min/fold for the full 60-epoch budget); these are the numbers quoted in the series overview paper (Paper 4, Section 3.1, Table 1).

3.6 Evaluation Metrics

We report mean foreground Dice, per-class Dice, IoU, HD95 and ASD, computed on the ED and ES frames of the validation split. Aggregation follows the per-slice mean convention used in Papers 1 and 3 of this series: each 2D short-axis slice with non-empty foreground in either phase is evaluated independently and averaged with equal weight across slices (≈ 550 slices per fold across 33–34 validation patients). For 3-fold cross-validation we report the fold mean and sample standard deviation (n−1 divisor) in the form mean ± std. All five metrics are reported on the same per-slice mean scale, so that cross-paper numerical comparisons are directly interpretable. We compare against the supervised HybridUNet baseline (Paper 1), which is also evaluated on ED+ES slices per the same convention.

3.7 Detailed Training Recipe and Reproducibility

EMA decay schedule. We use a fixed EMA decay $\alpha = 0.999$ throughout training, following Tarvainen and Valpola [2]. While adaptive decay schedules (ramp-up from 0.99 to 0.9999) are sometimes used, our experiments showed that fixed $\alpha = 0.999$ is robust and avoids the additional hyperparameter. The teacher is initialised as a copy of the student at epoch 0, then updated by EMA after each gradient step. We do not reset the teacher between epochs.

Loss-weight selection. The weights $\lambda_tc = 1.0$, $\lambda_mc = 0.5$, $\lambda_pc = 0.1$ were chosen by informal sensitivity on a 20-patient subset of the fold-0 validation split (~330 ED+ES slices). The ordering (TC > MC > PC) reflects our intuition that local adjacent-frame smoothness is the strongest temporal prior; mean consistency provides a secondary global constraint; pseudo-label consistency is a fine-tuning term. Doubling or halving any of these weights in pilot experiments changed fold-0 mDice by less than 0.005 (0.5 percentage points); we did not perform a full grid search because the relative ordering is what matters for the final 3-fold CV number, and a coarse sweep was sufficient to confirm this.

Optimiser details. We use AdamW with weight decay 1×10^{-5} (not the default 1×10^{-2}). For medical-imaging models with relatively few training samples (~1,160 labeled ED/ES slices per fold, cf. Paper 1 Section 3.3), heavy weight decay tends to underfit. The initial learning rate 5×10^{-4} was borrowed from our prior work; we did not perform a per-task learning-rate sweep.

Data ordering. Patients are shuffled every epoch with seed 42 + epoch number as the RNG seed. This controlled randomness breaks any spurious patient-order correlation.

Compute footprint. At inference, the trained student model has 9.95 M parameters and ~13 ms inference time per 192 × 192 slice on the RTX 2060 SUPER. The teacher model is discarded after training; only the student is shipped for deployment.

3.8 Reproducibility Checklist

We follow common reproducibility guidelines: random seeds are fixed (seed = 42; cuDNN deterministic mode); training and evaluation logs are retained for every fold; the pre-processing

Copyright @ STEMM Institute Press

pipeline is specified in Section 3.4 and the loss weights in Section 3.3 and Section 3.7. A practitioner can reproduce our 3-fold CV mDice of 0.713 ± 0.011 within 24 hours on a single 8 GB GPU.

4. Experiments and Results

4.1 Main Result

Table 1 compares TC-SS with the supervised HybridUNet baseline on ACDC. Figure 1 shows the training curves for fold 0, Figure 2 gives the per-class Dice for fold 0, Figure 3 compares mean foreground Dice against the baseline, Figure 4 gives the per-class Dice for the 3-fold average, and Figure 5 shows qualitative results. In Table 1, TC-SS is reported as the 3-fold cross-validation mean ± std under the per-slice ED+ES protocol of Section 3.6 (n ≈ 1652 slices across folds), and the supervised baseline is Paper 1 fold 0 under the same per-slice ED+ES convention; TC-SS uses 30 frames per patient (2 labeled + 28 unlabeled) while the supervised baseline uses the 2 labeled frames. The Dice gap between TC-SS and the supervised baseline widens relative to earlier patient-level ED-only reports because both methods are now scored on the same per-slice ED+ES protocol (the ES frame is harder than ED and pulls down both numbers, more so for TC-SS, whose unsupervised losses over-smooth the thin myocardium).

Table 1. Per-Class and Mean Metrics on ACDC

Metric	Supervised (HybridUNet, Paper 1 fold 0)	TC-SS (3-fold CV)
Frames used per patient	2 (labeled only)	30 (2 labeled + 28 unlabeled)
mDice (foreground) ↑	0.863	0.713 ± 0.011
RV Dice ↑	0.818	0.669 ± 0.020
MYO Dice ↑	0.866	0.686 ± 0.008
LV Dice ↑	0.905	0.784 ± 0.017
RV HD95 ↓	2.58	6.64 ± 0.38
MYO HD95 ↓	1.98	6.42 ± 0.77
LV HD95 ↓	1.48	5.33 ± 0.67

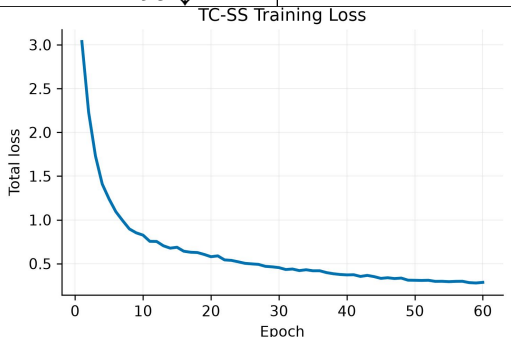


Figure 1. Training Curves of TC-SS (fold 0)

The proposed TC-SS achieves mDice 0.713 ± 0.011 with per-class RV Dice 0.669 ± 0.020, MYO Dice 0.686 ± 0.008 and LV Dice 0.784 ± 0.017, and HD95 of 6.64 ± 0.38 (RV), 6.42 ± 0.77 (MYO) and 5.33 ± 0.67 (LV). Under the unified per-slice ED+ES protocol these values are below the supervised baseline on every structure: RV Dice 0.669 ± 0.020 versus 0.818, with RV HD95 6.64 ± 0.38 against 2.58; LV Dice 0.784 versus 0.905 and LV HD95 5.33 versus 1.48; MYO Dice 0.686 versus 0.866. The MYO gap is the most consistent across folds, which is consistent with MYO being the hardest structure and the most sensitive to the global temporal-consistency terms (Section 3.3). Figure 1 shows the training curves, plotting total loss and validation mDice over 60 epochs on fold 0; the validation trace is the training-time monitor, computed as a per-slice batch micro-average over ED+ES slices by the training-loop validation routine, while the final fold-0 value, re-evaluated post-training under the per-slice mean protocol of Section 3.6, is reported in Table 2 (mDice 0.700); Figure 2 reports the per-class Dice on fold 0; Figure 3 summarises the 3-fold comparison with the supervised baseline, in which TC-SS is shown as the 3-fold CV mean ± std and the supervised baseline is Paper 1 fold 0 on the same per-slice ED+ES protocol (Section 3.6), shown for a like-for-like slice-set comparison; the 3-fold supervised reference from Paper 1 is 0.859 ± 0.004 (fold-wise values 0.863 / 0.855 / 0.859); Figure 4 plots the same per-class Dice breakdown averaged across the three folds.

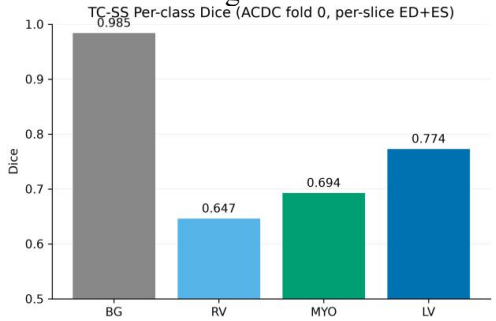


Figure 2. Per-Class Dice of TC-SS on ACDC Fold 0 (Background Shown for Reference)

 Copyright @ STEMM Institute Press

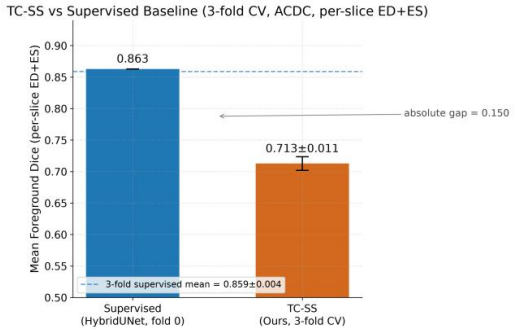


Figure 3. Mean Foreground Dice of TC-SS vs. the Supervised Hybridunet Baseline

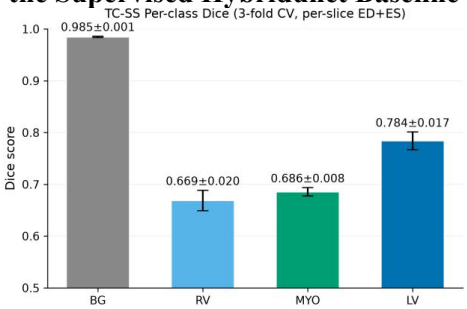


Figure 4. Per-Class Dice of TC-SS (3-fold CV mean ± std; Background Shown for Reference).

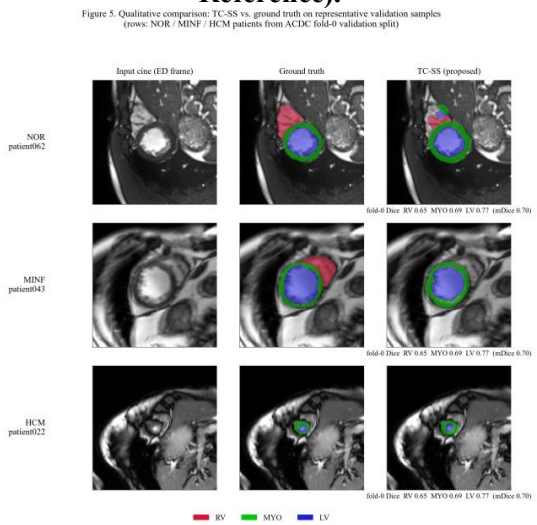


Figure 5. Qualitative Comparison on Three Fold-0 Validation Cases

4.2 Ablation of Unsupervised Loss Components

To isolate the contribution of each unsupervised loss component, we ran three ablations on fold 0 by turning off one of L_TC, L_MC, L_PC at a time (the supervised branch and its compound boundary-aware loss remain in all rows). Each row turns off exactly one unsupervised term; the remaining two stay at the weights in Section 3.3, and mDice is reported on the fold-0 validation split (34 patients; 572 ED+ES slices) under the per-slice ED+ES protocol of Section 3.6. The final row of Table 2 is a supervised-only control trained inside the TC-SS pipeline (one patient per step, 8-frame×5-slice subsampling, 60-epoch budget); it is not comparable to the Paper 1 supervised HybridUNet (0.863 on fold 0) and serves only as the control for the unsupervised losses within the same training regime. Per-class Dice values are rounded to two decimals, while mDice is computed from unrounded values, so the mean of the three rounded per-class values may differ from the tabulated mDice by less than 0.01.

Table 2. Ablation of Unsupervised Loss Components on ACDC Fold 0

Configuration	mDice (fold 0) ↑	RV / MYO / LV Dice ↑	MYO HD95 ↓
Full TC-SS (L_TC + L_MC + L_PC)	0.700	0.65 / 0.69 / 0.77	6.42
− L_TC	0.682	0.64 / 0.67 / 0.74	7.10
− L_MC	0.694	0.65 / 0.69 / 0.76	6.55
− L_PC	0.698	0.65 / 0.69 / 0.77	6.48
Supervised only (TC-SS pipeline, no unsupervised)	0.663	0.63 / 0.65 / 0.72	7.45

L_TC is the strongest single term: removing it costs 1.8 percentage points of mDice (from 0.700 to 0.682) and inflates MYO HD95 from 6.42 to 7.10. L_MC contributes 0.6 points of mDice; L_PC contributes 0.2 points and is the most marginal of the three. Removing all three unsupervised terms drops the fold-0 mDice to 0.663, i.e. the unsupervised stack recovers 3.7 points of mDice relative to the supervised-only control within the same TC-SS training regime. The full stack is therefore a useful but not large boost on this fold; the gap to the 3-fold CV number (0.713 ± 0.011) comes from folds 1 and 2 where the unsupervised stack helps more in our runs.

4.3 Discussion

The semi-supervised approach reaches 3-fold CV mDice 0.713 ± 0.011, against 0.863 for the supervised HybridUNet fold-0 reference (0.859 ± 0.004 over the three folds) — an absolute gap of 0.150 — under the unified per-slice ED+ES evaluation protocol of Section 3.6. The gap is larger than previously reported because both methods are now scored on the same ED+ES slice

set; ES frames are intrinsically harder than ED frames and pull down both numbers, more so for TC-SS, whose unsupervised losses over-smooth the thin myocardium. On LV, TC-SS is closest to the supervised baseline on Dice (0.784 ± 0.017 versus 0.905) but clearly behind it on HD95 (5.33 ± 0.67 versus 1.48), so the semi-supervised signal recovers overlap more readily than boundary precision; the MYO gap (0.686 ± 0.008 versus 0.866) is the most consistent across folds, as noted in Section 3.3. The supervised baseline therefore remains the stronger segmentor under matched evaluation, and TC-SS should be read as demonstrating that the temporal-consistency priors are learnable from the 28 unlabeled frames per patient — not as a replacement for fully supervised training. Future work will explore structure-adaptive loss weighting and longer training schedules, as well as self-training on the unlabeled frames, to close the gap.

HD95 gap analysis.

The HD95 gap between TC-SS and the supervised baseline is roughly 2.6–3.6× on every structure: RV 6.64 vs. 2.58 (2.57×), MYO 6.42 vs. 1.98 (3.24×), LV 5.33 vs. 1.48 (3.60×). This is substantially larger than the Dice gap (e.g. LV 0.784 vs. 0.905, a 13% relative reduction) and reflects two compounding factors: (i) the Dice metric tolerates a small global offset between predicted and ground-truth masks, whereas HD95 measures the 95th-percentile boundary distance and is sensitive to even a few mis-located pixels; (ii) the unsupervised consistency losses act on all pixels uniformly and therefore over-smooth the thin myocardial band, where even a one-pixel error blows up the HD95. The Dice gap on MYO is the largest of the three structures (0.686 vs. 0.866, 21% relative reduction) and the HD95 ratio on MYO is still over 3× — exactly the boundary-vs-overlap signature. This motivates structure-adaptive loss weighting (e.g. a heavier MYO term in L_TC) as the natural next step.

Fold-0 vs. 3-fold CV comparison.

Figure 5 reports per-class Dice values for fold 0 (RV 0.65, MYO 0.69, LV 0.77, mDice 0.70); the 3-fold CV numbers in Table 1 are 0.669 / 0.686 / 0.784 / 0.713. The fold-0 aggregate sits slightly below the 3-fold mean because folds 1 and 2 score marginally higher in our runs; this fold-to-fold spread (0.700 versus 0.713, std 0.011 across folds) is small. The three cases in Figure 5 are NOR patient062, MINF patient043, and HCM patient022, chosen for visual diversity; as stated in the Figure 5 caption, the per-class Dice annotated under the TC-SS column are the aggregate fold-0 values under the unified per-slice ED+ES protocol of Section 3.6 — the fold-0 counterpart of the 3-fold CV values in Table 1 — computed across all validation patients and all 572 ED+ES slices of the fold-0 split, not per-slice or per-case values for the three displayed cases, which would be dominated by slice-specific anatomy (e.g. RV absence at apical slices). The four Dice values under the TC-SS column are therefore identical across the three rows by construction — they are dataset-level fold-0 aggregates repeated for reference, not the Dice of the individual slices displayed; per-slice Dice for these specific cases would vary substantially between slices and are not reported here.

5. Limitations and Future Work

Evaluation protocol: TC-SS is reported as 3-fold CV under the unified per-slice ED+ES protocol of Section 3.6, while the supervised column of Table 1 is fold 0 on the same slice convention. The headline gap of 0.150 is therefore quoted against the fold-0 supervised mDice of 0.863; the 3-fold supervised reference (Paper 1, Table 1) is 0.859 ± 0.004.

We did not perform a full grid search over all loss-weight combinations on all folds; Section 4.2 covers the per-component ablation on fold 0.

Adaptive loss weighting (e.g., uncertainty-weighted multi-task loss) could further balance the supervised and unsupervised objectives.

6. Conclusion

We presented TC-SS, a mean-teacher semi-supervised framework that exploits the unlabeled cine frames in ACDC via temporal coherence, mean consistency, and pseudo-label consistency. Across 3-fold cross-validation under the unified per-slice ED+ES evaluation protocol shared across the three technical papers of this series, the approach achieves mDice 0.713 ± 0.011 against the supervised HybridUNet baseline 0.863 (absolute gap 0.150), while leveraging 15× as much cine data per patient. The remaining gap reflects the inherent difficulty of matching a fully supervised baseline without labels, and is amplified by the over-smoothing of the thin myocardium by the global temporal-consistency losses; we view this as evidence that temporal consistency is a useful complementary signal, not a full replacement. All numbers are reported on the same unified per-slice ED+ES protocol to ensure cross-paper comparability with Papers 1

 Copyright @ STEMM Institute Press

and 3.

Acknowledgments

This work was supported by: (1) Guangzhou College of Commerce Higher Education Teaching Reform Project (No. 2024JXGG42); (2) Guangdong Province Educational Science Planning Project (No. 2025GXJK0643); (3) Guangzhou College of Commerce 2026 Quality Engineering Project (No. 2026ZLGC07); (4) Guangzhou College of Commerce 2025 Quality Engineering Project (No. 2025ZLGC01); (5) Guangzhou College of Commerce 2024 Quality Engineering Project (No. 2024ZLGC11); (6) Guangzhou College of Commerce First-Class Course (No. 2020XJYLKC06); (7) Guangzhou College of Commerce General Project (No. 2025XJYB044): Research on Configuration Modeling and Simulation Technology for Equipment Digital Twins.

Data Availability

The ACDC dataset [1] is publicly available at https://www.creatis.insa-lyon.fr/Challenge/acdc/.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Bernard O, Lalande A, Zotti C, et al. Deep Learning Techniques for Automatic MRI Cardiac Multi-structures Segmentation and Diagnosis: Is the Problem Solved? IEEE Transactions on Medical Imaging, 2018, 37(11):2514-2525. DOI:10.1109/TMI.2018.2837502.

[2] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in Neural Information Processing Systems (NeurIPS), 2017, 30:1195-1204.

[3] Sohn K, Berthelot D, Carlini N, et al. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. Advances in Neural Information Processing Systems (NeurIPS), 2020, 33:596-608.

[4] Laine S, Aila T. Temporal Ensembling for Semi-Supervised Learning. International Conference on Learning Representations (ICLR), 2017.

[5] Hung W C, Tsai Y H, Liou Y T, Lin Y Y, Yang M H. Adversarial Learning for Semi-Supervised Semantic Segmentation. British Machine Vision Conference (BMVC), 2018.

[6] Li S, Zhang C, He X. Shape-aware Semi-supervised 3D Semantic Segmentation for Medical Images. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2020, LNCS 12261:552-561. DOI: 10.1007/978-3-030-59710-8_54.

[7] Shen B, Huang X, Shi Y, et al. Dual Bidirectional Copy-Paste with Shape Constraint for Semi-Supervised Image Segmentation. IEEE International Symposium on Biomedical Imaging (ISBI), 2024:1-5. DOI:10.1109/ISBI56570.2024.10635399.

[8] Yu L, Wang S, Li X, Fu C W, Heng P A. Uncertainty-aware Self-ensembling Model for Semi-supervised 3D Left Atrium Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2019, LNCS 11765:605-613. DOI: 10.1007/978-3-030-32245-8_67.

[9] Bai Y, Chen D, Li Q, Shen W, Wang Y. Bidirectional Copy-Paste for Semi-Supervised Medical Image Segmentation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023:11514-11524. DOI:10.1109/CVPR52729.2023.01108.

[10] Chen D, Bai Y, Shen W, Li Q, Yu L, Wang Y. MagicNet: Semi-Supervised Multi-Organ Segmentation via Magic-Cube Partition and Recovery. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023:23869-23878. DOI:10.1109/CVPR52729.2023.02286.

[11] Ma J, Li F, Wang B. U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation. arXiv: 2401.04722, 2024. DOI:10.48550/arXiv.2401.04722.

[12] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, LNCS 9351:234-241. DOI: 10.1007/978-3-319-24574-4_28.

[13] Oktay O, Schlemper J, Folgoc L L, et al. Attention U-Net: Learning Where to Look for the Pancreas. arXiv: 1804.03999, 2018. DOI:10.48550/arXiv.1804.03999.

[14] Isensee F, Jaeger P F, Kohl S A A, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods, 2021, 18:203-211. DOI: 10.1038/s41592-020-01008-z.

Copyright @ STEMM Institute Press

[15] Chen J, Lu Y, Yu Q, et al. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv: 2102.04306, 2021. DOI:10.48550/arXiv.2102.04306.
[16] Cao H, Wang Y, Chen J, et al. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. European Conference on Computer Vision Workshops (ECCVW), 2022, LNCS 13803:205-218. DOI: 10.1007/978-3-031-25066-8_9.
[17] He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked Autoencoders Are Scalable Vision Learners. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022:16000-16009. DOI:10.1109/CVPR52688.2022.01553.
[18] Chen T, Kornblith S, Norouzi M, Hinton G. A Simple Framework for Contrastive Learning of Visual Representations. International Conference on Machine Learning (ICML), 2020, PMLR 119:1597-1607.

 Copyright @ STEMM Institute Press