Received: 11 January 2025 | Revised: 19 August 2025 | Accepted: 3 October 2025

DOI: 10.1002/mp.70094

RESEARCH ARTICLE

MEDICAL PHYSICS

Semi-supervised cine cardiac MRI segmentation via joint registration and temporal attention perceiver

Yinqi Qin[1] | **Fumin Guo**[1] | **Ziyin Wang**[1] | **Sa Xiao**[2] | **Lei Zhang**[2] | **Xin Zhou**[1,2]

[1]Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, China

[2]State Key Laboratory of Magnetic Resonance and Atomic and Molecular Physics, National Centre for Magnetic Resonance in Wuhan, Wuhan Institute of Physics and Mathematics, Innovation Academy for Precision Measurement Science and Technology, Chinese Academy of Sciences, Wuhan, China

Correspondence
Xin Zhou, National Center for Magnetic Resonance in Wuhan, Wuhan Institute of Physics and Mathematics, Innovation Academy for Precision Measurement Science and Technology, 30 West Xiaohongshan, Wuhan 430071, China.
Email: xinzhou@wipm.ac.cn

Funding information
National Key Research and Development Program of China, Grant/Award Numbers: 2018YFA0704000, 2022YFC2410000; National Natural Science Foundation of China, Grant/Award Numbers: 82572361, 82127802, 21921004, U21A20392; Key Research Program of Frontier Science, Chinese Academy of Sciences, Grant/Award Number: ZDBS-LY-JSC004; Hubei Provincial Key Technology Foundataion of China, Grant/Award Numbers: 2021ACA013, 2023BAA021; Natural Science Foundation of Hubei Province, Grant/Award Number: 2023AFB1061

Abstract

Background: Segmentation of cardiac structures is essential for cardiac function evaluation using cine magnetic resonance imaging (MRI). Deep learning can be used to segment cardiac structures in cine cardiac MRI with high accuracy, but this approach requires fully annotated datasets for training, which are difficult to obtain. Semi-supervised segmentation methods provide a way to alleviate the burden of manual labeling by using labeled and unlabeled data for training. However, these methods generally provide suboptimal segmentation accuracies.

Purpose: To develop a semi-supervised method that utilizes relatively small training datasets and under-annotations for improved cine cardiac MRI segmentation.

Methods: The proposed approach consists of deformable registration, fully and weakly supervised segmentation, and a temporal attention perceiver (TAP). The registration module was trained to warp labeled frames to generate pseudo labels for unlabeled frames. The warped labeled images were used to train the fully supervised segmentation network. The unlabeled images and the pseudo label were used to train the weakly supervised segmentation model, and the segmentation prediction was compared with the input pseudo label as an auxiliary loss to the registration module. The TAP module was employed to generate optimized features for the warped labeled and the original unlabeled images both paired with the original labeled image. Consistency between the resulting features was enforced to refine cross-instance feature alignment to facilitate the registration. One hundred, twenty, and ten subjects from the Automatic Cardiac Diagnosis Challenge (ACDC) and seventy-five, thirty, and fifteen cases from the Multi-Vendor & Multi-Disease (M&Ms) Cardiac Image Segmentation Challenge were used for training, each with random end-systolic (ES)/end-diastolic (ED) frames labeled. The optimized models were used to segment the remaining 50 ACDC and 50 M&Ms subjects. The proposed approach was compared with several commonly used semi-supervised segmentation methods in terms of Dice-similarity-coefficients (DSC), average-symmetric-surface-distance (ASSD), and Hausdorff-distance (HD) for left (LV) and right (RV) ventricular cavity and myocardium (Myo). A Unet trained on the same subjects each with both frames labeled was used as an upper bound (Unet_UB).

Results: Using 100 ACDC training subjects, our approach yielded DSC = 0.910±0.063, ASSD = 1.37±0.63 mm, and HD = 6.38±2.99 mm for RV, DSC = 0.894±0.024, ASSD = 1.20±1.12 mm, and HD = 4.67±3.22 mm for Myo, and DSC = 0.934±0.056, ASSD = 1.25±1.63 mm, and HD = 3.97±5.76 mm for LV. A bidirectional copy-paste (BCP) method performed the best among the com-

Yinqi Qin and Fumin Guo contributed equally to this work.

parative methods and generated DSC = 0.902±0.060, ASSD = 1.45±0.60 mm, and HD = 7.50±3.20 mm for RV, DSC = 0.885±0.030, ASSD = 1.28±0.80 mm, and HD = 5.80±2.80 mm for Myo, and DSC = 0.920±0.068, ASSD = 1.15±0.40 mm, and HD = 4.20±3.30 mm for LV. For Unet_UB, these were 0.905±0.068, 1.48±0.61 mm, and 6.35±2.85 mm for RV, 0.895±0.030, 1.05±0.45 mm, and 4.40±3.09 mm for Myo, and 0.941±0.044, 1.02±0.34 mm, and 3.17±1.63 mm for LV. Similar trends were observed when using 75 M&Ms training subjects. For all the experiments, our approach outperformed BCP in general and yielded segmentation accuracies comparable to Unet_UB.

Conclusions: The proposed approach outperformed several commonly used semi-supervised segmentation methods and yielded segmentation accuracies on par with fully supervised Unet using various relatively small datasets and under annotations for training.

KEYWORDS

cine cardiac MRI, joint registration and segmentation, semi-supervised segmentation, temporal attention perceiver

1 | INTRODUCTION

Cine cardiac magnetic resonance imaging (MRI) is a sequence of individual frames that correspond to different timepoints in a cardiac cycle. Cine MRI is typically acquired by dividing a cardiac cycle into multiple frames under electrocardiogram triggering, and each frame contains information gathered over multiple heartbeats, resulting in a series of frames of the heart that can be displayed as a movie.[1] Cine MRI has been established as the gold standard for evaluating cardiac function, including chamber volumes, ejection fraction, myocardial thickness and mass, and wall motion dynamics, with high spatial and temporal resolution and no ionizing radiation.[2] Segmentation of cardiac structures is required to generate these cardiac functional measurements. Although manual segmentation remains an option, this approach requires significant expertise, and is labor-intensive, time-consuming, and user-dependent, incompatible with efficient and high-throughput clinical workflow. Recently, deep learning algorithms have demonstrated numerous promise in cardiac image segmentation when trained with large-scale and carefully annotated datasets in a fully supervised manner.[3–5] However, collecting large-scale cine MRI datasets requires substantial investment of medical resources, and manual annotation is labor-intensive, time-consuming, variable, and demands significant expertise.[6] Accordingly, there is an urgent need to develop segmentation algorithms that require relatively small training datasets and fewer annotations for widespread application of deep learning in clinical cardiac MRI workflow.

1.1 | Related work

Semi-supervised learning provides a way to alleviate the critical requirements of fully annotated dataset by utilizing information learned from both labeled and unlabeled data. In general, semi-supervised segmentation methods may be categorized into three classes: pseudo labeled-based methods, consistency regularization-based methods, and atlas-based methods.

Pseudo label-based methods usually adopt a model, for example, a pre-trained model, to generate pseudo label for unlabeled input data, which is used as new sample alone or mixed with labeled data for further training.[7] Thompson et al.[8] proposed to generate super pixel maps and leverage their features to improve the accuracy of pseudo label during training. Wang et al.[9] introduced a network that utilizes a re-weighting method to evaluate the loss function and select the pseudo label with high confidence. Shi et al.[10] developed a framework that involves uncertainty estimation based on subnetwork prediction inconsistency and a self-training strategy to generate more reliable predictions. Kalluri et al.[11] employed pixel-wise entropy regularization to align deep feature representations across different domain to generate low entropy pseudo label for unlabeled data. Chen et al.[12] developed a cross pseudo supervision (CPS) framework that comprises two differently initialized segmentation networks. The pseudo label generated by one network was used to supervise the other, and bidirectional consistency between the two networks' output was enforced. Cui et al.[13] estimated segmentation prediction uncertainty during training to rectify the pseudo label for unsupervised segmentation. Sohn et al.[14] introduced FixMatch, where the predictions, with probability above a threshold, of weakly augmented images were used as supervision for the same input images undergoing strong augmentation. Yang et al.[15] augmented input images using different strong perturbations, and unified image and feature perturbations in independent streams, dubbed unified dual-stream perturbations approach (UniMatch), to improve FixMatch.[14]

Bai et al.[16] randomly cropped the foreground in a labeled image and copy-pasted the cropped region onto the corresponding region in an unlabeled image. The same operation was then reversed, also known as bidirectional copy–paste (BCP), resulting in augmented inputs to a student network by mixing the original labeled and unlabeled images. Pseudo label generated by a teacher model was similarly mixed with manual masks and used as supervision for the student network. Song et al.[17] extended the BCP architecture[16] by incorporating two student models to learn heterogeneous features. Discrepancies between the two student networks' outputs were corrected by minimizing the distances for correct segmentation voxels and maximizing the entropy for erroneous segmentation voxels.

Consistency regularization-based algorithms typically utilize multiple models and encourage the models to generate invariant predictions for the same input at different views or under different perturbations. Mean-teacher methods[18,19] optimize a student model by encouraging consistency between the predictions of a teacher model and a student model with noise added to the input. Typically, the teacher model estimates segmentation uncertainty using Monte Carlo dropout, filtering out unreliable predictions and allowing the student model to focus on reliable cases. Xie et al.[20] developed a method that consists of a segmentation network (S-Net) and a pairwise relation network (PR-Net) that was trained to exploit the semantic consistency for each pair of images in feature space. The shared encoders of the two networks exploit the learned image representation ability of the PR-Net to improve the performance of the S-Net. Zhang et al.[21] employed a deep adversarial networks (DAN) to segment both labeled and unlabeled data and then used an evaluation network to differentiate segmentation quality for labeled and unlabeled images, which was used to fool the evaluation network and to improve the segmentation quality. Qiao et al.[22] employed deep co-training to train multiple neural networks using unlabeled images in different views and encouraged differences between networks' output by minimizing the Jensen–Shannon divergence between these predictions. Luo et al.[23] utilized a pyramid-prediction network to generate segmentation predictions for unlabeled images at multiple scales and minimized the differences between each prediction and their average weighted by multi-scale uncertainty. Ouali et al.[24] developed a network that consists of an encoder and multiple different decoders to generate various outputs as pseudo label for cross supervision with consistency training (CCT). Luo et al.[25] introduced an uncertainty rectified pyramid consistency (URPC) framework that comprises a convolutional neural network and a transformer, and the predictions from one network were used to train the other. Ma et al.[26] incorporated Mamba as the backbone of a Unet (MambaUnet), and employed pixel-level cross supervision and contrastive learning to enhance feature learning.

Atlas-based semi-supervised segmentation methods[27,28] typically employ a registration unit to align labeled images with unlabeled data. The input labeled images are also used to train a segmentation model to segment the unlabeled images, which provides a way to improve the registration through evaluation of the warped label and the unlabeled image segmentation. Xu et al.[27] proposed a DeepAtlas framework that employs a registration module to warp labeled images to unlabeled images and a segmentation module to generate predictions of unlabeled data. The segmentation results were then used to supervise the registration based on the anatomical similarity. Wang et al.[29] employed a CNN to learn voxel-wise correspondences between an atlas image and unlabeled images under a forward-backward consistency constraint. Elmahdy et al.[30] developed a 3D adversarial network for joint registration and segmentation. The generator estimated and applied deformation field to the moving image and the segmentation. A discriminator network was used to evaluate the alignment of the moving image and the segmentation with the fixed image. Dinsdale et al.[31] used an adapted spatial transformer network to learn the deformation field and resample the initial mask to create the final segmentation. Learning better registration to learn better segmentation (BRBS) involved deforming labeled images into unlabeled data to generate pseudo label.[28] Warped label and pseudo label were then used to train a segmentation model, and the segmentation predictions of labeled and unlabeled images were used to facilitate registration by providing anatomical guidance.

Although effective, existing pseudo-label and consistency regularization-based methods heavily rely on original manual label as the only source of semantic guidance and are limited by inefficient utilization of labeled data in general. In atlas-based methods, manual masks warped using various deformation field generated by image registration are used as pseudo labels for unlabeled data, effectively enhancing the diversity of semantic guidance and enlarging the size of training samples by treating the numerous intermediate warped labeled images as augmented samples. However, registration may result in mismatch between the pseudo labels and the unlabeled data, which are used together with the warped labeled images to train a single segmentation network, leading to confirmation bias and potentially impaired segmentation performance for existing atlas-based methods. In addition, registration is predominantly performed and constrained in image space. Although constraints in feature space have been explored, these methods typically rely on single-image representation that introduces biases. We think that exploring inter-image relationships for feature representation provides a way to reduce encoding biases and

generate optimized features for robust feature alignment and registration constraint. This may result in improved registration accuracy, which in turn enhances the segmentation performance. Accordingly, we proposed to address some of these issues and develop an atlas-based approach based on the DeepAtlas architecture for improved semi-supervised cine cardiac MRI segmentation.

1.2 | Contributions

We summarize the novelty and main contributions of our work as follows:

1. We incorporated two segmentation branches, one for fully supervised segmentation trained on matched warped labeled images generated by a registration module and the other for weakly supervised segmentation trained using the resulting pseudo label and original unlabeled images. The dual segmentation mechanism promotes effective optimization of each segmentation network that is designed to handle different matched and mismatched data and capture heterogeneous features collaboratively, alleviating confirmation bias and leading to improved segmentation performance.
2. We paired the warped labeled and original unlabeled images with the original labeled data to generate two image pairs. We proposed a temporal attention perceiver (TAP) to extract features from the two images in an input image pair independently and modulate the channel-wise weights based on the relationships between the two feature instances. The estimated channel-wise weights were used to scale and aggregate the two features to generate unbiased feature for the input image pair. Consistency between the resulting features for the two image pairs was enforced to further optimize the registration to generate more reliable pseudo label and spatial prior for improved segmentation.
3. We comprehensively evaluated the effects the dual segmentation and TAP modules as well as the entire algorithm pipeline. For two public cine MRI datasets, our approach outperformed several state-of-the-art semi-supervised segmentation methods and approached fully supervised training when using relatively small labeled and unlabeled datasets for training.

2 | METHODS

2.1 | Overview of algorithm pipeline

Our objective is to segment perhaps the most widely used end-systolic (ES) and end-diastolic (ED) frames in cine MRI using relatively small datasets with manual annotation at either the ED or ES frame for training. Figure 1 shows the schematic of the proposed semi-supervised segmentation framework, which consists of: 1) a registration module that warps the moving labeled image I_m and the segmentation S_m to the fixed unlabeled image I_f, yielding warped moving image I'_m and label S'_m; 2) a dual segmentation module that performs fully and weakly supervised segmentation of the warped labeled (I'_m and S'_m) and original unlabeled (I_f, S'_m) images, respectively, and provides semantic knowledge to facilitate the registration in 1); 3) a temporal attention perceiver (TAP) that encourages spatial and temporal consistency between the original image pair (I_m, I_f) and the warped pair (I_m, I'_m) in feature space to improve the registration in 1).

2.2 | Deformable image registration

For each subject, we randomly selected the ED or ES frame as unlabeled fixed image I_f and the other as labeled moving image I_m. We employed VoxelMorph[32] as the registration network *Reg*, which was pre-trained using the training dataset without any label. For each pair of input images I_m and I_f, the predicted displacement field $D = Reg(I_m, I_f)$ was used to generate the deformation field ϕ by incorporating identify transform Id, that is, $\phi = D + Id$. The resulting deformation field ϕ was used to warp the moving image I_m and its manual label S_m to match I_f. The warped moving image I'_m and label S'_m were obtained using: $I'_m = I_m \circ \phi$ and $S'_m = S_m \circ \phi$, where "$\circ$" denotes a differentiable interpolation operation implemented based on spatial transformer networks (STN)[33]. Briefly, STN consists of a localization network to regress transformation parameters, a grid generator to generate coordinates corresponding to image pixels, and a sampler that applies the transformation parameters to the input image. In particular, for a voxel p in I_m, its new location p' was calculated using: $p' = p + D(p)$. Since image signal intensities are only defined at integer locations, we linearly interpolated the signal intensities at the neighboring voxels of p' in I_m following previous work[32]: $I_m \circ \phi(p) = \sum_{q \in N(p')} I_m(q) \prod_{d \in \{x,y,z\}} (1 - \left| p'_d - q_d \right|)$, where $N(p')$ represents the neighboring voxels of p', d indicates dimensions, and $|\cdot|$ denotes absolute distance between two locations. For S'_m calculation, we split S_m into multiple binary images each representing one class and transformed each of them using the same operation as $I_m \circ \phi$. Note that the warped image I'_m and label S'_m could be viewed as augmented I_m and S_m, respectively, which were adopted to train a segmentation network in a supervised manner.

The registration network *Reg* was optimized by minimizing the similarity loss L_{Reg_sim} between the warped moving image I'_m and the fixed image I_f. In addition, a

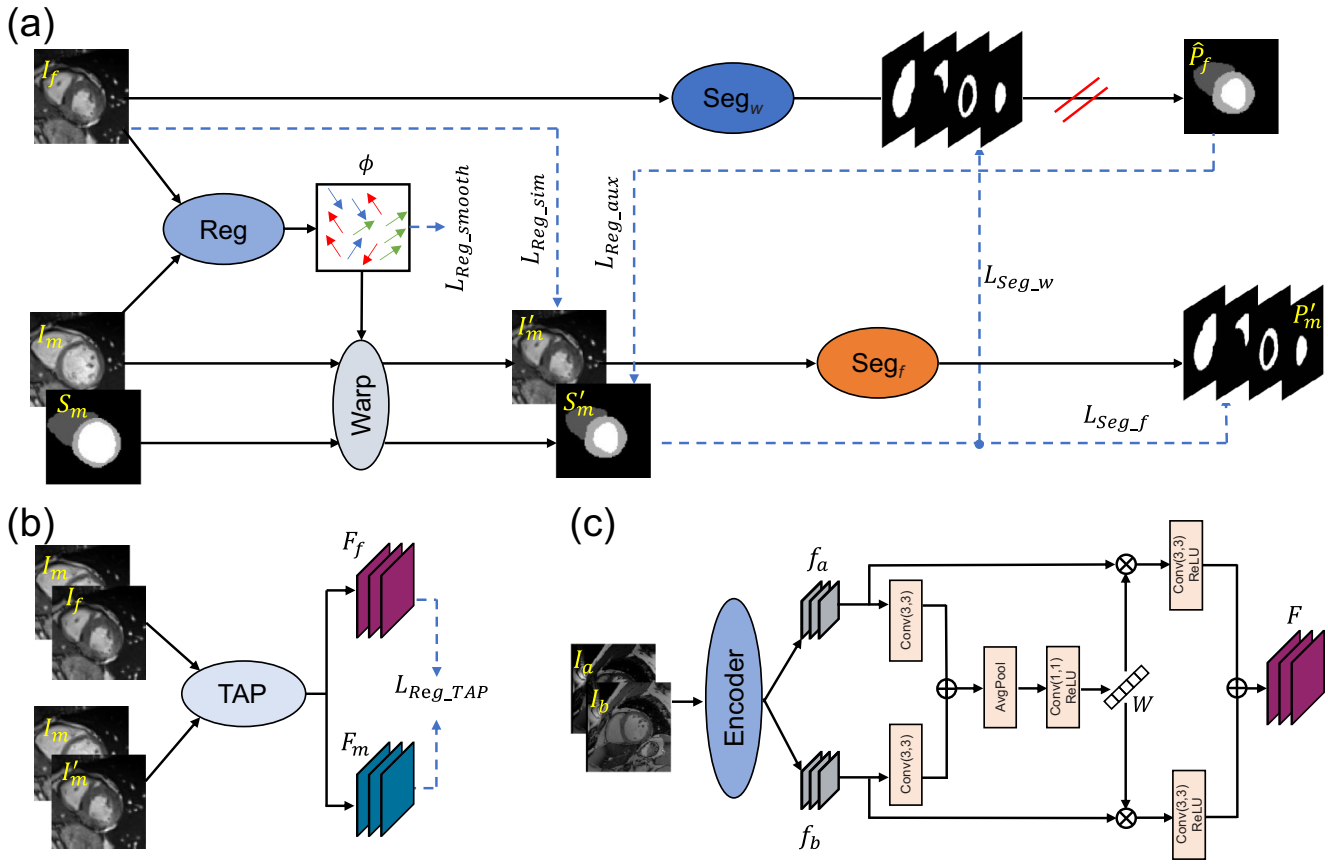

FIGURE 1 Overview of the proposed registration-based semi-supervised segmentation algorithm. (a) Joint registration and segmentation. Labeled moving image I_m and the segmentation S_m were aligned with unlabeled fixed image I_f through deformable registration using *Reg*. Seg_f was trained using I'_m and S'_m in fully supervised manner. Seg_w was trained on I_f and S'_m in a weakly supervised manner to generate prediction $\hat{P}_f$, which was used to facilitate the registration of I_m and I_f. (b) TAP that encourages the spatial and temporal consistency between warped image pair I_m and I'_m and original image pair I_m and I_f. (c) Architecture of TAP. Two input images I_a and I_b were independently entered into an encoder to extract features f_a and f_b, respectively, which were used to generate channel-wise attention weights W and the final spatial-temporal feature F that encodes information within and between I_a and I_b. ⊕: element-wise summation, ⊗: element-wise multiplication. TAP, temporal attention perceiver.

regularization loss L_{Reg_smooth} was adopted to promote smoothness of the estimated deformation field ϕ. These two terms were combined and used as the baseline loss of the registration network:

$$L_{Reg\_baseline}(I_m, I_f) = L_{Reg\_sim}(I_m \circ \phi, I_f) + \alpha \cdot L_{Reg\_smooth}(\phi), \quad (1)$$

where α denotes the weight of the regularization term.

We employed mean squared error for the similarity loss L_{Reg_sim} and diffusion regularizer on the differences between neighboring pixels in the deformation field for the regularization loss L_{Reg_smooth}, that is, $L_{Reg\_smooth}(\phi) = \sum_{x \in \Omega} ||\nabla \phi(x)||^2$, where x and Ω indicate a pixel and the image domain, respectively.

2.3 | Dual segmentation and feedback on registration

In the DeepAtlas framework[27], warped moving images I'_m and warped label S'_m, and fixed images I_f and the pseudo label S'_m were used together to train a single segmentation network. However, registration error results in mismatch between I_f and S'_m, and the use of matched and mismatched data together to train a single network leads to confirmation bias, potentially impairing the segmentation performance. In this work, we proposed to train a fully-supervised network Seg_f using the matched warped labeled data, and an independent weakly supervised network Seg_w using the potentially mismatched unlabeled images and the pseudo label. This mechanism provides a way to optimize each network effectively using different matched and mismatched data, and offers additional benefits because previous studies[12] show that using multiple independent networks enables leveraging the inherent heterogeneity in learning dynamics, enhancing the robustness of feature learning and network optimization. In the proposed approach, Seg_f and Seg_w share identical encoder–decoder structures and were independently initialized. Seg_f parameterized by θ_f takes the warped moving images I'_m and warped label S'_m

as input and generates segmentation prediction $P'_m = Seg_f(I'_m; \theta_f)$. Cross entropy ($CE(,)$) between S'_m and P'_m was used as the loss for Seg_f:

$$L_{Seg\_f}(P'_m, S'_m) = CE(P'_m, S'_m). \quad (2)$$

Let C be the number of classes. Given manual label $P \in \{0, 1\}^C$ and network softmax segmentation output $S \in [0, 1]^C$, $CE(P, S)$ is calculated as: $CE(P, S) = \sum_{c=1}^{C} -P^c \cdot log(S^c)$, where $P^c \in \{0, 1\}$ is the one-hot encoded vector of P for class c, $S^c \in [0, 1]$ is the vectorized cth component of network output S, satisfying $\sum_{c=1}^{C} S^c(x) = 1, \forall x \in \Omega$.

Seg_w was introduced to segment the unlabeled fixed images I_f to facilitate optimization of the registration network. As shown in Figure 1, Seg_w parameterized by θ_w takes the unlabeled fixed images I_f and the pseudo label S'_m as input, and produces segmentation $P_f = Seg_w(I_f; \theta_w)$. Similarly, $CE(,)$ was used for weakly supervised training of Seg_w:

$$L_{Seg\_w}(P_f, S'_m) = CE(P_f, S'_m). \quad (3)$$

During each training iteration, the parameters of Seg_f were frozen when optimizing Seg_w, and Seg_f was optimized in the same manner. Subsequently, Seg_w was applied to unlabeled fixed images I_f and the resulting probability maps P_f was used to generate the segmentation $\widehat{P}_f$ through *argmax*. Mean absolute error between network prediction $\widehat{P}_f$ and the input warped moving label S'_m was calculated to provide auxiliary supervision for the registration module as follows:

$$L_{Reg\_aux}(S'_m, \widehat{P}_f) = \frac{1}{|\Omega|} \sum_{x \in \Omega} \left| S'_m(x) - \widehat{P}_f(x) \right|. \quad (4)$$

We think that *Reg* and Seg_w can benefit each other and improve the overall registration and semi-supervised segmentation performance. The pseudo label S'_m generated by *Reg* was used to supervise the training of Seg_w and the quality of S'_m affects the optimization of Seg_w. For example, a decent registration by *Reg* would generate pseudo label S'_m that matches the unlabeled fixed images I_f well, facilitating optimization of Seg_w and resulting in satisfactory segmentation prediction $\widehat{P}_f$ for I_f, which in turn promotes *Reg* to generate improved S'_m through the loss L_{Reg_aux}. In contrast, a problematic registration and the resulting unacceptable pseudo label S'_m would cause difficulties in optimizing Seg_w, leading to inaccurate segmentation prediction $\widehat{P}_f$ that adversely affects the quality of S'_m, which further hampers the performance of *Reg*. In this work, the deformable registration network *Reg* was optimized with losses encompassing $L_{Reg\_baseline}$ that encourages the similarity between warped labeled images I'_m and unlabeled images I_f, suggesting that the resulting pseudo label S'_m is appropriate for optimization of Seg_w. Therefore, we think the registration module *Reg* and semi-supervised segmentation network Seg_w benefit each other and their interaction contributes to the improved performance for both components.

2.4 | Temporal attention perceiver

Commonly used registration models typically enforce similarity of fixed and warped moving images at image signal intensity level. Although some studies attempted to enforce constraints in latent feature space, the features are mainly encoded for a single image, introducing potential biases. Here, we proposed to extract unbiased features from fixed and moving images by generating image pairs and exploring inter-image relationships. We then enforced consistency between the resulting optimized features to constrain the alignment of the fixed and warped moving images in higher dimensions for improved registration accuracy, which in turn enhances the segmentation performance. For example, the fixed and warped moving images, when well-aligned, are not only similar between themselves in spatial dimensions, but also similar with respect to the original moving image in latent space. Accordingly, we paired the warped labeled I'_m and original unlabeled I_f images with the original labeled data I_m to generate two image pairs (I_m, I'_m) and (I_m, I_f). We proposed a TAP that takes an image pair as input and generates final features F_m and F_f that represent the spatial and temporal embeddings within and between the two images in (I_m, I'_m) and (I_m, I_f), respectively. Note that F_f extracted from (I_m, I_f) encodes both the spatial and the temporal information between I_m and I_f, and should be consistent with F_m extracted from I_m and I'_m. To this end, we enforced consistency between F_f and F_m as a way to constrain the registration of I_m and I_f by minimizing the loss $L_{Reg\_TAP}(F_f, F_m)$ as follows:

$$L_{Reg\_TAP}(F_f, F_m) = \frac{1}{|\Omega|} \sum_{x \in \Omega} \|F_f(x) - F_w(x)\|^2, \quad (5)$$

which penalizes the discrepancy between (I_m, I'_m) and (I_m, I_f) in feature space. The feature consistency loss $L_{Reg\_TAP}(F_f, F_m)$ refines cross-instance feature alignment for the registration task, and facilitates learning discriminative features and optimizing the registration to generate more reliable pseudo label and spatial prior for improved segmentation.

Figure 1c shows the structure of the TAP that consists of an encoder and a temporal attention unit. Two input images I_a and I_b from an image pair were independently entered into the encoder to generate spatial features f_a

and f_b, respectively. We employed a temporal attention mechanism to generate channel-wise weights W. In particular, we convolved and concatenated features f_a and f_b, and applied average pooling along spatial dimensions followed by 1×1 convolution to generate channel-wise attention W that encodes the relative weight of each channel. f_a and f_b were then multiplied with the resulting channel weights W, and convolved and summed to produce the final cross-instance feature F for image pair (I_a, I_b). The proposed TAP module can be formulated as:

$$W = Conv_{1\times1}(AvgPool(Conv(f_a) \oplus Conv(f_b))), \quad (6)$$

$$F = Conv(f_a \otimes W) \oplus Conv(f_b \otimes W), \quad (7)$$

where $Conv$ and $Conv_{1\times1}$ represent 3 × 3 and 1×1 convolution, respectively, $AvgPool$ indicates average pooling, $\oplus$ and $\otimes$ denote element-wise summation and multiplication, respectively.

We note that temporal attention has also been used in squeeze-and-excitation (SE) networks[34], whereby interdependences between feature channels are explicitly modeled and used to re-weight the channels of a single feature instance. However, channel-wise weights extracted from a single feature instance, as in SE networks, may be sub-optimal due to the intrinsic biases in feature encoding. Compared with SE[34] and related networks, the proposed approach employed temporal attention in a substantially different manner. In particular, two images from an image pair were independently entered into an encoder to extract features subject to similar level of biases; the resulting two feature instances were fused and used to adaptively modulate the channel-wise weights conditioned on the relationships between the two feature instances (Figure 1c), potentially alleviating the encoding biases and resulting in more robust channel weight estimation. Subsequently, the estimated inter-instance channel weights were used to scale and aggregate the two input features to generate the final cross-instance feature for the input image pair. We then enforced consistency between the resulting features as additional constraints to the registration task. This strategy also differs from other commonly used registration models, which typically enforce similarity between two input images directly in image space.

Combining Equations (1), (4), and (5), we can formulate the total registration loss as:

$$Loss_{Reg}(I_m, I_f) = L_{Reg\_baseline}(I_m, I_f) + L_{Reg\_aux}(S'_m, \widehat{P}_f) + L_{Reg\_TAP}(F_f, F_m). \quad (8)$$

ALGORITHM 1 Training Process of the Proposed Semi-supervised Segmentation Algorithm

Input: Reg, Seg_f, Seg_w, TAP; X cine cardiac MRI dataset each containing an ED and an ES frame with a random frame manually labeled.
Output: Trained Seg_f and Reg for segmentation and registration inference

Pre-train Reg with X cardiac cine MRI subjects each with an ED and an ES frame without manual label
Pre-train TAP with X cardiac cine MRI subjects each with an ED and an ES frame without manual label
while *not converge* **do**
  Sample batch data I_m and S_m, and I_f;
  // train the registration network Reg to generate augmented labeled images and pseudo label
  Freeze Seg_f and Seg_w, train Reg to register I_m and S_m to I_f by minimizing $L_{Reg\_baseline}$ in Equation (1) to generate I'_m and S'_m;
  // train the fully supervised segmentation Seg_f
  Freeze Seg_w and Reg, train Seg_f using I'_m and S'_m by minimizing Equation (2) to generate segmentation prediction P'_m;
  // train the weakly supervised segmentation Seg_w
  Freeze Seg_f and Reg, train Seg_w using I_f and S'_m following Equation (3);
  // apply Seg_w to I_f to provide semantic information to Reg
  Freeze Seg_w to segment I_f and generate segmentation prediction P_f and $\widehat{P}_f$, calculate the auxiliary loss L_{Reg_aux} in Equation (4);
  // calculate the TAP loss to aid in optimization of Reg
  Calculate the TAP loss L_{Reg_TAP} in Equation (5);
  // calculate the total registration loss to update Reg
  Activate and update Reg by minimizing Equation (8): $L_{Reg\_baseline} + L_{Reg\_aux} + L_{Reg\_TAP}$;
end

2.5 | Algorithm training process

The registration network, two segmentation networks, and the TAP module were trained sequentially as shown in Algorithm 2.5. In particular, the registration network and the TAP module were pre-trained using all the subjects in the training dataset without manual annotations. For each training epoch, one of the networks/module was activated and the others were

frozen, and the parameters were updated with respect to the corresponding loss function. Upon training convergence, the trained segmentation Seg_f and registration *Reg* modules were employed to segment and register test dataset.

3 | EXPERIMENTS

3.1 | Cine cardiac MRI dataset

3.1.1 | ACDC

The Automatic Cardiac Diagnosis Challenge dataset (ACDC, https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html) consists of 150 subjects recruited at the University Hospital of Dijon (France). These subjects were evenly distributed in five categories of cardiac pathologies, including heart failure with myocardial infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy, abnormal right ventricle, and healthy volunteers.[35] For each subject, 2D short-axis cine images were acquired from the base to the apex at 1.5T or 3.0T (Siemens Aera and Siemens Trio, Siemens Medical Solutions, Germany) under breath-hold conditions using an SSFP sequence (in-plane voxel size = 1.34–1.68 mm^2, slice thickness = 5–10 mm, number of slices = 6–18, 28–40 frames/slice covering a partial or complete cardiac cycle). Manual segmentation of the left ventricular cavity (LV), myocardium (Myo), and right ventricular cavity (RV) in the ED and ES frames was performed by an experienced expert.

3.1.2 | M&Ms challenge

The Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge (M&Ms, https://www.ub.edu/mnms/) dataset consists of 375 subjects, of which 345 cases were scanned by Canon ($n = 50$), Philips ($n = 125$), Siemens ($n = 95$), and GE MR systems ($n = 75$), and were manually segmented for LV, Myo, and RV at the ED and ES frames by an expert clinician.[36] The 125 subjects scanned by Philips systems were recruited at two different clinical centers, and included healthy volunteers and patients with diverse cardiac pathologies, for example, hypertrophic cardiomyopathy, hypertensive heart disease, and dilated cardiomyopathy. Therefore, we used these subjects as our study cohort. For each subject, 2D short-axis cine images were acquired from the base to apex at 1.5T (Philips Archieva, The Neitherlands) with the following parameters: in-plane voxel size = 1.20–1.45 mm^2, slice thickness = 9.9 mm, number of slices = 10–11, 26–30 frames/slice.

For both datasets, participants' consent was collected in the respective studies and was exempted in this work.

3.2 | Algorithm implementation

The 150 subjects in ACDC dataset were divided into 100 subjects for training and 50 cases for testing, both of which were composed of equal numbers of subjects in the five pathology categories. Similarly, the 125 M&Ms subjects were randomly split into 75 and 50 cases for training and testing, respectively. Prior to network training, all the datasets were resampled in *X*–*Y* plane to a voxel size of 1.2 × 1.2 mm^2, normalized to zero mean and unit variance, and cropped/padded to a uniform matrix size of 224 × 224 ×18.

We also investigated the utility of the proposed approach when trained with smaller dataset. In particular, we further randomly selected two and four cases in each of the five cardiac pathology categories from the entire 100 ACDC training subjects, resulting in another two ACDC training datasets of 10 and 20 subjects, respectively. Similarly, we randomly selected 15 and 30 subjects from the entire 75 M&Ms training cases as two additional M&Ms training dataset. Note that both the ACDC and M&Ms dataset originally contain manual segmentation at both the ED and ES frames for each subject. For the three training sessions using different numbers of training subjects in each dataset, we intentionally discarded manual label at random ED or ES frame for each subject for semi-supervised learning. For both datasets, testing was performed on both the ED and ES frames for the same subjects.

The proposed approach was implemented with Python 3.7.1 and Pytorch 1.7.1 framework on an NVIDIA V100l graphics processing unit (GPU, NVIDIA Corp., Santa Clara, CA, USA) provided by Digital Research Alliance of Canada. VoxelMorph[32] was adopted for the deformable registration. 2D Unet with ResNet50 as the backbone of the encoder was employed for the fully and weakly supervised segmentation. As shown in Figure 1 and Algorithm 2.5, the registration network *Reg* and segmentation networks Seg_f and Seg_w were optimized sequentially with the other two fixed during each training iteration with the following parameters: optimizer = ADAM, initial learning rate = 4e-4, polynomial decay rate = 5e-4, batch size = 20, number of epochs = 400. Note that the TAP was pre-trained and kept frozen during the following algorithm training process. During Seg_f and Seg_w training, data augmentation including random rotation (-60∼+60°), scaling (0.5∼1.5 times), translation (-60∼+60 pixels), and intensity variation (0.5∼1.5 times) were performed in parallel.

3.3 | Evaluation methods

Algorithm segmentation was compared with manual delineation using the Dice similarity coefficient (DSC, [0,1]), the average symmetric surface distance (ASSD,

mm), and the Hausdorff distance (HD, mm). Continuous variables were presented as mean ± standard deviation (Mean ± SD).

In addition, we calculated LV end-diastolic volume (LVEDV, ml), end-systolic volume (LVESV, ml), stroke volume (LVSV, ml), ejection fraction (LVEF,%), myocardial mass (LVMM, g), and RV ejection fraction (RVEF,%) using the segmentation provided by our approach and compared these measurements with manual analyses. For LVMM calculation, a density of 1.05 g/mL was used.[37]

3.4 | Comparison with state-of-the-art

We compared the segmentation accuracy of the proposed approach (Ours) with several recently developed semi-supervised segmentation methods based on: 1) pseudo label, for example, CPS,[12] UniMatch,[15] and BCP,[16] 2) consistency regularization, for example, uncertainty aware mean teacher (UAMT),[19] DAN,[21] URPC,[25] CCT,[24] and MambaUnet,[26] and 3) atlases: for example, DeepAtlas[27] and BRBS.[28] In addition, we trained a 2D Unet with the same architecture as Seg_f and Seg_w using only one labeled frame and both labeled frames for each subject in each training dataset. These were used as a lower bound (Unet_LB) and an upper bound (Unet_UB) of the segmentation accuracy, respectively. We also trained the proposed approach using the same training datasets in a fully supervised manner. In particular, for each training epoch, we randomly assigned one of the ED and ES frames as labeled image and the other frame as unlabeled image to train the proposed approach in a semi-supervised manner. We then reversed the assignment to train the proposed approach a second time in the same epoch. This led to a fully supervised version of the proposed approach (Ours_FS) using the complete labeled dataset.

For fair comparison, all the methods were trained and tested on the same subjects and frame(s) without any post-processing.

3.5 | Ablation studies

We investigated the effects of the weakly supervised segmentation, the auxiliary loss as feedback on the registration, and the TAP module in the proposed algorithm framework for both segmentation and registration (Loss = $L_{Reg\_baseline} + L_{Seg\_f} + L_{Seg\_w} + L_{Reg\_aux} + L_{Reg\_TAP}$). This leads to five ablation methods: 1) joint registration and the fully supervised segmentation for warped labeled images only (Loss = $L_{Reg\_baseline} + L_{Seg\_f}$) as a baseline; 2) joint registration and fully and weakly supervised segmentation without the auxiliary loss as feedback on registration (Loss = $L_{Reg\_baseline} + L_{Seg\_f} + L_{Seg\_w}$); 3) joint registration and fully and weakly supervised segmentation with the auxiliary loss as feedback on registration (Loss = $L_{Reg\_baseline} + L_{Seg\_f} + L_{Seg\_w} + L_{Reg\_aux}$); 4) joint registration and fully supervised segmentation for both unlabeled and warped labeled images with the auxiliary loss (Loss = $L_{Reg\_baseline} + L_{Seg\_f} + L_{Reg\_aux}$), which is identical to DeepAtlas; 5) joint registration and fully supervised segmentation for warped labeled images with the TAP (Loss = $L_{Reg\_baseline} + L_{Seg\_f} + L_{Reg\_TAP}$). Note that weights of different terms in the loss function were omitted for simplicity.

3.6 | Statistical analyses

Paired *t*-test was performed to compare: 1) DSC provided by the proposed approach and the other semi-supervised segmentation methods in Section 3.4, and 2) segmentation and registration DSC provided by the proposed approach and the third ablation method in Section 3.5 to evaluate the effectiveness of the TAP module. Normality of data distribution was determined using Shapiro–Wilk test and when data did not satisfy normal distribution, Wilcoxon signed rank test was performed for non-parametric data.

Pearson correlation coefficients (r) were used to determine the relationships between: 1) the final segmentation and registration accuracies for the ablation methods in Section 3.5, and 2) the algorithm and manual cardiac function measurements. For each ablation method, relationships between the segmentation and registration DSC for RV, Myo, and LV were calculated for all the subjects in each test dataset. In addition, we calculated the mean segmentation and registration DSC for RV, Myo, and LV for all the test subjects for each ablation method, and explored the relationships of the mean segmentation and registration DSC across all the ablation algorithms ($n = 6$), including the proposed framework. Agreement between the algorithm and manual cardiac function measurements was evaluated using the Bland–Altman method (bias, 95% limits of agreement [LoA]). All of the statistical analyses were performed with GraphPad Prism v9.5.0 (GraphPad Software Inc., San Diego, CA, USA). Results were considered significant when the probability of making a Type I error was less than 5% ($p<0.05$).

4 | RESULTS

4.1 | ACDC segmentation

Figure 2 shows representative semi-supervised segmentation of a slice of an ACDC test subject provided by various algorithms trained on 10, 20, and 100 ACDC subjects each with a random ED/ES frame labeled. Qualitatively, the proposed approach generated RV, Myo,

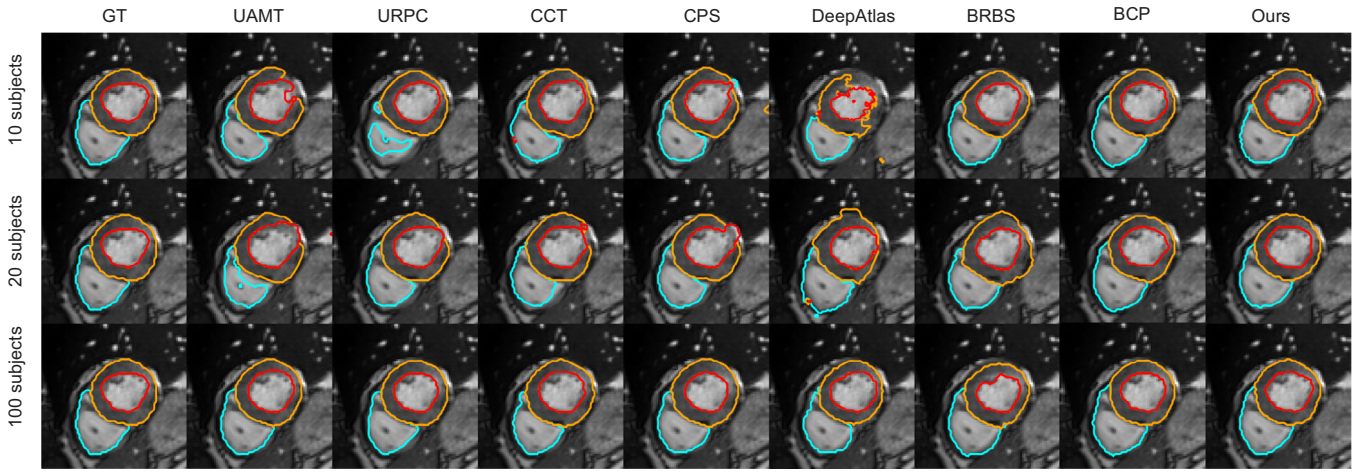

FIGURE 2 Representative segmentation of a slice of an ACDC test subject provided by manual observer (GT) and various algorithms trained on 10 (1st row), 20 (2nd row), and 100 (3rd row) subjects each with a random ED/ES frame labeled. ACDC, automatic cardiac diagnosis challenge; ED, end-diastolic; ES, end-systolic.

and LV segmentation with much less shrinkage, leakage, and more complete coverage and smoother contours in consistency with manual annotation. Of note, the other algorithms appeared to be more sensitive to the number of training subjects and the amount of manual annotations. This was evidenced by the discrepancies between algorithm and manual segmentation when using 10 and 20 training subjects, which were much improved when the number of the training subjects was increased to 100 for the other methods. In contrast, the proposed approach was less sensitive to the size of training subjects and manual annotations, and demonstrated relatively high degree of agreement with manual annotation.

Figure 3 shows the segmentation accuracy on the ACDC dataset ($n = 50$ subjects) provided by various algorithms using 10, 20, and 100 ACDC subjects each with randomly selected ED/ES label for semi-supervised training. For the three training sessions, the proposed approach markedly outperformed Unet_LB, UAMT, URPC, CCT, CPS, and BRBS by providing greater DSC and lower ASSD and HD with generally lower SD for RV, Myo, and LV. Of note, our approach achieved higher segmentation accuracy than BCP and MambaUNet, which further outperformed UniMatch and the widely used DeepAtlas, and yielded segmentation accuracy on par with Unet_UB.

As shown in Table 1, our approach achieved DSC of {0.910±0.063, 0.894±0.024, 0.934±0.056}, ASSD of {1.37±0.063, 1.20±1.12, 1.25±1.63} mm, and HD of {6.38±2.99, 4.67±3.22, 3.97±5.76} mm for {RV, Myo, LV} when trained using 100 subjects. The RV DSC was greater ($p<0.05$) and the Myo and LV DSC were lower ($p<0.05$) than Unet_UB, and these were consistently greater than DeepAtlas ($p<0.05$ for 3/3 of the cases) and BCP ($p<0.05$ for 3/3 of the comparison). Table 2 shows that the performance of all the algorithms, including the proposed approach, was reduced as expected when using 20 training subjects. Compared with BCP, our approach yielded consistent improvements of 0.005–0.009 for DSC ($p<0.05$ for 2/3 of the cases), 0.11–0.016 mm for ASSD and 0.34–1.40 mm for HD. Similar to that using 100 subjects for training as in Table 1, our approach yielded comparable toward somewhat lower DSC ($p<0.05$ for 1/3 of the cases) and higher ASSD and HD for RV, Myo, and LV compared with Unet_UB. Very similar trend was observed when the number of training subjects was reduced to 10 (Table 3). As expected, the segmentation accuracy for all the methods was decreased except for Unet_LB, which surprisingly witnessed improved accuracy. Interestingly, our approach outperformed Unet_UB ($p<0.05$ for 3/3 of the DSC comparison) and yielded the highest accuracies among all the comparative methods ($p<0.05$ for 39/39 of the DSC comparison). For these experiments, BCP yielded the highest accuracies than the other comparative methods. Our approach outperformed BCP in the majority of the cases ($p<0.05$ for 8/9 of the DSC measurements), and achieved similar segmentation accuracies as Unet_UB when using 100 training subjects (Table 1) and higher accuracies when using 10 training cases (Table 3). Notably, the proposed approach trained in a fully supervised manner, that is, Ours_FS, performed similarly well to Ours and Unet_UB when using 100 ACDC subjects for training (Table 1), and much better than the two methods when trained on 20 (Table 2) and 10 (Table 3) ACDC subjects.

4.2 | M&Ms challenge segmentation

Figure 4 illustrates a slice of an M&Ms test subject segmented by various algorithms trained using 15, 30, and

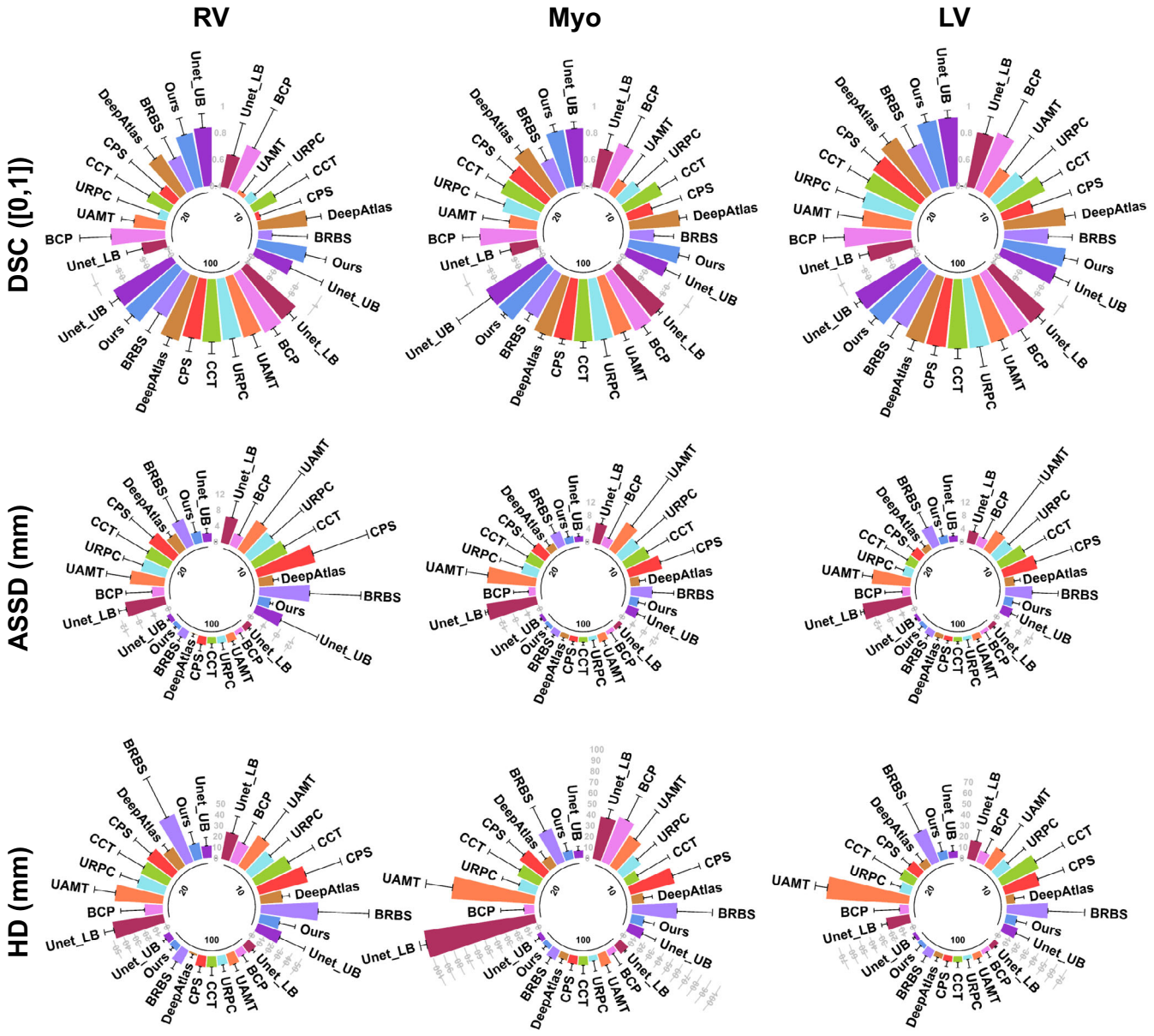

FIGURE 3 ACDC segmentation results ($n = 50$ subjects) provided by various algorithms trained on 10, 20, and 100 ACDC subjects each with a random ED/ES frame labeled. Error bar represents the standard deviation of the data. ACDC, automatic cardiac diagnosis challenge; ED, end-diastolic; ES, end-systolic.

75 M&Ms subjects each with randomly selected ED/ES label for training. Similarly, all the comparative methods experienced various degrees of problematic segmentation, and these issues were alleviated with the increase of training subjects. In contrast, our approach demonstrated very similar segmentation accuracies regardless of the sizes of training datasets.

Figure 5 illustrates the segmentation accuracy on 50 M&Ms test subjects for various algorithms trained using 15, 30, and 75 M&Ms training subjects. For the three training sessions, UniMatch, BCP, MambaUNet, and DeepAtlas generally outperformed Unet_LB, CPS, UAMT, DAN, URPC, CCT, and BRBS. BCP yielded higher DSC and lower ASSD and HD compared with UniMatch and MambaUNet, and higher DSC toward slightly greater ASSD and HD than DeepAtlas. Our approach, in general, noticeably improved DeepAtlas segmentation accuracies and outperformed BCP toward slightly lower accuracies than Unet_UB.

Table 4 shows that, compared with BCP, our approach yielded DSC improvements of 0.008, 0.008, and 0.006 for RV, Myo, and LV, respectively, using 75 training subjects. These improvements were 0.009, 0.007, and 0.006 when using 30 training subjects (Table 5), and were 0.015, 0.010, and 0.007 when using 15 training

TABLE 1 Algorithm segmentation of $n = 50$ ACDC subjects for various methods trained on the same 100 ACDC training subjects each with random ED/ES frames labeled.

Method	RV			Myo			LV		
	DSC([0,1])	ASSD(mm)	HD(mm)	DSC([0,1])	ASSD(mm)	HD(mm)	DSC([0,1])	ASSD(mm)	HD(mm)
Unet_LB	$0.887^{<0.0001}_{0.058}$	$1.91_{0.67}$	$9.28_{2.51}$	$0.861^{<0.0001}_{0.028}$	$1.40_{0.47}$	$10.18_{1.19}$	$0.920^{0.0042}_{0.057}$	$1.36_{0.57}$	$4.89_{1.06}$
CPS	$0.855^{<0.0001}_{0.101}$	$2.73_{2.44}$	$10.32_{5.30}$	$0.862^{0.0003}_{0.038}$	$1.70_{1.23}$	$8.07_{5.87}$	$0.915^{0.0024}_{0.070}$	$1.48_{0.90}$	$5.50_{4.83}$
UniMatch	$0.888^{0.0062}_{0.095}$	$1.78_{0.82}$	$8.30_{4.03}$	$0.872^{0.0049}_{0.085}$	$1.56_{1.46}$	$6.07_{4.18}$	$0.912^{0.0064}_{0.105}$	$1.63_{2.12}$	$5.16_{7.49}$
BCP	$0.902^{0.0435}_{0.060}$	$1.45_{0.60}$	$7.50_{3.20}$	$0.885^{0.0277}_{0.030}$	$1.28_{0.80}$	$5.80_{2.80}$	$0.920^{0.0028}_{0.068}$	$\mathbf{1.15}_{0.40}$	$4.20_{3.30}$
UAMT	$0.874^{<0.0001}_{0.102}$	$2.58_{2.40}$	$11.35_{6.98}$	$0.856^{<0.0001}_{0.067}$	$2.49_{3.70}$	$12.28_{11.94}$	$0.916^{0.0062}_{0.085}$	$1.95_{3.04}$	$5.90_{5.47}$
DAN	$0.849^{<0.0001}_{0.106}$	$2.52_{1.49}$	$11.36_{6.49}$	$0.848^{<0.0001}_{0.062}$	$1.95_{1.97}$	$8.89_{6.99}$	$0.913^{0.0072}_{0.077}$	$1.51_{0.88}$	$5.35_{4.15}$
URPC	$0.860^{<0.0001}_{0.136}$	$2.12_{2.69}$	$8.93_{6.01}$	$0.865^{0.0013}_{0.128}$	$1.34_{2.62}$	$5.13_{9.49}$	$0.922^{0.0154}_{0.150}$	$1.23_{2.75}$	$\mathbf{3.83}_{7.81}$
CCT	$0.872^{<0.0001}_{0.087}$	$2.15_{1.45}$	$9.96_{7.72}$	$0.867^{0.0015}_{0.049}$	$1.62_{1.20}$	$8.33_{7.86}$	$0.921^{0.0049}_{0.076}$	$1.46_{0.91}$	$5.80_{4.66}$
MambaUNet	$0.896^{0.0171}_{0.085}$	$1.64_{0.76}$	$7.66_{3.68}$	$0.877^{0.0126}_{0.075}$	$1.44_{1.34}$	$5.60_{3.86}$	$0.918^{0.0184}_{0.095}$	$1.50_{1.96}$	$4.76_{6.91}$
DeepAtlas	$0.898^{0.0297}_{0.063}$	$1.72_{0.98}$	$7.67_{3.39}$	$0.867^{0.0006}_{0.038}$	$1.46_{0.99}$	$6.30_{4.04}$	$0.919^{0.0093}_{0.077}$	$1.43_{0.90}$	$4.82_{3.35}$
BRBS	$0.771^{<0.0001}_{0.160}$	$2.84_{1.70}$	$14.64_{7.37}$	$0.775^{<0.0001}_{0.056}$	$2.06_{0.83}$	$10.20_{6.22}$	$0.859^{<0.0001}_{0.120}$	$2.20_{1.52}$	$8.42_{6.20}$
Ours	$\mathbf{0.910}_{0.063}$	$\mathbf{1.37}_{0.63}$	$\mathbf{6.38}_{2.99}$	$\mathbf{0.894}_{0.024}$	$\mathbf{1.20}_{1.12}$	$\mathbf{4.67}_{3.22}$	$\mathbf{0.934}_{0.056}$	$1.25_{1.63}$	$3.97_{5.76}$
Ours_FS	$0.906^{<0.0001}_{0.060}$	$1.44_{0.69}$	$6.68_{3.04}$	$0.893^{0.9488}_{0.029}$	$1.12_{0.73}$	$4.61_{3.65}$	$0.940^{<0.0001}_{0.050}$	$1.13_{1.11}$	$3.63_{3.73}$
Unet_UB	$0.905^{<0.0001}_{0.068}$	$1.48_{0.61}$	$6.35_{2.85}$	$0.895^{0.0010}_{0.030}$	$1.05_{0.45}$	$4.40_{3.09}$	$0.941^{<0.0001}_{0.044}$	$1.02_{0.34}$	$3.17_{1.63}$

Note: Results are shown for $\text{Mean}^{p-\text{value}}_{\text{SD}}$ and bold indicates higher performance.
Abbreviations: ACDC, automatic cardiac diagnosis challenge; ED, end-diastolic; ES, end-systolic.

TABLE 2 Algorithm segmentation of $n = 50$ ACDC subjects using various methods trained on the same 20 ACDC training subjects each with random ED/ES frames labeled.

Method	RV			Myo			LV		
	DSC([0,1])	ASSD(mm)	HD(mm)	DSC([0,1])	ASSD(mm)	HD(mm)	DSC([0,1])	ASSD(mm)	HD(mm)
Unet_LB	$0.583^{<0.0001}_{0.120}$	$12.57_{2.91}$	$46.93_{8.86}$	$0.609^{<0.0001}_{0.080}$	$15.79_{1.02}$	$102.46_{5.97}$	$0.737^{<0.0001}_{0.070}$	$15.31_{1.31}$	$21.57_{5.03}$
CPS	$0.543^{<0.0001}_{0.315}$	$9.63_{10.95}$	$28.94_{19.56}$	$0.741^{<0.0001}_{0.151}$	$5.31_{6.25}$	$27.42_{18.82}$	$0.839^{<0.0001}_{0.152}$	$3.40_{2.85}$	$13.42_{8.84}$
UniMatch	$0.785^{<0.0001}_{0.120}$	$4.81_{4.99}$	$19.67_{14.48}$	$0.809^{<0.0001}_{0.085}$	$2.78_{2.75}$	$10.97_{9.31}$	$0.883^{0.0041}_{0.095}$	$3.11_{3.53}$	$10.61_{10.32}$
BCP	$0.805^{0.0372}_{0.095}$	$3.90_{3.55}$	$16.50_{12.00}$	$0.825^{0.0205}_{0.038}$	$2.25_{2.20}$	$9.30_{7.80}$	$0.902^{0.0582}_{0.065}$	$2.55_{2.85}$	$8.50_{8.20}$
UAMT	$0.638^{<0.0001}_{0.234}$	$10.79_{9.77}$	$44.03_{21.87}$	$0.610^{<0.0001}_{0.129}$	$15.33_{5.59}$	$110.95_{21.18}$	$0.769^{<0.0001}_{0.154}$	$12.13_{6.45}$	$75.68_{23.27}$
DAN	$0.440^{<0.0001}_{0.307}$	$10.24_{9.14}$	$32.48_{19.52}$	$0.643^{<0.0001}_{0.200}$	$7.58_{8.20}$	$37.10_{24.70}$	$0.764^{<0.0001}_{0.189}$	$4.54_{3.80}$	$17.77_{11.27}$
URPC	$0.470^{<0.0001}_{0.325}$	$8.00_{7.71}$	$26.07_{17.45}$	$0.677^{<0.0001}_{0.224}$	$5.24_{8.89}$	$16.99_{16.56}$	$0.794^{<0.0001}_{0.218}$	$3.03_{3.21}$	$10.52_{9.22}$
CCT	$0.592^{<0.0001}_{0.288}$	$8.17_{13.34}$	$26.93_{26.59}$	$0.737^{<0.0001}_{0.147}$	$5.29_{9.06}$	$21.86_{33.22}$	$0.823^{<0.0001}_{0.170}$	$3.36_{9.50}$	$14.49_{29.64}$
MambaUNet	$0.792^{0.0038}_{0.110}$	$4.44_{4.61}$	$18.13_{13.35}$	$0.817^{0.0057}_{0.075}$	$2.57_{2.54}$	$10.13_{8.60}$	$0.892^{0.0162}_{0.085}$	$2.87_{3.26}$	$9.79_{9.52}$
DeepAtlas	$0.747^{<0.0001}_{0.110}$	$5.94_{1.96}$	$20.19_{6.78}$	$0.807^{0.0037}_{0.055}$	$2.39_{2.26}$	$10.97_{8.02}$	$0.897^{0.0317}_{0.078}$	$\mathbf{2.01}_{2.23}$	$\mathbf{7.81}_{8.45}$
BRBS	$0.672^{<0.0001}_{0.174}$	$8.98_{9.95}$	$46.77_{40.35}$	$0.655^{<0.0001}_{0.106}$	$4.79_{3.41}$	$32.90_{24.51}$	$0.772^{<0.0001}_{0.173}$	$6.72_{6.30}$	$32.49_{28.27}$
Ours	$\mathbf{0.810}_{0.129}$	$\mathbf{3.70}_{3.84}$	$\mathbf{15.10}_{11.10}$	$\mathbf{0.834}_{0.056}$	$\mathbf{2.14}_{2.12}$	$\mathbf{8.44}_{7.17}$	$\mathbf{0.909}_{0.062}$	$2.39_{2.72}$	$8.16_{7.93}$
Ours_FS	$0.859^{<0.0001}_{0.091}$	$2.33_{1.55}$	$10.30_{6.60}$	$0.861^{<0.0001}_{0.053}$	$1.46_{0.76}$	$6.54_{4.47}$	$0.919^{0.0086}_{0.061}$	$1.65_{1.29}$	$5.89_{5.46}$
Unet_UB	$0.837^{<0.0001}_{0.107}$	$2.73_{1.69}$	$11.50_{7.38}$	$0.831^{0.0850}_{0.051}$	$1.73_{1.03}$	$7.06_{3.79}$	$0.911^{0.0569}_{0.058}$	$1.88_{1.02}$	$6.46_{5.13}$

Note: Results are shown for $\text{Mean}^{p-\text{value}}_{\text{SD}}$ and bold indicates higher performance.
Abbreviations: ACDC, automatic cardiac diagnosis challenge; ED, end-diastolic; ES, end-systolic.

subjects (Table 6). For the three training sessions, our approach yielded accuracies that were approximately 0.01 lower in DSC ($p<0.05$ for 3/9 of the cases), 0.3 mm higher in ASSD, and 1.3 mm higher in HD, compared with Unet_UB. Although the segmentation accuracies for all the algorithms were decreased when the number of training subjects was reduced from 75 to 15, BCP, DeepAtlas, and our approach exhibited lower decline than the other methods and our approach outperformed BCP ($p<0.05$ for 2/9 of the DSC comparison) and DeepAtlas ($p<0.05$ for 9/9 of the DSC comparison) in general. For these cases, the proposed approach trained in a fully supervised manner, that is, Ours_FS, yielded somewhat overall comparable accuracies to Unet_UB and slightly higher performance than Ours.

TABLE 3 Algorithm segmentation of $n = 50$ ACDC subjects using various methods trained on the same 10 ACDC training subjects each with random ED/ES frames labeled.

Method	RV			Myo			LV		
	DSC([0,1])	ASSD(mm)	HD(mm)	DSC([0,1])	ASSD(mm)	HD(mm)	DSC([0,1])	ASSD(mm)	HD(mm)
Unet_LB	$0.649^{<0.0001}_{0.251}$	$8.44_{6.38}$	$25.23_{16.03}$	$0.692^{<0.0001}_{0.122}$	$6.32_{2.15}$	$39.45_{8.30}$	$0.815^{<0.0001}_{0.135}$	$3.87_{2.42}$	$17.75_{10.59}$
CPS	$0.434^{<0.0001}_{0.322}$	$18.71_{21.94}$	$46.00_{34.41}$	$0.594^{<0.0001}_{0.234}$	$10.71_{13.36}$	$41.74_{26.29}$	$0.640^{<0.0001}_{0.307}$	$10.92_{14.60}$	$32.80_{24.94}$
UniMatch	$0.750^{0.0023}_{0.158}$	$5.02_{5.28}$	$24.31_{19.10}$	$0.764^{0.0063}_{0.108}$	$3.65_{3.46}$	$15.55_{12.70}$	$0.848^{0.0076}_{0.123}$	$3.60_{3.42}$	$13.34_{12.16}$
BCP	$0.758^{0.0094}_{0.155}$	$4.19_{4.11}$	$20.00_{15.00}$	$0.771^{0.0143}_{0.082}$	$3.00_{2.80}$	$12.50_{10.00}$	$0.855^{0.0384}_{0.108}$	$3.00_{2.80}$	$11.00_{9.80}$
UAMT	$0.431^{<0.0001}_{0.327}$	$10.94_{13.62}$	$33.30_{21.50}$	$0.533^{<0.0001}_{0.249}$	$10.14_{14.80}$	$33.27_{24.79}$	$0.629^{<0.0001}_{0.300}$	$7.12_{7.86}$	$20.33_{13.64}$
DAN	$0.423^{<0.0001}_{0.322}$	$19.04_{24.32}$	$46.40_{33.13}$	$0.508^{<0.0001}_{0.249}$	$15.97_{22.73}$	$44.10_{32.25}$	$0.582^{<0.0001}_{0.325}$	$16.4_{24.86}$	$36.40_{29.71}$
URPC	$0.479^{<0.0001}_{0.347}$	$9.91_{15.33}$	$26.07_{22.59}$	$0.585^{<0.0001}_{0.288}$	$7.23_{14.51}$	$20.84_{21.99}$	$0.678^{<0.0001}_{0.326}$	$6.43_{14.66}$	$15.74_{20.38}$
CCT	$0.593^{<0.0001}_{0.288}$	$11.41_{13.34}$	$36.27_{26.59}$	$0.709^{<0.0001}_{0.147}$	$8.29_{9.06}$	$44.04_{33.22}$	$0.776^{<0.0001}_{0.170}$	$9.00_{9.50}$	$37.37_{29.64}$
MambaUNet	$0.752^{0.0031}_{0.148}$	$4.63_{4.87}$	$22.44_{17.63}$	$0.762^{0.0055}_{0.098}$	$3.37_{3.19}$	$14.35_{11.72}$	$0.851^{0.0132}_{0.113}$	$3.32_{3.16}$	$12.31_{11.22}$
DeepAtlas	$0.768^{0.0176}_{0.110}$	$4.57_{1.96}$	$20.19_{6.78}$	$0.779^{0.0283}_{0.055}$	$2.90_{2.26}$	$12.66_{10.44}$	$0.862^{0.0077}_{0.078}$	$3.01_{2.23}$	$11.59_{11.30}$
BRBS	$0.504^{<0.0001}_{0.263}$	$15.85_{19.30}$	$52.47_{46.62}$	$0.582^{<0.0001}_{0.147}$	$6.79_{8.11}$	$41.13_{32.43}$	$0.729^{<0.0001}_{0.180}$	$8.49_{9.77}$	$37.69_{33.73}$
Ours	$\mathbf{0.773}_{0.153}$	$\mathbf{3.86}_{4.06}$	$\mathbf{18.70}_{14.70}$	$\mathbf{0.786}_{0.078}$	$\mathbf{2.81}_{2.66}$	$\mathbf{12.00}_{9.77}$	$\mathbf{0.870}_{0.103}$	$\mathbf{2.77}_{2.63}$	$\mathbf{10.30}_{9.35}$
Ours_FS	$0.842^{<0.0001}_{0.082}$	$2.84_{1.78}$	$11.98_{7.17}$	$0.835^{<0.0001}_{0.054}$	$2.08_{1.65}$	$9.05_{7.81}$	$0.902^{<0.0001}_{0.069}$	$2.33_{2.84}$	$8.15_{9.20}$
Unet_UB	$0.696^{<0.0001}_{0.262}$	$8.78_{15.37}$	$23.11_{25.06}$	$0.726^{<0.0001}_{0.117}$	$3.69_{3.13}$	$14.57_{13.12}$	$0.841^{<0.0001}_{0.124}$	$3.42_{2.65}$	$13.66_{12.77}$

Note: Results are shown for $\text{Mean}^{p-\text{value}}_{\text{SD}}$ and bold indicates higher performance.
Abbreviations: ACDC, automatic cardiac diagnosis challenge; ED, end-diastolic; ES, end-systolic.

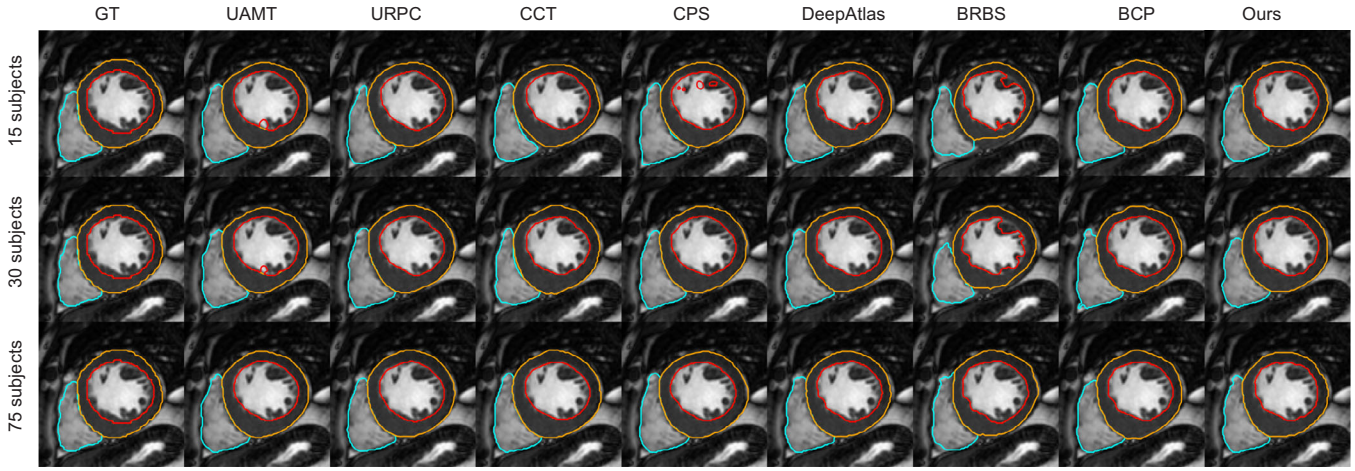

FIGURE 4 Representative segmentation of a slice of an M&Ms test subject provided by manual observer (GT) and various algorithms trained on 15 (1st row), 30 (2nd row), and 75 (3rd row) subjects each with a random ED/ES frame labeled. ED, end-diastolic; ES, end-systolic.

4.3 | Cardiac function measurements

For our approach trained using 100, 20, and 10 ACDC subjects, supplementary Figures S1–S3 show that there were strong ($r>0.89$) and significant ($p<0.0001$) correlations between algorithm and manual LV function measurements. Likewise, algorithm LV function measurements were strongly ($r>0.88$) and significantly ($p<0.0001$) correlated with manual analyses when using 75, 30, and 15 M&Ms subjects for training, as shown in supplementary Figures S4–S6. For both datasets, algorithm RVEF measurements were correlated ($r>0.72$, $p<0.0001$) with manual measurements except for algorithm training using 20 ACDC and 10 M&Ms subjects, which exhibited weak but significant correlations (ACDC: $r = 0.384, p = 0.0056$; M&Ms: $r = 0.390, p<0.0001$) and relatively low degree of agreement (ACDC: bias = -26.13%, 95%LoA = [-136.3%, 84.07%]; M&Ms: bias = -10.44%, 95%LoA = [-78.72%, 57.84%]). In general, Bland–Altman analyses indicated that there was promising agreement between algorithm and manual LVEDV, LVESV, LVSV, LVEF, LVMM, and RVEF measurements, with greater variances when the numbers of training subjects were reduced.

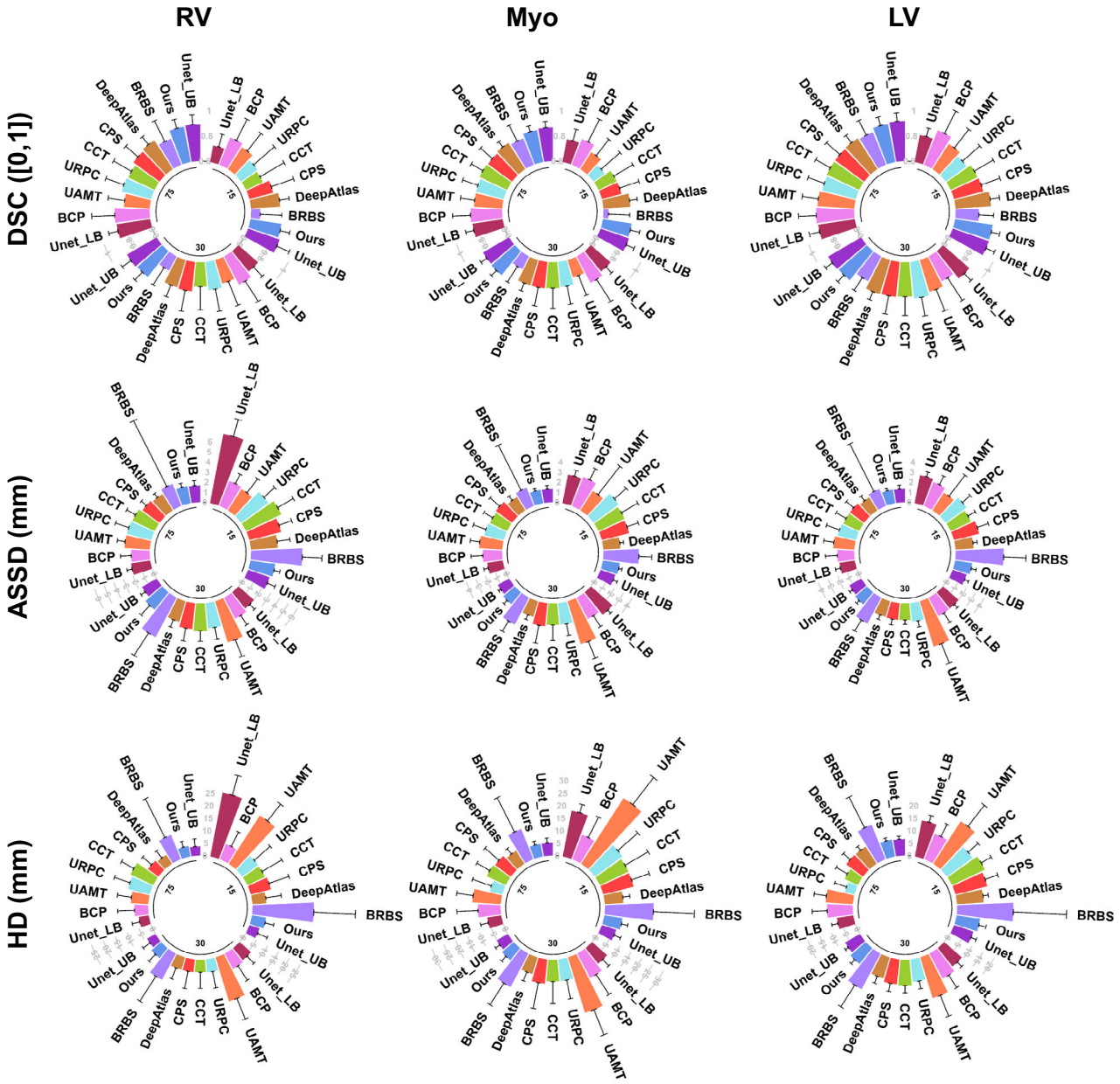

FIGURE 5 M&Ms segmentation results ($n = 50$ subjects) provided by various algorithms trained on 15, 30, and 75 M&Ms subjects each with a random ED/ES frame labeled. Error bar represents the standard deviation of the data. ED, end-diastolic; ES, end-systolic.

4.4 | Effects of Seg_w, TAP, and L_{Reg_aux} on segmentation and registration

Table 7 reveals the impact of various components on segmentation for different ablation methods trained on 20 ACDC and 30 M&Ms subjects. For the ACDC test dataset, Seg_w demonstrated effectiveness by improving the baseline ($Reg + Seg_f$) DSC by 0.056, 0.009, and -0.001 for RV, Myo, and LV, respectively. Likewise, these improvements were 0.007, 0.015, 0.007 for the M&Ms dataset. TAP ($Reg + Seg_f + TAP$) boosted the baseline DSC by 0.047, 0.005, and 0.006 for the ACDC cases, and by 0.027, 0.020, and 0.009 for the M&Ms subjects. Compared with $Reg + Seg_f + Seg_w + L_{Reg\_aux}$, exclusion of the auxiliary loss steadily and moderately reduced DSC by 0.004 to 0.017 for the ACDC and M&Ms dataset, and these were 0.003 to 0.052 when Seg_w was excluded, that is, DeepAtlas. Integration of the TAP module yielded similar effects as the Seg_w toward slightly greater overall DSC for the three heart structures. Among these combinations, the highest DSC was achieved when Seg_w, TAP, and L_{Reg_aux} were utilized.

Figure 6 illustrates the registration of a fixed and a moving image from an ACDC and an M&Ms challenge

TABLE 4 Algorithm segmentation of $n = 50$ M&Ms subjects using various methods trained on the same 75 M&Ms training subjects each with random ED/ES frames labeled.

	RV			Myo			LV		
Method	**DSC([0,1])**	**ASSD(mm)**	**HD(mm)**	**DSC([0,1])**	**ASSD(mm)**	**HD(mm)**	**DSC([0,1])**	**ASSD(mm)**	**HD(mm)**
Unet_LB	$0.867^{0.0083}_{0.074}$	$1.91_{1.01}$	$\mathbf{4.05}_{3.61}$	$0.840^{0.0049}_{0.064}$	$1.55_{0.83}$	$5.64_{3.68}$	$0.891^{0.0074}_{0.082}$	$1.54_{0.97}$	$6.66_{2.91}$
CPS	$0.855^{<0.0001}_{0.127}$	$2.08_{1.31}$	$5.09_{5.60}$	$0.834^{<0.0001}_{0.125}$	$1.97_{2.49}$	$7.76_{8.86}$	$0.883^{0.0063}_{0.140}$	$1.74_{1.88}$	$8.25_{4.33}$
UniMatch	$0.865^{<0.0001}_{0.107}$	$2.01_{1.10}$	$5.51_{3.95}$	$0.836^{<0.0001}_{0.099}$	$1.58_{0.80}$	$6.18_{2.20}$	$0.889^{0.0192}_{0.098}$	$1.64_{0.59}$	$8.23_{1.63}$
BCP	$0.875^{0.0687}_{0.075}$	$1.82_{0.85}$	$4.60_{3.10}$	$0.848^{0.0623}_{0.050}$	$1.42_{0.55}$	$5.50_{2.00}$	$0.900^{0.0694}_{0.072}$	$1.45_{0.55}$	$7.20_{1.50}$
UAMT	$0.803^{<0.0001}_{0.181}$	$2.56_{2.09}$	$5.45_{8.03}$	$0.823^{<0.0001}_{0.142}$	$1.90_{2.31}$	$8.98_{10.11}$	$0.890^{0.0062}_{0.139}$	$1.67_{2.34}$	$10.21_{7.15}$
DAN	$0.832^{<0.0001}_{0.131}$	$2.49_{1.55}$	$6.97_{6.73}$	$0.790^{<0.0001}_{0.112}$	$2.28_{1.86}$	$11.29_{7.84}$	$0.860^{<0.0001}_{0.128}$	$2.08_{2.00}$	$10.81_{5.93}$
URPC	$0.839^{<0.0001}_{0.150}$	$2.47_{2.75}$	$8.93_{7.81}$	$0.825^{<0.0001}_{0.128}$	$1.98_{2.62}$	$\mathbf{5.13}_{9.49}$	$0.873^{<0.0001}_{0.136}$	$1.84_{2.69}$	$\mathbf{3.83}_{6.01}$
CCT	$0.829^{<0.0001}_{0.163}$	$2.40_{2.00}$	$9.96_{6.67}$	$0.839^{0.0079}_{0.105}$	$1.85_{2.02}$	$8.33_{7.10}$	$0.882^{0.0075}_{0.132}$	$1.73_{2.23}$	$5.80_{6.14}$
MambaUNet	$0.864^{0.0175}_{0.097}$	$2.06_{0.95}$	$5.09_{3.41}$	$0.837^{0.0095}_{0.089}$	$1.61_{0.58}$	$6.70_{2.64}$	$0.887^{0.0081}_{0.088}$	$1.78_{0.92}$	$8.12_{2.20}$
DeepAtlas	$0.869^{0.0291}_{0.077}$	$1.91_{0.96}$	$4.27_{4.22}$	$0.845^{0.0347}_{0.052}$	$1.55_{1.02}$	$6.19_{3.28}$	$0.889^{0.0085}_{0.071}$	$1.51_{1.12}$	$8.14_{2.25}$
BRBS	$0.821^{<0.0001}_{0.181}$	$2.33_{14.51}$	$10.48_{20.84}$	$0.824^{<0.0001}_{0.139}$	$1.92_{9.93}$	$12.98_{18.39}$	$0.876^{0.0048}_{0.129}$	$1.79_{10.49}$	$14.87_{16.46}$
Ours	$\mathbf{0.883}_{0.072}$	$\mathbf{1.72}_{0.79}$	$4.25_{2.83}$	$\mathbf{0.856}_{0.046}$	$\mathbf{1.34}_{0.48}$	$5.15_{1.83}$	$\mathbf{0.906}_{0.068}$	$\mathbf{1.37}_{0.49}$	$6.86_{1.36}$
Ours_FS	$0.887^{0.0693}_{0.074}$	$1.63_{0.75}$	$6.54_{3.02}$	$0.864^{0.0472}_{0.046}$	$1.34_{0.62}$	$5.18_{2.37}$	$0.915^{0.0832}_{0.074}$	$1.40_{0.77}$	$4.05_{1.87}$
Unet_UB	$0.889^{0.0702}_{0.079}$	$1.65_{1.01}$	$3.62_{3.66}$	$0.865^{0.0664}_{0.048}$	$1.31_{0.68}$	$5.05_{2.33}$	$0.914^{0.0425}_{0.069}$	$1.32_{0.735}$	$6.54_{1.53}$

Note: Results are shown for $\text{Mean}^{p-\text{value}}_{\text{SD}}$ and bold indicates higher performance.
Abbreviations: ED, end-diastolic; ES, end-systolic.

TABLE 5 Algorithm segmentation of $n = 50$ M&Ms subjects using various methods trained on the same 30 M&Ms training subjects each with random ED/ES frames labeled.

	RV			Myo			LV		
Method	**DSC([0,1])**	**ASSD(mm)**	**HD(mm)**	**DSC([0,1])**	**ASSD(mm)**	**HD(mm)**	**DSC([0,1])**	**ASSD(mm)**	**HD(mm)**
Unet_LB	$0.806^{<0.0001}_{0.101}$	$\mathbf{2.04}_{1.34}$	$6.14_{4.64}$	$0.806^{<0.0001}_{0.063}$	$2.83_{0.78}$	$7.98_{4.08}$	$0.876^{<0.0001}_{0.076}$	$2.04_{0.84}$	$9.36_{2.88}$
CPS	$0.832^{0.0033}_{0.167}$	$2.57_{2.54}$	$5.16_{8.29}$	$0.808^{<0.0001}_{0.149}$	$2.28_{3.56}$	$9.74_{12.54}$	$0.876^{<0.0001}_{0.145}$	$1.65_{1.15}$	$10.00_{3.97}$
UniMatch	$0.843^{0.0073}_{0.101}$	$2.86_{3.27}$	$7.92_{6.90}$	$0.828^{0.0086}_{0.082}$	$1.96_{1.30}$	$8.02_{4.67}$	$0.879^{0.0090}_{0.083}$	$2.00_{1.19}$	$10.53_{3.50}$
BCP	$0.852^{0.0572}_{0.088}$	$2.38_{2.63}$	$5.50_{4.70}$	$0.836^{0.0847}_{0.060}$	$1.65_{1.05}$	$6.50_{3.20}$	$0.888^{0.0624}_{0.075}$	$1.65_{0.80}$	$8.50_{2.70}$
UAMT	$0.793^{<0.0001}_{0.190}$	$2.71_{2.18}$	$6.87_{8.25}$	$0.744^{<0.0001}_{0.206}$	$2.38_{2.54}$	$11.02_{11.35}$	$0.821^{<0.0001}_{0.211}$	$2.06_{2.17}$	$11.32_{7.05}$
DAN	$0.732^{<0.0001}_{0.203}$	$4.23_{3.56}$	$18.31_{10.60}$	$0.665^{<0.0001}_{0.166}$	$4.42_{3.58}$	$23.12_{12.94}$	$0.748^{<0.0001}_{0.182}$	$4.66_{4.07}$	$16.98_{10.70}$
URPC	$0.822^{<0.0001}_{0.178}$	$2.48_{2.35}$	$5.25_{7.91}$	$0.797^{<0.0001}_{0.147}$	$2.14_{2.30}$	$8.66_{9.13}$	$0.861^{<0.0001}_{0.149}$	$1.82_{1.21}$	$9.05_{3.74}$
CCT	$0.792^{<0.0001}_{0.185}$	$2.70_{2.03}$	$\mathbf{5.03}_{7.67}$	$0.804^{<0.0001}_{0.159}$	$2.12_{3.06}$	$8.71_{10.61}$	$0.871^{<0.0001}_{0.148}$	$1.62_{0.88}$	$10.43_{3.40}$
MambaUNet	$0.841^{0.0159}_{0.091}$	$2.64_{3.02}$	$6.11_{5.31}$	$0.823^{<0.0001}_{0.072}$	$1.81_{1.08}$	$7.40_{3.61}$	$0.874^{<0.0001}_{0.073}$	$1.85_{0.83}$	$9.72_{3.00}$
DeepAtlas	$0.824^{0.0024}_{0.098}$	$2.11_{1.53}$	$5.16_{5.62}$	$0.807^{<0.0001}_{0.065}$	$1.60_{0.94}$	$6.95_{4.48}$	$0.880^{0.0118}_{0.067}$	$1.60_{0.91}$	$8.78_{3.47}$
BRBS	$0.727^{<0.0001}_{0.160}$	$4.41_{5.25}$	$11.88_{14.07}$	$0.715^{<0.0001}_{0.145}$	$3.02_{2.33}$	$15.14_{12.27}$	$0.820^{<0.0001}_{0.118}$	$2.90_{2.49}$	$16.08_{12.28}$
Ours	$\mathbf{0.861}_{0.084}$	$2.20_{2.52}$	$5.15_{4.37}$	$\mathbf{0.843}_{0.056}$	$\mathbf{1.51}_{0.90}$	$\mathbf{6.17}_{3.01}$	$\mathbf{0.894}_{0.071}$	$\mathbf{1.54}_{0.69}$	$\mathbf{8.10}_{2.50}$
Ours_FS	$0.867^{0.0852}_{0.095}$	$2.40_{4.01}$	$9.43_{10.20}$	$0.853^{0.0174}_{0.052}$	$1.41_{0.65}$	$5.94_{2.91}$	$0.902^{0.0405}_{0.073}$	$1.39_{0.64}$	$4.32_{2.13}$
Unet_UB	$0.870^{0.0531}_{0.083}$	$1.66_{0.97}$	$3.78_{3.94}$	$0.837^{0.0703}_{0.056}$	$1.28_{0.45}$	$5.10_{2.44}$	$0.901^{0.0625}_{0.065}$	$1.35_{0.63}$	$6.83_{1.70}$

Note: Results are shown for $\text{Mean}^{p-\text{value}}_{\text{SD}}$ and bold indicates higher performance.
Abbreviations: ED, end-diastolic; ES, end-systolic.

subject provided by various combinations of L_{Reg_aux}, Seg_w, and TAP with the baseline model, which consists of *Reg* and Seg_f only. Table 8 shows the effects of Seg_w, TAP, and L_{Reg_aux} on ED and ES frames registration when these networks were trained on the same 20 ACDC and 30 M&Ms subjects as that in Table 7. Very similar trend was observed as for the impact of these modules on the segmentation task. For example, incorporation of Seg_w dramatically improved the baseline (*Reg* + Seg_f) registration DSC by 0.038, 0.071, 0.72 for RV, Myo, and LV, respectively, for the ACDC dataset, and these were 0.021, 0.005, and 0.015 for the M&Ms dataset. The registration accuracy dropped moderately when the feedback from the weakly supervised segmentation on the registration was disconnected, that is, without L_{Reg_aux}. The TAP module on top of the baseline

TABLE 6 Algorithm segmentation of $n = 50$ M&Ms subjects using various methods trained on the same 15 M&Ms training subjects each with random ED/ES frames labeled.

Method	RV			Myo			LV		
	DSC([0,1])	ASSD(mm)	HD(mm)	DSC([0,1])	ASSD(mm)	HD(mm)	DSC([0,1])	ASSD(mm)	HD(mm)
Unet_LB	$0.732^{<0.0001}_{0.102}$	$6.95_{3.45}$	$25.87_{9.65}$	$0.782^{<0.0001}_{0.093}$	$2.86_{1.13}$	$18.56_{4.03}$	$0.816^{<0.0001}_{0.068}$	$2.80_{0.95}$	$14.64_{3.25}$
CPS	$0.793^{<0.0001}_{0.185}$	$3.21_{2.64}$	$8.23_{10.02}$	$0.758^{<0.0001}_{0.161}$	$2.75_{2.33}$	$12.45_{8.94}$	$0.837^{<0.0001}_{0.156}$	$2.55_{2.29}$	$13.06_{6.34}$
UniMatch	$0.832^{0.0248}_{0.091}$	$2.83_{1.65}$	$8.55_{4.35}$	$0.813^{0.0079}_{0.082}$	$1.94_{1.25}$	$7.95_{3.85}$	$0.873^{0.0083}_{0.083}$	$2.25_{1.35}$	$10.45_{4.25}$
BCP	$0.831^{0.0204}_{0.082}$	$2.65_{1.30}$	$6.00_{4.50}$	$0.818^{0.0242}_{0.080}$	$1.80_{0.70}$	$7.70_{3.30}$	$0.892^{0.0793}_{0.071}$	$1.85_{1.10}$	$10.02_{3.70}$
UAMT	$0.805^{<0.0001}_{0.185}$	$3.32_{2.08}$	$6.53_{7.65}$	$0.759^{<0.0001}_{0.197}$	$3.07_{2.44}$	$10.51_{10.91}$	$0.836^{<0.0001}_{0.196}$	$2.53_{1.89}$	$10.95_{6.53}$
DAN	$0.753^{<0.0001}_{0.175}$	$2.49_{3.47}$	$22.94_{10.13}$	$0.644^{<0.0001}_{0.162}$	$2.28_{7.04}$	$31.02_{16.35}$	$0.723^{<0.0001}_{0.185}$	$2.08_{5.20}$	$20.19_{12.79}$
URPC	$0.759^{<0.0001}_{0.217}$	$3.32_{2.87}$	$7.72_{10.20}$	$0.708^{<0.0001}_{0.197}$	$3.07_{3.12}$	$13.64_{12.75}$	$0.803^{<0.0001}_{0.182}$	$2.53_{2.05}$	$12.62_{6.43}$
CCT	$0.745^{<0.0001}_{0.219}$	$3.83_{3.53}$	$7.69_{10.96}$	$0.757^{<0.0001}_{0.175}$	$2.83_{3.75}$	$12.56_{12.74}$	$0.833^{<0.0001}_{0.173}$	$2.19_{1.65}$	$14.43_{5.47}$
MambaUNet	$0.826^{0.0084}_{0.081}$	$2.95_{1.45}$	$7.25_{3.35}$	$0.808^{0.0085}_{0.072}$	$2.02_{1.04}$	$8.71_{3.63}$	$0.871^{0.0048}_{0.073}$	$2.05_{1.18}$	$11.34_{4.50}$
DeepAtlas	$0.830^{0.0176}_{0.077}$	$2.65_{2.54}$	$\mathbf{5.53}_{7.45}$	$0.812^{0.0143}_{0.059}$	$\mathbf{1.68}_{0.72}$	$7.39_{3.32}$	$0.879^{0.0126}_{0.074}$	$1.78_{0.78}$	$10.63_{2.53}$
BRBS	$0.673^{<0.0001}_{0.170}$	$5.05_{3.88}$	$24.41_{19.97}$	$0.639^{<0.0001}_{0.167}$	$3.49_{2.55}$	$19.35_{19.58}$	$0.783^{<0.0001}_{0.116}$	$4.73_{4.40}$	$22.39_{25.49}$
Ours	$\mathbf{0.846}_{0.077}$	$\mathbf{2.46}_{1.16}$	$5.68_{4.23}$	$\mathbf{0.828}_{0.075}$	$\mathbf{1.68}_{0.62}$	$\mathbf{7.26}_{3.03}$	$\mathbf{0.899}_{0.067}$	$\mathbf{1.71}_{0.98}$	$\mathbf{9.45}_{3.47}$
Ours_FS	$0.865^{0.0095}_{0.083}$	$2.26_{1.74}$	$8.87_{5.24}$	$0.841^{0.0086}_{0.053}$	$1.49_{0.71}$	$6.28_{3.74}$	$0.899^{0.5000}_{0.071}$	$1.56_{0.95}$	$5.09_{3.86}$
Unet_UB	$0.862^{0.0165}_{0.106}$	$2.24_{1.65}$	$4.20_{5.02}$	$0.836^{0.0244}_{0.066}$	$1.54_{0.83}$	$6.02_{3.27}$	$0.901^{0.2489}_{0.072}$	$1.46_{0.81}$	$8.11_{2.58}$

Note: Results are shown for $\text{Mean}^{p-\text{value}}_{\text{SD}}$ and bold indicates higher performance.
Abbreviations: ED, end-diastolic; ES, end-systolic.

TABLE 7 Effects of the major components on algorithm segmentation using 20 ACDC and 30 M&Ms challenge subjects for training.

Components			ACDC ($n = 50$)			M&Ms ($n = 50$)		
Seg_w	L_{Reg_aux}	TAP	RV	Myo	LV	RV	Myo	LV
			$0.747_{0.110}$	$0.807_{0.55}$	$0.897_{0.078}$	$0.824_{0.098}$	$0.807_{0.065}$	$0.880_{0.067}$
✓			$0.803_{0.092}$	$0.816_{0.072}$	$0.896_{0.056}$	$0.831_{0.078}$	$0.822_{0.094}$	$0.887_{0.104}$
✓	✓		$0.808^{0.036}_{0.052}$	$0.823^{<0.0001}_{0.089}$	$0.902^{0.0009}_{0.064}$	$0.848^{0.0004}_{0.082}$	$0.830^{<0.0001}_{0.078}$	$0.891^{0.6606}_{0.080}$
	✓		$0.756_{0.125}$	$0.812_{0.093}$	$0.899_{0.078}$	$0.828_{0.118}$	$0.810_{0.106}$	$0.881_{0.132}$
		✓	$0.794_{0.098}$	$0.812_{0.085}$	$0.903_{0.075}$	$0.851_{0.108}$	$0.827_{0.088}$	$0.889_{0.093}$
✓	✓	✓	$\mathbf{0.810}_{0.129}$	$\mathbf{0.834}_{0.056}$	$\mathbf{0.909}_{0.062}$	$\mathbf{0.861}_{0.084}$	$\mathbf{0.843}_{0.056}$	$\mathbf{0.894}_{0.071}$

Note: The baseline model consists of the registration and the fully supervised segmentation for warped labeled images only. Results are shown for $\text{Mean}^{p-\text{value}}_{\text{SD}}$ and bold indicates higher performance.
Abbreviations: ACDC, automatic cardiac diagnosis challenge; TAP, temporal attention perceiver.

TABLE 8 Effects of the major components on ED and ES frame registration using 20 ACDC and 30 M&Ms challenge subjects for training.

Components			ACDC (n = 50)			M&Ms (n = 50)		
Seg_w	L_{Reg_aux}	TAP	RV	Myo	LV	RV	Myo	LV
			$0.714_{0.182}$	$0.617_{0.256}$	$0.775_{0.132}$	$0.638_{0.198}$	$0.630_{0.153}$	$0.798_{0.128}$
✓			$0.753_{0.107}$	$0.666_{0.129}$	$0.831_{0.112}$	$0.631_{0.125}$	$0.626_{0.184}$	$0.791_{0.108}$
✓	✓		$0.752^{<0.0001}_{0.092}$	$0.688^{0.0337}_{0.105}$	$0.847^{<0.0001}_{0.082}$	$0.659^{0.0858}_{0.161}$	$0.635^{0.0059}_{0.168}$	$0.813^{0.0361}_{0.080}$
	✓		$0.731_{0.125}$	$0.630_{0.132}$	$0.796_{0.114}$	$0.639_{0.109}$	$0.628_{0.131}$	$0.804_{0.082}$
		✓	$0.760_{0.090}$	$0.677_{0.138}$	$0.849_{0.095}$	$0.657_{0.112}$	$0.636_{0.164}$	$0.817_{0.077}$
✓	✓	✓	$\mathbf{0.769}_{0.102}$	$\mathbf{0.692}_{0.113}$	$\mathbf{0.854}_{0.095}$	$\mathbf{0.662}_{0.128}$	$\mathbf{0.642}_{0.152}$	$\mathbf{0.825}_{0.088}$
TransMorph			$0.722_{0.114}$	$0.639_{0.103}$	$0.807_{0.130}$	$0.644_{0.127}$	$0.635_{0.133}$	$0.816_{0.107}$

Note: The baseline model consists of the registration and the fully supervised segmentation for warped labeled images only. Results are shown for $\text{Mean}^{p-\text{value}}_{\text{SD}}$ and bold indicates higher performance.
Abbreviations: ACDC, automatic cardiac diagnosis challenge; ED, end-diastolic; ES, end-systolic; TAP, temporal attention perceiver.

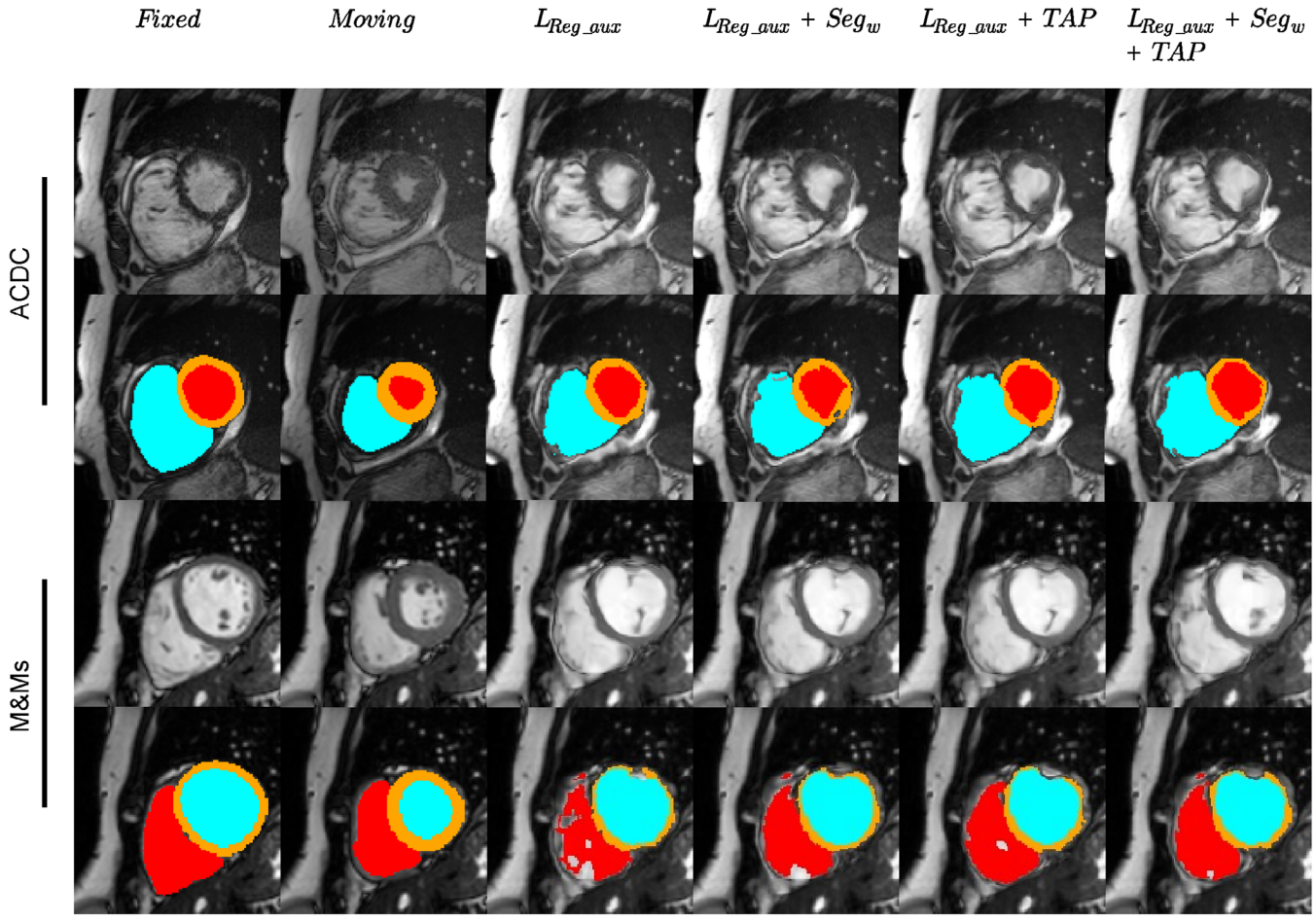

FIGURE 6 Illustration of ED and ES frame registration for an ACDC and an M&Ms subject using various ablation methods. The baseline model contains the registration and the fully supervised segmentation for warped labeled images only. Original and deformed manual masks were overlaid (2nd row) on the fixed, moving, and deformed moving images (1st row). ED, end-diastolic; ES, end-systolic.

framework yielded similar results as that of Seg_w. Combination of Seg_w, TAP, and L_{Reg_aux} yielded the highest DSC among these ablation methods and TransMorph.[38]

4.5 | Relationships between registration and segmentation accuracies

Table 9 shows the Pearson correlation coefficients between segmentation and registration DSC for various combinations of Reg, Seg_f, Seg_w, TAP, and L_{Reg_aux} trained on the same 20 ACDC and 30 M&Ms subjects as that in Tables 7 and 8. For 28/36 cases, there was no correlation ($p>0.05$) between segmentation and registration DSC. For the remainder, segmentation DSC was weakly ($n = 7, 0.200<r<0.399, p<0.05$) and moderately ($n = 1, 0.400<r<0.699, p<0.05$) correlated with registration DSC. Considering all the ablation methods and the proposed approach, there was strong correlation ($r>0.6$) between segmentation and registration DSC at a patient population level.

5 | DISCUSSION

Semi-supervised learning provides a way to alleviate the critical requirement of manual annotation for algorithm training, which requires numerous medical resources and hinders efficient clinical workflow. In this work, we developed an approach, based on the DeepAtlas framework, for semi-supervised cine cardiac MRI segmentation. Our technical contributions include: 1) incorporation of two independent segmentation models, one for fully supervised training using matched data provided by a registration module and the other for weakly supervised training using potentially mismatched pseudo label and original unlabeled images, 2) development of a temporal attention perceiver that explores inter-image relationships to generate optimized cross-instance features, enabling constraint of the alignment of fixed and warped moving images in higher dimensions for improved registration performance. For two public cine MRI datasets, we demonstrated: 1) effectiveness of the dual segmentation and the TAP module

TABLE 9 Pearson correlation coefficients r (p-value) between segmentation and registration DSC for RV, Myo, and LV for the ACDC and M&Ms test datasets for each ablation method and across all the ablation methods.

Components			ACDC (n = 50)			M&Ms (n = 50)		
Seg_w	L_{Reg_aux}	*TAP*	RV	Myo	LV	RV	Myo	LV
			0.349(0.013)	0.372(0.079)	0.156(0.278)	0.121(0.403)	0.304(0.032)	0.061(0.674)
✓			0.199(0.167)	0.377(0.007)	0.205(0.153)	0.164(0.254)	0.228(0.112)	0.053(0.714)
✓	✓		0.171(0.235)	0.306(0.031)	0.088(0.544)	0.167(0.028)	0.218(0.129)	0.013(0.931)
	✓		0.256(0.068)	0.349(0.038)	0.122(0.457)	0.188(0.725)	0.357(0.092)	0.117(0.245)
		✓	0.173(0.229)	0.272(0.056)	0.219(0.127)	0.144(0.317)	0.213(0.138)	0.215(0.134)
✓	✓	✓	0.311(0.028)	0.422(0.002)	0.062(0.667)	0.099(0.492)	0.168(0.241)	0.122(0.398)
Across all ablation methods			0.907(0.013)	0.723(0.105)	0.732(0.098)	0.920(0.009)	0.816(0.048)	0.628(0.181)

Abbreviations: ACDC, automatic cardiac diagnosis challenge; DSC, Dice similarity coefficient; LV, left ventricular cavity; Myo, myocardium; RV, right ventricular cavity; TAP, temporal attention perceiver.

as well as the entire pipeline for algorithm segmentation and registration; and 2) promising segmentation and registration accuracies that outperformed several state-of-the-art methods and approached fully supervised segmentation when using relatively small datasets and under annotations for training.

For both the ACDC and M&Ms datasets, we achieved higher segmentation accuracies using the proposed approach compared with UniMatch, BCP, MambaUNet, and DeepAtlas, which outperformed the other comparative semi-supervised segmentation methods. This may be because of the use of the dual segmentation networks for separate fully and weakly supervised segmentation and the TAP module for registration refinement in latent feature space in higher dimensions. In atlas-based methods, the same network was used to segment the labeled and the unlabeled images in a fully supervised manner, and anatomy similarity between the segmentation of the labeled and unlabeled images were enforced to improve the registration. However, registration error may lead to mismatch between the fixed unlabeled images and the resulting pseudo label, leading to issues in training the segmentation network. In contrast, we employed an additional segmentation network Seg_w that takes the fixed unlabeled images and the potentially mismatched pseudo label as input and trained Seg_w in a weakly supervised manner with cross entropy as the loss, as previously suggested,[39] potentially minimizing the issues in training Seg_f and Seg_w using a single network. Compared with DeepAtlas that uses a single segmentation network, the dual segmentation design yielded substantially greater DSC for both the ACDC and M&Ms datasets. As shown in Table 7, the incorporation of Seg_w improved DSC by 0.003–0.052 for the ACDC datesaet and 0.01–0.02 for the M&Ms dataset, compared with DeepAtlas. The BRBS[28] model provides a way to improve the registration of fixed and moving images leading to more accurate alignment between the fixed images and pseudo label. Although effective, this approach explored inter-subject registration and did not consider the spatial and temporal constraint between cine frames, resulting in sub-optimal registration that may impede the following segmentation. Here, we introduced a novel TAP model to generate optimized features for image pairs by leveraging the relationships between the feature instances extracted from two images. The resulting features related to the warped moving image and original fixed images were used to enforce spatial and temporal consistency and minimize the discrepancy between the original (I_m, I_f) and the warped (I_m, I'_m) image pairs at feature level, leading to improved registration and segmentation performance, as evidenced in Tables 8 and 7, respectively. In addition, we enforced the similarity between the segmentation of the fixed image and the pseudo label as an auxiliary loss to the registration module, further improving the registration and segmentation performance, as evidenced in Tables 8 and 7, respectively. This is likely because the weakly supervised segmentation network provides semantic information to the registration network, enhancing its perceiving ability of the heart region. As a result, the registration network focuses more on the foreground and generates more accurate registration and more realistic warped images and label for training the two segmentation networks.

As shown in Table 9, we observed weak toward no correlation between segmentation and registration DSC for each ablation method. This may be because the correlation was calculated at a subject level, and the registration module, which provides matched warped images and label that may be viewed as a way of data augmentation, is relatively independent from the fully supervised segmentation module. For example, there are cases where the registration of ED and ES frames is challenging yet the fully supervised segmentation remains relatively straightforward. When the segmentation and registration DSC were averaged for each test dataset of 50 subjects, we observed strong correlation (r = 0.628-0.920, Table 9) between the mean segmentation and registration DSC across the n = 6 variants of the proposed algorithm, suggesting an overall correlation between the registration and the fully supervised segmentation at a

population rather than individual subject level. We also investigated the correlation between registration and the fully supervised segmentation across the $n = 6$ ablation methods at an individual subject level. As expected, we observed weak-to-moderate correlation $r = 0.426\pm0.178, 0.403\pm0.168, 0.478\pm0.183, 0.461\pm0.316, 0.415\pm0.257$, and 0.484 ± 0.327 for RV, Myo, and LV in the ACDC and M&Ms test datasets, respectively (data not shown). These results suggest that, overall, improved registration facilitates the following segmentation at an individual subject level, and the trend was strengthened at a population level. As shown in Table 8, while the baseline framework comprised of the registration model and the fully supervised segmentation model yielded lower registration DSC for both ACDC and M&Ms test datasets compared with TransMorph, our proposed approach outperformed TransMorph by a large margin. These results suggest the utility of the weakly supervised segmentation and the TAP model in facilitating the deformable registration of ED and ES frames.

We also demonstrated the potential of our approach using relatively small datasets for algorithm training. As shown in Tables 1–3 and Tables 4–6, our approach yielded higher DSC and lower ASSD and HD than the other semi-supervised segmentation methods, and these are quite similar to fully supervised learning, which required twice the amount of manual label for training. As in Tables 2 and 5, we achieved greater DSC for Myo compared with Unet_UB when using 20 ACDC and 30 M&Ms subjects for training. Noticeably, our approach yielded uniformly greater DSC and lower ASSD and HD when using 10 ACDC subjects each with a random ED/ES frame labeled for training, compared with Unet_UB that used both ED and ES label for training. This may be because fully supervised learning methods, for example, Unet_UB, did not explicitly explore the relationships between ED and ES frames while our approach utilized the TAP module to improve the registration, which provided data to train the dual segmentation networks, implicitly leveraging the relationships between the two frames. In addition, our approach utilized deformable registration as a way to generate numerous intermediate warped images and labels that match unlabeled frames, leading to potentially more realistic and more effective data augmentation. Although the proposed approach outperformed Unet_UB when trained on 10 ACDC subjects (Table 3), different phenomenon was observed when using fewer subjects from the M&Ms dataset as in Table 6. This may be because of the use of a relatively greater number of training subjects from the M&Ms dataset, that is, 10 subjects from ACDC versus 15 cases from M&Ms. To verify these findings, we trained the proposed approach and Unet_UB by reducing the original 15 random M&Ms subjects to 10 cases. For the same 50 M&Ms test subjects, our approach yielded DSC of $\{0.823\pm0.111, 0.818\pm0.070, 0.889\pm0.072\}$, ASSD of $\{2.73\pm1.93, 1.72\pm0.71, 1.73\pm0.95\}$ mm, and HD of $\{11.21\pm7.01, 8.52\pm5.73, 6.12\pm4.16\}$ mm for {RV, Myo, LV}, superior to Unet_UB (DSC = $\{0.823\pm0.138, 0.782\pm0.097, 0.884\pm0.068\}$, ASSD = $\{3.15\pm2.52, 2.23\pm1.16, 2.12\pm1.43\}$ mm, and HD = $\{11.02\pm7.12, 9.38\pm5.68, 6.64\pm4.80\}$ mm for {RV, Myo, LV}). Other factors, including differences in cardiac pathologies, image acquisition center and protocols, inter-subject variations may contribute to various degrees of algorithm generalizability and registration and segmentation performance between the two datasets, which warrants further investigation. Overall, these results indicate that the proposed approach provides a way to reduce the burden of manual delineation of cine images for algorithm training, facilitating integration of deep learning segmentation models for efficient clinical workflow and adoption of the proposed approach in clinical scenarios where training images and annotations are limited.

Although promising results were achieved, we acknowledge several study limitations. In this work, we investigated semi-supervised segmentation of the widely used ED and ES frames in cine MRI. Extension and application of the proposed approach to all frames across an entire cardiac cycle is needed in the context of providing additional imaging measurements to improve clinical management of patients with cardiac disease. In addition, the proposed approach involves registering ED and ES frames that are structurally similar for the same subject to provide pseudo label for the unlabeled frames. This may limit application of the proposed approach to broader clinical situations, whereby manual annotation requires numerous resources and the vast majority of medical images used for training are unannotated. Although a relatively small portion of training subjects may be (partially) annotated, registering these subjects to other unlabeled cases that are structurally dissimilar can be challenging due to various factors, including anatomical and functional variability, differences in imaging protocols, image noise, contrast, resolution, pathology and disease status. These challenges highlight the need for improving the registration and segmentation performance of the proposed approach to accommodate broader clinical scenarios. Furthermore, the ACDC and M&Ms dataset comprise healthy volunteers and patients with limited spectrum of cardiac pathologies recruited at one and two clinical centers, respectively, and the segmentation algorithm was developed and evaluated on 2D slices with potential overfitting to specific pathologies. Investigation of larger 3D volumetric cardiac cine MRI datasets acquired from subjects with diverse cardiac pathologies using different MR systems and imaging protocols across multiple healthcare centers will be prioritized in the future. We observed sub-optimal degree of Pearson correlation and Bland–Altman agreement for algorithm versus manual RVEF measurements. This may be

because RV is generally more difficult to segment due to the relatively small size, complex shape and morphology, and blurry boundaries with surrounding tissues. Although the segmentation accuracies were generally promising, small differences in RV volumes at either ED or ES frames can lead to substantial variations in RVEF calculation. Further optimization of algorithm performance is required for translating the proposed approach to clinical deployment. Moreover, we employed perhaps the most widely used VoxelMorph and Unet as the registration and segmentation module, respectively, in the proposed algorithm framework. We note that the focus of this study is to develop and investigate the effects of the dual segmentation and the TAP modules, and to integrate these components to generate an effective joint segmentation and registration algorithm pipeline, rather than to develop single segmentation and registration networks. We think the proposed algorithm framework applies to many recently developed advanced segmentation and registration networks, for example, TransMorph, RSegNet, and U-ReSNet. These limitations represent the direction of our future work.

6 | CONCLUSION

We developed an approach that integrates deformable registration, fully and weakly supervised segmentation with feedback on registration, and a temporal attention mechanism for semi-supervised cine cardiac MRI segmentation. For two public cardiac MRI datasets, our approach outperformed several state-of-the-art methods by a large margin and closely approached fully supervised learning when using relatively small datasets and under annotations for training. These observations suggest that our approach could support the integration of deep learning models in clinical cardiac MR imaging workflow.

ACKNOWLEDGMENTS

We acknowledge the use of the computational facilities of Digital Research Alliance of Canada. This work was supported by National Key Research and Development Program of China (2018YFA0704000 and 2022YFC2410000), National Natural Science Foundation of China (82572361, 82127802, 21921004, and U21A20392), Key Research Program of Frontier Sciences CAS (ZDBS-LY-JSC004), Hubei Provincial Key Technology Foundation of China (2021ACA013 and 2023BAA021), and Natural Science Foundation of Hubei Province (2023AFB1061).

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts to disclose.

DATA AVAILABILITY STATEMENT

The data used in this study are publicly available at: https://www.ub.edu/mnms/ and https://www.creatis.insalyon.fr/Challenge/acdc/databases.html.

REFERENCES

1. Kramer CM, Barkhausen J, Bucciarelli-Ducci C, Flamm SD, Kim RJ, Nagel E. Standardized cardiovascular magnetic resonance imaging (CMR) protocols: 2020 update. *J Cardiovasc Magn Reson*. 2020;22:17.
2. Schulz-Menger J, Bluemke DA, Bremerich J, et al. Standardized image interpretation and post-processing in cardiovascular magnetic resonance-2020 update: society for cardiovascular magnetic resonance (SCMR): board of trustees task force on standardized post-processing. *J Cardiovasc Magn Reson*. 2020;22:19.
3. Ammar A, Bouattane O, Youssfi M. Automatic cardiac cine MRI segmentation and heart disease classification. *Comput Med Imaging Graph*. 2021;88:101864.
4. Hu H, Pan N, Wang J, Yin T, Ye R. Automatic segmentation of left ventricle from cardiac MRI via deep learning and region constrained dynamic programming. *Neurocomputing*. 2019;347:139-148.
5. Fu Z, Zhang J, Luo R, Sun Y, Deng D, Xia L. TF-Unet: An automatic cardiac MRI image segmentation method. *Math Biosci Eng*. 2022;19:5207-5222.
6. Wang S, Li C, Wang R, et al. Annotation-efficient deep learning for automatic medical image segmentation. *Nat Commun*. 2021;12:5915.
7. Lee D-H. Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. In: *Workshop on challenges in representation learning, ICML*. vol. 3, Atlanta; 2013:896.
8. Thompson BH, Caterina GD, Voisey JP. Pseudo-label refinement using superpixels for semi-supervised brain tumour segmentation. In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE; 2022:1-5.
9. Wang X, Yuan Y, Guo D, et al. SSA-Net: Spatial self-attention network for COVID-19 pneumonia infection segmentation with semi-supervised few-shot learning. *Med Image Anal*. 2022;79:102459.
10. Shi Y, Zhang J, Ling T, et al. Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation. *IEEE Trans Med Imaging*. 2021;41:608-620.
11. Kalluri T, Varma G, Chandraker M, Jawahar C. Universal semi-supervised semantic segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019:5259-5270.
12. Chen X, Yuan Y, Zeng G, Wang J. Semi-supervised semantic segmentation with cross pseudo supervision. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021:2613-2622.
13. Cui B, Zhang M, Xu M, Wang A, Yuan W, Ren H. Rectifying noisy labels with sequential prior: Multi-scale temporal feature affinity learning for robust video segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2023:90-100.
14. Sohn K, Berthelot D, Carlini N, et al. Fixmatch: simplifying semi-supervised learning with consistency and confidence. *Adv Neural Inf Process Syst*. 2020;33:596-608.
15. Yang L, Qi L, Feng L, Zhang W, Shi Y. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023:7236-7246.
16. Bai Y, Chen D, Li Q, Shen W, Wang Y. Bidirectional copy-paste for semi-supervised medical image segmentation. In: *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023:11514-11524.
17. Song B, Wang Q. SDCL: Students discrepancy-informed correction learning for semi-supervised medical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2024:567-577.
18. Tarvainen A, Valpola H. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. *Adv Neural Inf Process Syst*. 2017:1195-1204.
19. Yu L, Wang S, Li X, Fu C-W, Heng P-A. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer; 2019:605-613.
20. Xie Y, Zhang J, Liao Z, Verjans J, Shen C, Xia Y. Pairwise relation learning for semi-supervised gland segmentation. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*. Springer; 2020:417-427.
21. Zhang Y, Yang L, Chen J, Fredericksen M, Hughes DP, Chen DZ. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*. Springer; 2017:408-416.
22. Qiao S, Shen W, Zhang Z, Wang B, Yuille A. Deep co-training for semi-supervised image recognition. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018:135-152.
23. Luo X, Wang G, Liao W, et al. Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *Med Image Anal*. 2022;80:102517.
24. Ouali Y, Hudelot C, Tami M. Semi-supervised semantic segmentation with cross-consistency training. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020:12674-12684.
25. Luo X, Hu M, Song T, Wang G, Zhang S. Semi-supervised medical image segmentation via cross teaching between CNN and transformer. In: *International Conference on Medical Imaging with Deep Learning*. PMLR; 2022:820-833.
26. Ma C, Wang Z. Semi-Mamba-UNet: Pixel-level contrastive and cross-supervised visual Mamba-based UNet for semi-supervised medical image segmentation. *Knowledge-Based Syst*. 2024;300:112203.
27. Xu Z, Niethammer M. DeepAtlas: Joint semi-supervised learning of image registration and segmentation. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer; 2019:420-429.
28. He Y, Ge R, Qi X, et al. Learning better registration to learn better few-shot medical image segmentation: authenticity, diversity, and robustness. *IEEE Trans Neural Net Learn Syst*. 2022;35:2588-2601.
29. Wang S, Cao S, Wei D, et al. LT-Net: label transfer by learning reversible voxel-wise correspondence for one-shot medical image segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020:9162-9171.
30. Elmahdy MS, Wolterink JM, Sokooti H, Išgum I, Staring M. Adversarial optimization for joint registration and segmentation in prostate CT radiotherapy. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*. Springer; 2019:366-374.
31. Dinsdale N, Jenkinson M, Namburete A. Spatial warping network for 3D segmentation of the hippocampus in MR images. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*. Springer; 2019:284-291.
32. Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, Dalca AV. Voxelmorph: a learning framework for deformable medical image registration. *IEEE Trans Med Imaging*. 2019;38:1788-1800.
33. Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K. Spatial transformer networks. *Adv Neural Inf Process Syst*. 2015:2017-2028
34. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018:7132-7141.
35. 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 Trans Med Imaging*. 2018;37:2514-2525.
36. Campello VM, Gkontra P, Izquierdo C, et al. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge. *IEEE Trans Med Imaging*. 2021;40:3543-3554.
37. Grothues F, Smith GC, Moon JC, et al. Comparison of interstudy reproducibility of cardiovascular magnetic resonance with two-dimensional echocardiography in normal subjects and in patients with heart failure or left ventricular hypertrophy. *Am J Cardiol*. 2002;90:29-34.
38. Chen J, Frey EC, He Y, Segars WP, Li Y, Du Y. Transmorph: Transformer for unsupervised medical image registration. *Med Image Anal*. 2022;82:102615.
39. He Y, Li T, Ge R, et al. Few-shot learning for deformable medical image registration with perception-correspondence decoupling and reverse teaching. *IEEE J Biomed Health Inf*. 2021;26:1177-1187.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Qin Y, Guo F, Wang Z, Xiao S, Zhang L, Zhou X. Semi-supervised cine cardiac MRI segmentation via joint registration and temporal attention perceiver. *Med Phys*. 2025;52:e70094.
https://doi.org/10.1002/mp.70094