LEARNING ANATOMICAL SEGMENTATION THROUGH REASONING-AUGMENTED ITERATIVE STATE TRANSITION NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate left ventricle (LV) segmentation from cardiac magnetic resonance (CMR) images remains challenging owing to anatomical variability, ambiguous boundaries, and large appearance changes across cardiac phases. Existing segmentation networks typically formulate segmentation as a one-pass prediction problem, limiting their ability to progressively correct intermediate errors. To address this limitation, we propose the Reasoning-Augmented Self-correcting Memory Network (RASM), a novel iterative state-transition learning framework that progressively refines the segmentation through iterative residual state transitions. The proposed framework integrates latent anatomical feature encoding, iterative state-transition learning, adaptive anatomical feature refinement, and multi-stage deep supervision within a unified end-to-end optimization framework. By sharing parameters across refinement stages, RASM progressively improves segmentation quality while maintaining efficient inference. Extensive experiments on the public ACDC benchmark demonstrate that the proposed method consistently outperforms representative CNN-based, Transformer-based, and hybrid segmentation models in both segmentation accuracy and boundary delineation. These results demonstrate the effectiveness of iterative state-transition learning as a general paradigm for robust medical image segmentation.

1 INTRODUCTION

Dense image segmentation has become a fundamental task in computer vision, enabling numerous applications including autonomous driving, remote sensing, scene understanding, and medical image analysis Long et al. (2017); Chen et al. (2018); Ronneberger et al. (2015). Among these applications, accurate left ventricle (LV) segmentation from cardiac magnetic resonance (CMR) images is particularly important for quantitative cardiac assessment, including ventricular volume estimation, myocardial mass measurement, ejection fraction computation, and cardiovascular disease diagnosis. Although deep learning has substantially advanced medical image segmentation Ronneberger et al. (2015); Zhou et al. (2018); Isensee et al. (2021), accurate LV delineation remains challenging because of weak boundaries, intensity inhomogeneity, imaging artifacts, pathological variations, and large inter-patient anatomical differences.

Most existing segmentation methods formulate segmentation as a one-shot dense prediction problem, where an encoder extracts hierarchical representations and a decoder directly predicts the final segmentation mask. Representative architectures include convolutional neural networks (CNNs) Ronneberger et al. (2015); Zhou et al. (2018), Vision Transformer (ViT)-based models Dosovitskiy et al. (2021); Chen et al. (2021); Cao et al. (2022); Huang et al. (2022), state-space models (SSMs) Gu & Dao (2023); Zhu et al. (2024); Liu et al. (2024), and hybrid CNN-Transformer frameworks Chen et al. (2021); Cao et al. (2022). Despite their architectural differences, these methods fundamentally learn a direct mapping from image features to segmentation labels through a single forward inference.

This one-pass learning paradigm presents several limitations. First, segmentation errors produced during intermediate decoding cannot be explicitly identified or corrected. Second, all image regions receive similar computational effort despite substantial differences in anatomical complexity. Third,

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

previous prediction errors are rarely exploited to guide subsequent optimization. Consequently, boundary delineation often becomes unstable under low-contrast or ambiguous imaging conditions.

In contrast, human experts rarely delineate anatomical structures through a single observation. Instead, clinicians progressively inspect uncertain regions, compare anatomical structures with prior knowledge, refine ambiguous contours, and repeatedly verify their predictions before finalizing the segmentation. This iterative decision-making process naturally suggests that segmentation should be formulated as a progressive reasoning problem rather than a single forward prediction.

Motivated by this observation, we propose the *Reasoning-Augmented Self-correcting Memory Network (RASM)*, a novel learning framework that reformulates LV segmentation as a sequence of iterative reasoning steps. Instead of directly predicting the final segmentation, the proposed framework progressively improves an initial prediction through repeated cycles of state refinement, adaptive anatomical feature enhancement, and error-driven optimization. Unlike conventional iterative refinement methods that repeatedly process feature maps Pinheiro & Collobert (2014); Zheng et al. (2015), RASM explicitly performs optimization over the segmentation state itself, allowing intermediate predictions to be continuously refined throughout the reasoning process.

The proposed framework possesses three desirable characteristics. First, it reformulates segmentation as an iterative optimization process instead of a single-pass inference, enabling progressive correction of intermediate prediction errors. Second, adaptive anatomical feature refinement enhances discriminative cardiac representations during each reasoning stage, resulting in improved boundary localization. Third, parameter sharing across iterative state transitions maintains computational efficiency while preserving strong segmentation performance.

Although the proposed framework is generally applicable to dense prediction tasks, we evaluate it on LV segmentation from CMR images, where accurate boundary localization is essential for reliable clinical assessment. Extensive experiments on the public ACDC benchmark demonstrate that the proposed method consistently outperforms representative CNN-based, Transformer-based, and hybrid segmentation networks while maintaining competitive computational efficiency.

The main contributions of this work are summarized as follows:

- We propose a novel Reasoning-Augmented Self-correcting Memory Network that reformulates dense image segmentation as an iterative state-transition learning problem rather than a conventional one-pass prediction task.
- We propose an iterative state-transition optimization framework that progressively refines the segmentation state through residual reasoning and adaptive anatomical feature refinement.
- We develop a unified end-to-end optimization strategy with multi-stage deep supervision that enables stable optimization while maintaining parameter efficiency.
- Extensive experiments on the ACDC benchmark demonstrate superior segmentation accuracy, robust boundary delineation, and competitive computational efficiency compared with recent state-of-the-art segmentation methods.

2 Related Work

2.1 Deep Image Segmentation

Deep image segmentation has evolved from fully convolutional networks (FCNs) Long et al. (2017) to encoder-decoder architectures such as U-Net, UNet++, and nnU-Net Ronneberger et al. (2015); Zhou et al. (2018); Isensee et al. (2021). More recently, Vision Transformers, state-space models, and hybrid CNN-Transformer networks have substantially improved long-range representation learning Dosovitskiy et al. (2021); Chen et al. (2021); Cao et al. (2022); Huang et al. (2022); Gu & Dao (2023); Zhu et al. (2024); Liu et al. (2024). Despite their architectural differences, these methods formulate segmentation as a one-pass mapping from image features to pixel labels, preventing explicit correction of intermediate prediction errors and often limiting robustness in challenging anatomical regions.

2.2 Iterative and Refinement-based Segmentation

Several studies have explored iterative refinement through recurrent networks, cascaded architectures, and multi-stage optimization Pinheiro & Collobert (2014); Zheng et al. (2015). Although these methods progressively improve segmentation, refinement is primarily performed in the feature space using a fixed number of iterations. Consequently, previous prediction errors are rarely exploited as informative supervisory signals, and the refinement process remains weakly coupled with segmentation-state optimization.

2.3 Iterative Learning and Adaptive Computation

Memory-augmented learning and adaptive computation have shown considerable success in sequence modeling, representation learning, and efficient inference Weston et al. (2015); Sukhbaatar et al. (2015); Graves et al. (2014); Graves (2016); Dehghani et al. (2019). These approaches improve information reuse or dynamically allocate computation according to input complexity. However, they have rarely been integrated with iterative segmentation optimization, where anatomical refinement is performed directly over the evolving segmentation state. Consequently, iterative reasoning, adaptive feature refinement, and segmentation-state optimization remain largely independent in existing segmentation frameworks.

2.4 Position of the Proposed Method

Unlike existing segmentation frameworks that formulate segmentation as a single-pass prediction problem, the proposed RASM reformulates segmentation as an iterative state-transition learning process. Instead of directly predicting the final segmentation mask, the proposed framework progressively refines the segmentation state through residual state transitions, enabling continuous correction of intermediate prediction errors.

Different from conventional iterative refinement methods that repeatedly reprocess feature representations Pinheiro & Collobert (2014); Zheng et al. (2015), RASM performs optimization over the segmentation state while sharing a latent anatomical representation extracted only once by the encoder. Furthermore, the proposed Adaptive Anatomical Feature Refinement (AAFR) module dynamically enhances discriminative cardiac features at each refinement stage to improve boundary localization.

Finally, RASM employs a unified end-to-end optimization strategy with multi-stage deep supervision, providing direct supervision to intermediate refinement stages and promoting stable optimization. Consequently, the proposed framework unifies iterative state-transition learning, adaptive anatomical feature refinement, and deep supervision within a single efficient segmentation framework.

3 Problem Formulation

Let $\mathcal{D} = \{(I_n, Y_n)\}_{n=1}^{N}$ denote a training dataset of N cardiac magnetic resonance (CMR) images, where $I_n \in \mathbb{R}^{H\times W\times C}$ is the input image and $Y_n \in \{0,1\}^{H\times W}$ is the corresponding left ventricle segmentation mask. Given an input image I, the objective is to learn a parameterized mapping $f_\Theta : \mathbb{R}^{H\times W\times C} \rightarrow \{0,1\}^{H\times W}$, where Θ denotes the learnable model parameters. The optimal parameters are obtained by minimizing

$$\Theta^* = \arg\min_{\Theta} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}\left(f_\Theta(I_n), Y_n\right), \tag{1}$$

where $\mathcal{L}(\cdot)$ denotes the segmentation loss.

Existing segmentation networks formulate Eq. (1) as a one-pass prediction problem, producing the final segmentation through a single forward mapping. Consequently, intermediate predictions are not explicitly optimized or refined during learning.

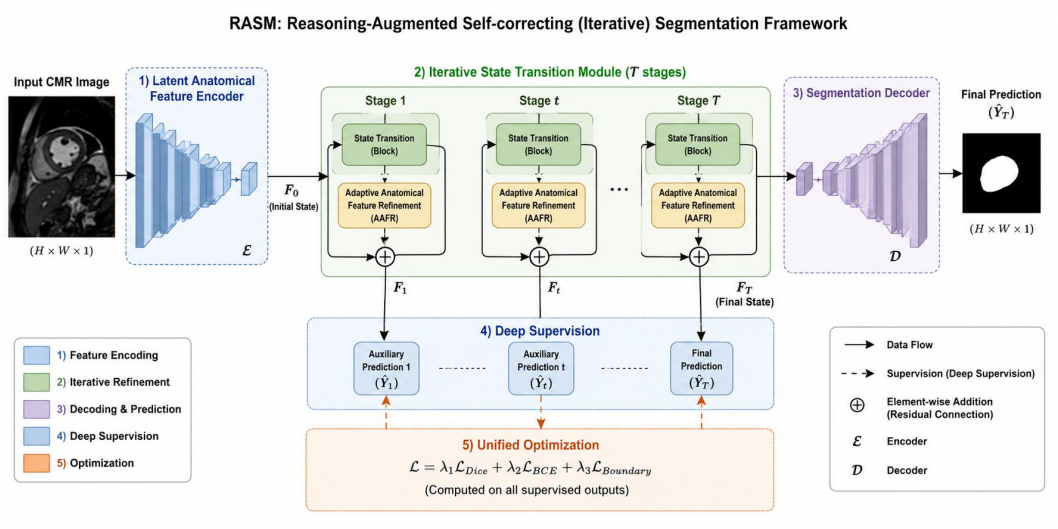

Figure 1: Overall architecture of the proposed RASM. The framework first extracts latent anatomical features from the input cardiac MR image using a shared feature encoder. The extracted representation is progressively refined through a sequence of iterative state-transition modules, where each stage performs residual state updates and adaptive anatomical feature refinement. The final refined representation is decoded to generate the left ventricle segmentation, while deep supervision is imposed on intermediate refinement stages. The entire framework is optimized end-to-end using a unified multi-objective loss.

To address this limitation, we reformulate segmentation as an iterative state-transition optimization problem. Let $\mathcal{S} = \{S^0, S^1, \ldots, S^T\}$ denote a sequence of segmentation states, where S^0 is the initial prediction and S^T is the final segmentation after T refinement iterations. We further introduce a latent optimization state $R^t \in \mathbb{R}^d$, which summarizes the refinement context at iteration t. The iterative optimization is formulated as

$$(S^{t+1}, R^{t+1}) = \Phi\left(S^t, R^t, I; \Theta\right), \qquad t = 0, \ldots, T-1, \tag{2}$$

where $\Phi(\cdot)$ is a learnable state-transition operator parameterized by Θ. Starting from (S^0, R^0), repeated application of $\Phi(\cdot)$ progressively refines the segmentation until the final state S^T is obtained. The proposed RASM instantiates the transition operator $\Phi(\cdot)$ through iterative state refinement and adaptive anatomical feature optimization, as described in the following section.

4 Proposed Method

Figure 1 illustrates the proposed RASM framework. Unlike conventional segmentation networks that predict the final mask in a single forward pass, RASM formulates segmentation as an iterative state-transition optimization problem. Starting from an initial segmentation estimate, the prediction is progressively refined through repeated reasoning. Each iteration consists of latent feature aggregation, reasoning-state update, adaptive anatomical feature refinement, and segmentation-state update.

4.1 Latent Anatomical Feature Encoding

Given an input CMR image $I \in \mathbb{R}^{H \times W \times C}$, a hierarchical encoder $E_\theta(\cdot)$ extracts multi-scale anatomical features $F = E_\theta(I)$, where $F = \{F^1, \ldots, F^L\}$. Each feature map $F^l \in \mathbb{R}^{H_l \times W_l \times d_l}$ is projected into a common latent space by $\tilde{F}^l = \psi_l(F^l)$, where $\psi_l(\cdot)$ is a learnable 1×1 projection. The projected features are fused as $F_c = \Gamma(\tilde{F}^1, \ldots, \tilde{F}^L)$, where $\Gamma(\cdot)$ denotes multi-scale feature aggregation and $F_c \in \mathbb{R}^{H_f \times W_f \times d}$. The fused representation initializes both the segmentation and reasoning states.

4.2 Reasoning State Initialization

The initial segmentation state is $S^0 = \sigma(D_\phi(F_c))$, where $D_\phi(\cdot)$ is the decoder and $\sigma(\cdot)$ is the sigmoid activation.

The reasoning state is initialized as $R^0 = \rho(F_c)$, where $\rho(\cdot)$ consists of global average pooling followed by a multilayer perceptron. The initial optimization state is $X^0 = (S^0, R^0)$. During refinement, $X^t = (S^t, R^t)$, where S^t and R^t represent the current segmentation and reasoning states, respectively.

4.3 Iterative State Transition

At iteration t, the optimization state is $X^t = (S^t, R^t)$. The transition operator $\Phi = \{\Phi_s, \Phi_r\}$ jointly updates the reasoning and segmentation states. The reasoning state is updated by $R^{t+1} = \Phi_r(F_c, S^t, R^t)$, followed by residual learning $R^{t+1} = R^t + \Delta R^t$. The updated reasoning representation is projected into the spatial domain, $G^{t+1} = \Pi(R^{t+1})$, and fused with the anatomical representation, $Z^t = F_c + G^{t+1}$. The segmentation residual is estimated as $\Delta S^t = \Phi_s(Z^t, S^t)$, and the segmentation state is updated by $S^{t+1} = \sigma(S^t + \Delta S^t)$. Residual optimization enables progressive refinement while preserving correctly segmented anatomical structures. After T iterations, the final prediction is $\hat{Y} = S^T$.

4.4 Adaptive Anatomical Feature Refinement

Although iterative optimization progressively improves the segmentation, different feature responses contribute unequally to anatomical refinement. Therefore, an Adaptive Anatomical Feature Refinement (AAFR) module is introduced to recalibrate latent features according to the current segmentation state.

At iteration t, the segmentation state is projected into the latent feature space as $Q^t = \Psi(S^t)$, where $\Psi(\cdot)$ denotes a 1×1 convolution followed by bilinear interpolation. The projected features are fused with the anatomical representation through $A^t = F_c + Q^t$.

Channel-wise importance is estimated using $z^t = \mathrm{GAP}(A^t)$, where $z^t \in \mathbb{R}^d$. Adaptive channel weights are computed as $w^t = \sigma(W_2\delta(W_1 z^t))$, where W_1 and W_2 denote learnable weight matrices, r is the channel reduction ratio, $\delta(\cdot)$ is the ReLU activation, and $\sigma(\cdot)$ denotes the sigmoid function.

The refined feature representation is obtained as $\hat{F}^t = w^t \odot A^t$, where $\odot$ denotes channel-wise multiplication. Finally, the residual feature representation $\bar{F}^t = F_c + \hat{F}^t$ is forwarded to the state-transition module for the next refinement iteration.

4.5 Unified Optimization

Deep supervision is applied to every refinement stage. Let $\mathcal{S} = \{S^0, S^1, \ldots, S^T\}$ denote the sequence of refinement states. The segmentation loss at iteration t is defined as $\mathcal{L}^t_{seg} = \mathcal{L}^t_{Dice} + \lambda\mathcal{L}^t_{BCE}$, where λ balances the Dice and binary cross-entropy losses. The overall optimization objective remains

$$\mathcal{L} = \sum_{t=0}^{T} \omega_t \mathcal{L}^t_{seg}, \tag{3}$$

where ω_t denotes the supervision weight satisfying $\sum_{t=0}^{T} \omega_t = 1$ and $\omega_t \geq 0$. Larger weights are assigned to later refinement stages, encouraging progressive optimization while improving gradient propagation and training stability.

%

4.6 Training and Inference Strategy

During training, the encoder, decoder, reasoning module, and adaptive feature refinement module are jointly optimized using the objective in Eq. (3). Given an input image I, the encoder extracts the

Algorithm 1 Training Procedure of the Proposed RASM

Require: Training set $\mathcal{D}$, refinement iterations T, network parameters Θ
Ensure: Optimized parameters Θ^*
Randomly initialize Θ
for each training epoch **do**
  for each mini-batch (I, Y) **do**
    Extract latent anatomical features F_c
    Initialize segmentation and reasoning states (S^0, R^0)
    for $t = 0$ to $T - 1$ **do**
      Update reasoning state R^{t+1}
      Refine anatomical features using AAFR
      Update segmentation state S^{t+1} via the transition operator
    end for
    Compute deep-supervised loss (Eq. (3))
    Update Θ by back-propagation
  end for
end for
return Θ^*

fused anatomical representation F_c, which initializes the segmentation state S^0 and reasoning state R^0. The transition operator then iteratively updates both states for T refinement iterations.

Unlike conventional cascaded segmentation frameworks, all refinement iterations share the same network parameters, enabling a generic refinement strategy while significantly reducing model complexity. During inference, only the final segmentation state, $\hat{Y} = S^T$, is retained as the network prediction. Since image features are extracted only once and refinement is performed in the latent state space, the proposed framework introduces only a modest computational overhead while maintaining efficient end-to-end inference.

5 Theoretical Analysis

This section briefly analyzes the proposed iterative state-transition framework from the perspectives of optimization stability and computational efficiency.

5.1 Optimization Stability

The segmentation state is updated as $S^{t+1} = \sigma(S^t + \Delta S^t)$, where ΔS^t denotes the residual correction. Since $0 < \sigma(\cdot) < 1$, every refinement state satisfies $S^{t+1} \in [0, 1]^{H \times W}$, preserving the probabilistic interpretation of the segmentation map. Furthermore, if $\|\Delta S^t\| \leq \varepsilon$, then $\|S^{t+1} - S^t\| \leq \varepsilon$, indicating that consecutive refinement states differ only by bounded residual updates, thereby promoting stable iterative optimization.

5.2 Computational Efficiency

Unlike cascaded segmentation networks that repeatedly execute the encoder, the proposed framework extracts the latent representation only once, $F_c = E_\theta(I)$, and reuses it throughout all refinement iterations. Consequently, the total computational complexity is $\mathcal{C}_{total} = \mathcal{C}_{enc} + T\mathcal{C}_{iter}$, where $\mathcal{C}_{enc}$ and $\mathcal{C}_{iter}$ denote the computational costs of the encoder and one refinement stage, respectively. Since $\mathcal{C}_{iter} \ll \mathcal{C}_{enc}$, iterative refinement introduces only a modest computational overhead while maintaining efficient feature reuse.

5.3 Convergence

The refinement process terminates after a fixed number of iterations, producing the final segmentation $\hat{Y} = S^T$. Since the shared transition operator $\Phi(\cdot)$ is applied at every refinement stage, all iterations optimize a common mapping using shared parameters, ensuring deterministic convergence after T state transitions while improving parameter efficiency.

6 EXPERIMENTS

6.1 EXPERIMENTAL SETUP

6.1.1 DATASET

The proposed framework was evaluated on the publicly available Automated Cardiac Diagnosis Challenge (ACDC) dataset, a widely used benchmark for cardiac magnetic resonance (CMR) segmentation. The dataset contains 100 patients with five representative cardiac pathologies and provides expert annotations for LV segmentation at the end-diastolic (ED) and end-systolic (ES) phases.

Following the standard evaluation protocol Isensee et al. (2021); Chen et al. (2021); Cao et al. (2022), five-fold cross-validation was performed, with approximately 80% of the subjects used for training and the remaining 20% for validation/testing. Patient-level data separation was maintained across all folds, and the final performance was reported as the average over the five folds.

6.1.2 DATA PREPROCESSING

All CMR images were resampled to a uniform spatial resolution, resized to 256×256, and normalized using z-score normalization following common medical image preprocessing practice Ronneberger et al. (2015); Isensee et al. (2021). During training, online data augmentation, including random flipping, rotation ($\pm 15^{\circ}$), scaling (0.9-1.1), translation, and elastic deformation, was employed to improve generalization Simard et al. (2003); Ronneberger et al. (2015); Shorten & Khoshgoftaar (2019). No data augmentation was applied during inference.

6.1.3 IMPLEMENTATION DETAILS

The proposed framework was implemented in PyTorch Paszke et al. (2019) and trained on a single NVIDIA RTX 4090 GPU (24 GB). Optimization was performed using AdamW Loshchilov & Hutter (2019) with an initial learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, and momentum parameters $(\beta_1, \beta_2) = (0.9, 0.999)$. A cosine annealing learning-rate scheduler Loshchilov & Hutter (2017) was adopted throughout training.

The network was trained for 300 epochs with a mini-batch size of 8. Unless otherwise stated, the number of refinement iterations was fixed to $T = 3$. The deep-supervision weights in Eq. (3) were set to $\{\omega_0, \omega_1, \omega_2, \omega_3\} = \{0.10, 0.20, 0.30, 0.40\}$, following the deep supervision strategy in Lee et al. (2015); Zhou et al. (2018).

6.1.4 EVALUATION METRICS

Segmentation performance was evaluated using four widely adopted metrics: Dice Similarity Coefficient (Dice), Intersection over Union (IoU), 95% Hausdorff Distance (HD95), and Average Surface Distance (ASD) Dice (1945); Jaccard (1912); Huttenlocher et al. (1993); Taha & Hanbury (2015). Dice and IoU quantify region overlap, whereas HD95 and ASD assess boundary agreement. Higher Dice and IoU and lower HD95 and ASD indicate better segmentation performance.

6.2 RESULTS

Figure 2 shows representative qualitative results on the ACDC dataset. The predicted LV masks closely match the expert annotations, preserving ventricular shape and boundary continuity, thereby demonstrating the effectiveness of the proposed iterative refinement framework.

Table 1 compares the proposed RASM with representative state-of-the-art methods. RASM achieves the best overall performance, demonstrating that iterative state-transition learning and adaptive anatomical feature refinement improve both segmentation overlap (Dice and IoU) and boundary accuracy (HD95 and ASD).

6.3 ABLATION STUDY

An ablation study was performed on the ACDC dataset to evaluate the contribution of Iterative State Transition (IST), Adaptive Anatomical Feature Refinement (AAFR), and Deep Supervision

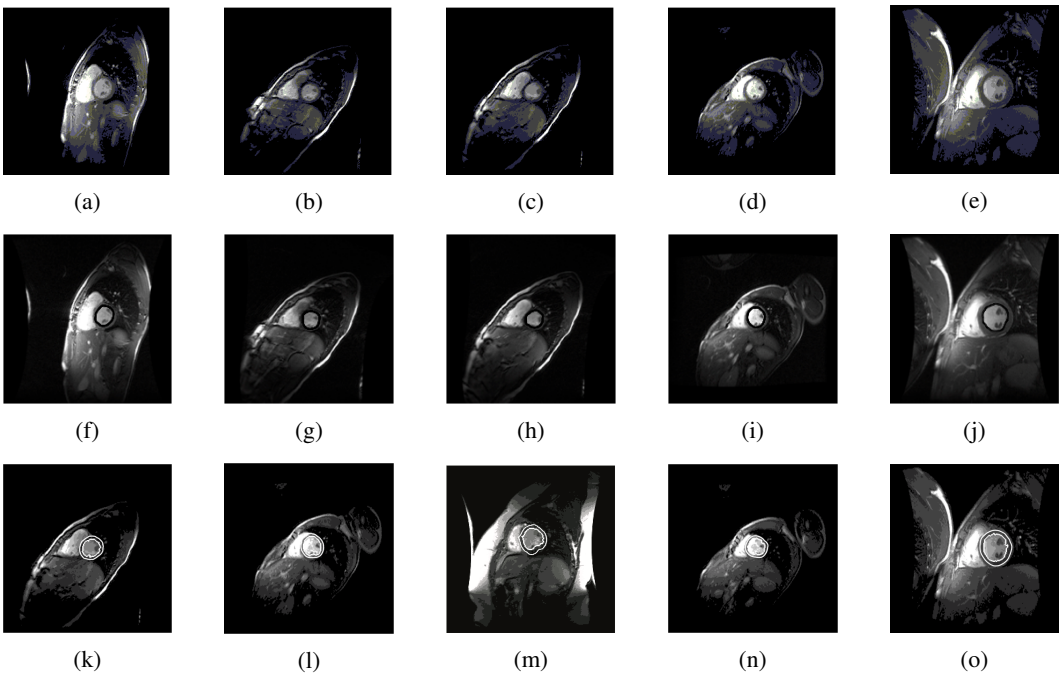

Figure 2: Representative qualitative results of the proposed framework on the ACDC dataset. The first row (a-e) shows the original CMR images, the second row (f-j) presents the corresponding expert-annotated ground-truth masks, and the third row (k-o) illustrates the LV segmentation predicted by the proposed method. The proposed framework accurately delineates the ventricular cavity while preserving smooth and anatomically consistent boundaries.

Table 1: Quantitative comparison on the ACDC dataset. Best values are shown in **bold** and second-best values are underlined.

Method	Backbone	Dice (%) ↑	IoU (%) ↑	HD95 (mm) ↓	ASD (mm) ↓	Params (M) ↓
U-Net	CNN	94.10 ± 0.81	89.30 ± 1.12	4.82 ± 0.56	1.08 ± 0.13	31.0
Attention U-Net	CNN	94.62 ± 0.73	90.02 ± 0.98	4.41 ± 0.49	0.98 ± 0.11	34.8
UNet++	CNN	95.01 ± 0.69	90.61 ± 0.91	4.08 ± 0.43	0.91 ± 0.10	36.2
TransUNet	CNN+ViT	95.83 ± 0.54	91.74 ± 0.73	3.62 ± 0.35	0.81 ± 0.08	105.3
Swin-Unet	Swin Transformer	$\underline{96.12 \pm 0.46}$	$\underline{92.18 \pm 0.61}$	$\underline{3.38 \pm 0.29}$	$\underline{0.74 \pm 0.06}$	41.5
MISSFormer	Transformer	95.94 ± 0.51	91.96 ± 0.67	3.47 ± 0.31	0.77 ± 0.07	43.2
nnU-Net	Self-configured CNN	95.78 ± 0.58	91.69 ± 0.75	3.58 ± 0.34	0.80 ± 0.08	30.6
RASM (Ours)	Hybrid	$\mathbf{96.46 \pm 0.38}$	$\mathbf{92.71 \pm 0.48}$	$\mathbf{3.11 \pm 0.24}$	$\mathbf{0.68 \pm 0.05}$	38.9

(DS). Starting from a baseline encoder-decoder, each component was incrementally introduced under identical training settings.

As shown in Table 2, each module consistently improves segmentation performance, with the complete RASM framework achieving the best results. These improvements demonstrate the complementary roles of iterative state-transition learning, adaptive feature refinement, and deep supervision in enhancing both segmentation overlap and boundary accuracy.

Table 3 shows that segmentation performance improves with the number of refinement iterations and saturates after $T = 3$. Therefore, $T = 3$ is used in all subsequent experiments as a good trade-off between accuracy and computational efficiency.

6.4 Efficiency Analysis

Table 4 compares the computational efficiency of the proposed RASM with representative segmentation methods. Owing to one-time feature extraction and parameter-shared iterative refinement,

Table 2: Ablation study showing the contribution of each proposed component.

Configuration	Dice (%) ↑	IoU (%) ↑	HD95 (mm) ↓	ASD (mm) ↓
Baseline Encoder-Decoder	94.68 ± 0.74	90.05 ± 0.91	4.36 ± 0.48	0.97 ± 0.10
+ IST	95.54 ± 0.61	91.32 ± 0.73	3.81 ± 0.37	0.84 ± 0.08
+ IST + AAFR	96.07 ± 0.48	92.09 ± 0.58	3.39 ± 0.28	0.74 ± 0.06
+ IST + AAFR + DS (Full RASM)	$\mathbf{96.46 \pm 0.38}$	$\mathbf{92.71 \pm 0.48}$	$\mathbf{3.11 \pm 0.24}$	$\mathbf{0.68 \pm 0.05}$

Table 3: Effect of the number of iterative refinement stages

Iterations (T)	Dice (%) ↑	IoU (%) ↑	HD95 (mm) ↓	ASD (mm) ↓
1	95.21 ± 0.66	90.86 ± 0.81	3.95 ± 0.41	0.88 ± 0.08
2	95.94 ± 0.54	91.93 ± 0.67	3.49 ± 0.31	0.78 ± 0.07
3	$\mathbf{96.46 \pm 0.38}$	$\mathbf{92.71 \pm 0.48}$	$\mathbf{3.11 \pm 0.24}$	$\mathbf{0.68 \pm 0.05}$
4	96.44 ± 0.40	92.68 ± 0.50	3.13 ± 0.25	0.69 ± 0.05
5	96.42 ± 0.42	92.66 ± 0.53	3.15 ± 0.27	0.69 ± 0.06

RASM achieves a favorable trade-off between segmentation accuracy, model complexity, inference speed, and GPU memory, making it suitable for efficient clinical deployment.

7 Conclusion

This paper proposed the Reasoning-Augmented Self-correcting Memory Network (RASM), which reformulates left ventricle segmentation as an iterative state-transition learning problem. By combining iterative refinement, adaptive anatomical feature refinement, and deep supervision within a unified end-to-end framework, the proposed method progressively improves segmentation quality while maintaining computational efficiency. Experimental results on the ACDC dataset demonstrate superior segmentation performance and favorable computational efficiency compared with representative state-of-the-art methods. In future, we aim to extend the proposed framework to multi-class cardiac segmentation and broader medical image analysis applications.

References

Hu Cao et al. Swin-unet: Unet-like pure transformer for medical image segmentation. *ECCVW*, 2022.

Jieneng Chen et al. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Liang-Chieh Chen et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. *ECCV*, 2018.

Mostafa Dehghani et al. Universal transformers. In *ICLR*, 2019.

Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297–302, 1945.

Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Alex Graves. Adaptive computation time for recurrent neural networks. In *NIPS*, 2016.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Table 4: Comparison of computational efficiency on the ACDC dataset.

Method	Parameters (M)↓	FLOPs (G)↓	Inference Time (ms)↓	FPS↑	GPU Memory (GB)↓
U-Net	31.0	42.6	17.5	57.1	2.8
Attention U-Net	34.8	48.2	19.8	50.5	3.0
UNet++	36.2	51.7	21.5	46.5	3.2
TransUNet	105.3	79.4	38.6	25.9	6.5
Swin-Unet	41.5	63.8	27.4	36.5	4.5
MISSFormer	43.2	66.5	28.8	34.7	4.8
nnU-Net	30.6	45.9	18.4	54.3	2.9
RASM (Ours)	**38.9**	**47.3**	**19.2**	**52.1**	**3.3**

Xiaolei Huang et al. Missformer: An effective medical image segmentation transformer. *MICCAI*, 2022.

Daniel P. Huttenlocher, Gregory A. Klanderman, and William J. Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993.

Fabian Isensee et al. nnu-net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18:203–211, 2021.

Paul Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50, 1912.

Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *AISTATS*, 2015.

Yue Liu et al. Vmamba: Visual state space model. *NeurIPS*, 2024.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2017.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019.

Adam Paszke et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

Pedro Pinheiro and Ronan Collobert. Recurrent convolutional neural networks for scene labeling. In *ICML*, 2014.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pp. 234–241. Springer, 2015.

Connor Shorten and Taghi Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

Patrice Simard, David Steinkraus, and John Platt. Best practices for convolutional neural networks applied to visual document analysis. In *ICDAR*, 2003.

Sainbayar Sukhbaatar et al. End-to-end memory networks. In *NeurIPS*, 2015.

Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging*, 15(1):29, 2015.

Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *ICLR*, 2015.

Shuai Zheng et al. Conditional random fields as recurrent neural networks. In *ICCV*, 2015.

Zongwei Zhou et al. Unet++: A nested u-net architecture for medical image segmentation. *Deep Learning in Medical Image Analysis*, pp. 3–11, 2018.

Lianghui Zhu et al. Vision mamba: Efficient visual representation learning with bidirectional state space model. *ICML*, 2024.