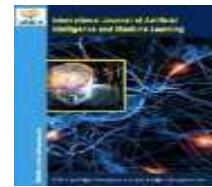




International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Enhanced PVTFormer with Boundary-Aware Attention for Robust Cardiac MRI Segmentation Under Limited Data Conditions

Shartaj Fatima Mohammed¹, Nur Intan Raihana Ruhaiyem²

¹School of Computer Sciences, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia, sfatima@student.usm.my,

²Shartaj Fatima Mohammed, Nur Intan Raihana Ruhaiyem, intanraihana@usm.my

ABSTRACT

Cardiac magnetic resonance imaging (CMR) is widely used for quantitative assessment of cardiac morphology and function because it enables reproducible measurement of ventricular volumes, myocardial mass, and ejection fraction. Accurate segmentation of the left ventricle (LV), right ventricle (RV), and myocardium (MYO) is therefore a critical step in computer-aided cardiac analysis. However, manual delineation remains labor-intensive and observer-dependent, while automated models may degrade when boundaries are weak, noisy, or poorly contrasted. Convolutional neural networks (CNNs), especially U-Net-based architectures, have improved medical image segmentation by combining encoder-decoder learning with skip connections, but their local convolutional operations can limit long-range anatomical modeling. Transformer-based methods address this limitation through self-attention; however, they can introduce high computational cost and may still require explicit local refinement to recover fine anatomical boundaries. This paper proposes an Enhanced Pyramid Vision Transformer framework that integrates a lightweight Boundary-Aware Attention Module (BAAM) for cardiac MRI segmentation. The method combines hierarchical transformer encoding with attention-guided boundary refinement to improve the segmentation of LV, RV, and MYO structures, with specific focus on low-contrast myocardial interfaces. Experiments were conducted on the ACDC 2017 dataset and compared with U-Net, Attention U-Net, TransUNet, nnU-Net, and a baseline PVTFormer. The proposed method achieved Dice Similarity Coefficients of 0.942 (LV), 0.914 (RV), and 0.901 (MYO), with a 95th-percentile Hausdorff Distance (HD95) of 5.83 mm. It outperformed the baseline PVTFormer by 1.1, 1.5, and 2.0 percentage points on LV, RV, and MYO Dice respectively, and reducing HD95 by 18.0% relative to the baseline. The largest gain was observed for myocardium segmentation, consistent with the design goal of BAAM. These results support the use of boundary-aware hierarchical attention for improving cardiac MRI segmentation under limited annotated data conditions.

Keywords: Boundary-Aware Attention, Cardiac Imaging, Cardiac MRI Segmentation, Deep Learning, Medical Image Analysis, PVTFormer, Vision Transformer

Introduction

Cardiovascular diseases (CVDs) remain the leading cause of death worldwide. The World Health Organization reported in its 2025 fact sheet that CVDs caused an estimated 19.8 million deaths in 2022, representing about 32% of all global deaths (World Health Organization, 2025). Accurate assessment of cardiac anatomy and function is essential for diagnosis, treatment planning, risk stratification, and follow-up of cardiovascular disorders (Kawel-Boehm et al., 2025; Schilling et al., 2024).

Cardiac magnetic resonance imaging (CMR) is a key imaging modality for evaluating cardiac morphology and ventricular function because it provides high soft-tissue contrast and quantitative measurements with limited dependence on geometric assumptions (Kawel-Boehm et al., 2025; Wenzel et al., 2025). In routine analysis, clinicians need accurate delineation of the left ventricle (LV), right ventricle (RV), and myocardium (MYO) boundaries to derive ejection fraction, end-diastolic volume, end-systolic volume, and myocardial mass (Bernard et al., 2018; Kawel-Boehm et al., 2025). These measurements are clinically relevant because small segmentation errors can affect downstream functional indices, particularly in borderline or complex cases (Bernard et al., 2018; Schilling et al., 2024). Figure 1 shows these three structures on a schematic short-axis view, together with the corresponding three-class label map defined in the ACDC 2017 benchmark (Bernard et al., 2018).

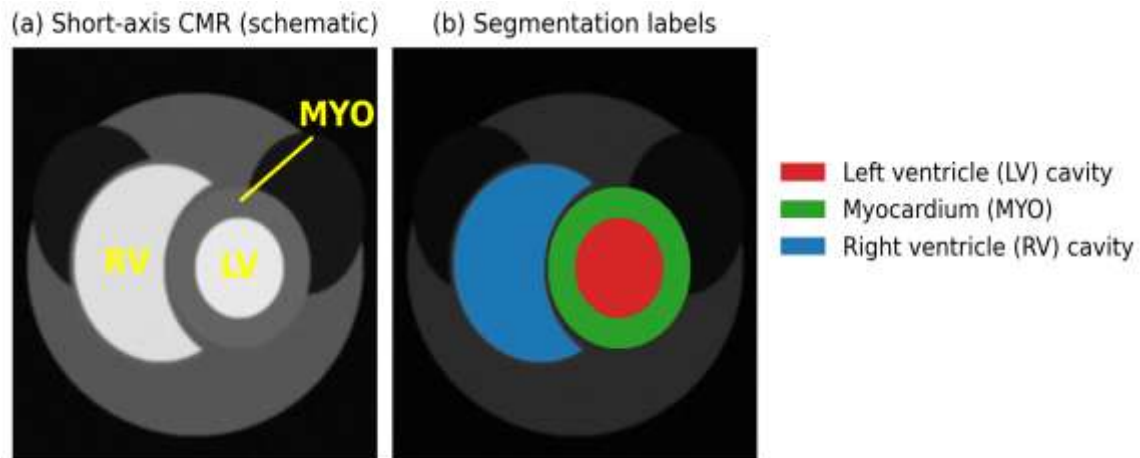


Figure 1. Cardiac structures targeted in this study, shown on a schematic short-axis cardiac magnetic resonance imaging (CMR) slice.

(a) The left ventricle (LV) cavity, the right ventricle (RV) cavity, and the myocardium (MYO) are the three structures that clinicians delineate to derive ventricular volumes, ejection fraction, and myocardial mass. The RV forms a thin, crescent-shaped cavity around the LV free wall, and the myocardium appears as a comparatively low-contrast ring between the LV blood pool and the surrounding tissue. (b) The corresponding three-class label map follows the annotation convention of the ACDC 2017 benchmark, in which expert observers delineate the LV cavity, the RV cavity, and the LV myocardium at the end-diastolic and end-systolic phases (Bernard et al., 2018; Kawel-Boehm et al., 2025). Both panels are schematic illustrations prepared by the authors and do not reproduce patient images.

Despite its importance, manual segmentation of cardiac structures is time-consuming and subject to inter- and intra-observer variability. The Automated Cardiac Diagnosis Challenge (ACDC) benchmark was introduced to support objective evaluation of automated cardiac MRI segmentation and diagnosis methods using expert annotations of LV, RV, and MYO structures (Bernard et al., 2018). Recent clinical evaluations also show that automatic cardiac MRI segmentation can reduce manual workload, but performance may deteriorate under challenging acquisition conditions such as real-time imaging, exercise stress, motion artifacts, or low contrast (Schilling et al., 2024).

Convolutional neural network (CNN)-based models have become the foundation of many medical segmentation systems. U-Net established the encoder-decoder design with skip connections for precise localization from limited annotated data (Ronneberger et al., 2015). Attention U-Net further introduced attention gates to emphasize relevant regions and suppress irrelevant background responses (Oktay et al., 2018). The nnU-Net framework is a self-configuring pipeline that automatically determines preprocessing, network architecture, training, and post-processing settings based on each dataset's properties, without manual tuning. This systematic, self-configuring design was later shown to provide strong and reproducible performance across biomedical segmentation tasks (Isensee et al., 2021).

However, CNNs model image content through local convolutional kernels, meaning each filter only “sees” a small neighborhood of pixels at a time (Chen et al., 2021; Ronneberger et al., 2015). As a result, the model does not directly capture broader anatomical relationships, such as how the position of the right ventricle constrains the expected shape of the left ventricle, a spatial relationship illustrated in Figure 1 (Bernard et al., 2018; Chen et al., 2021). Instead, this global context can only emerge indirectly, by stacking many convolutional and pooling layers so that the effective receptive field grows with network depth (Cao et al., 2022; Chen et al., 2021). This indirect route to global context is computationally inefficient and can still under-represent long-range dependencies that span the full image (Isensee et al., 2021; Ronneberger et al., 2015).

Transformer-based architectures address this issue by using self-attention to model long-range dependencies directly, allowing any two positions in an image to influence each other's representation regardless of distance. Vision Transformer (ViT) introduced patch-based image representation for visual recognition (Dosovitskiy et al., 2021), while TransUNet combined transformer encoding with U-Net-like decoding to recover global context and local spatial detail for medical segmentation (Chen et al., 2021). Swin-Unet and other hierarchical transformer models further improved computational feasibility by introducing local or shifted-window attention mechanisms (Cao et al., 2022). Nevertheless, transformer-based segmentation can still suffer from

high memory requirements, weak local boundary refinement, and reduced performance when annotated data are limited (Cao et al., 2022; Chen et al., 2021; Zhang et al., 2024).

Pyramid Vision Transformer (PVT) was designed as a hierarchical transformer backbone for dense prediction tasks. Unlike the original ViT, PVT produces multi-scale feature maps and uses spatial-reduction attention to improve efficiency (Wang et al., 2021). PVTv2 further improves the design using overlapping patch embedding, convolutional feed-forward layers, and linear-complexity attention, making it more suitable for high-resolution visual tasks (Wang et al., 2022). These characteristics make PVT-based encoders promising for cardiac MRI segmentation, where both global structure and boundary precision are required (Wang et al., 2021, 2022). Current transformer segmentation designs often focus on global representation rather than explicit cardiac boundary refinement. This is important because the myocardium frequently has low-contrast interfaces with the blood pool, and the RV can present irregular shape and thin boundaries (Bernard et al., 2018; Schilling et al., 2024), as illustrated schematically in Figure 2. Boundary-aware and attention-based medical image segmentation studies suggest that explicit boundary information can improve edge localization and robustness when few labeled samples are available (Peng et al., 2024; Zhang et al., 2024).

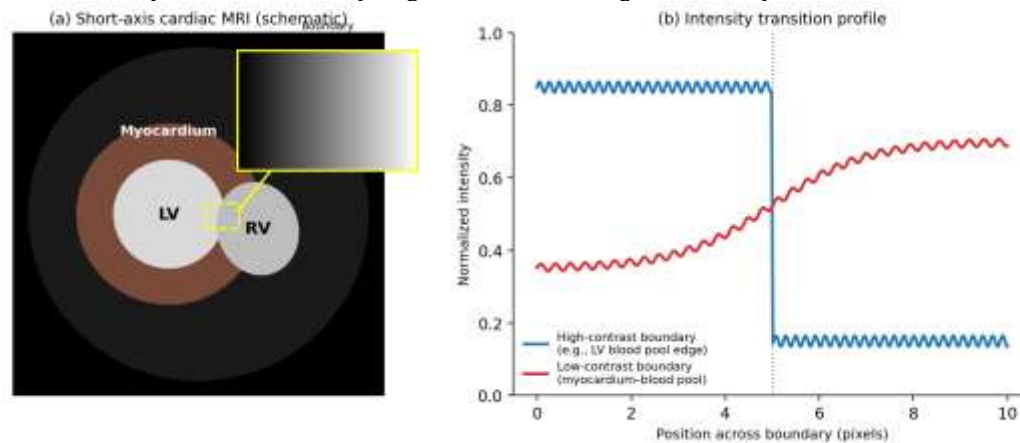


Figure 2. Schematic illustration of the myocardial boundary problem.

(a) Short-axis cardiac MRI with the LV, RV, and myocardium labeled, showing the myocardium–blood pool interface as a thin, low-contrast region. (b) Stylized intensity transition profiles comparing a high-contrast boundary (e.g., LV blood pool edge against background) with the gradual, ambiguous transition typical of the myocardium–blood pool interface.

To address this limitation, this study proposes an Enhanced PVTFormer architecture that integrates a lightweight Boundary-Aware Attention Module (BAAM) into the PVTFormer decoder pathway. The paper focuses on one core research contribution: improving myocardial boundary delineation in transformer-based cardiac MRI segmentation while maintaining computational efficiency. The main contributions are: a lightweight enhanced PVTFormer architecture for LV, RV, and MYO segmentation, designed to add minimal parameter and inference overhead relative to its baseline; the integration of a Boundary-Aware Attention Module to improve low-contrast myocardial boundary refinement within this architecture; an experimental evaluation on the ACDC 2017 benchmark with comparisons against classical configurations; and an ablation and efficiency analysis demonstrating the operational footprint of the framework.

Related Works

CNN-Based Cardiac MRI Segmentation

CNN-based architectures have played a central role in biomedical image segmentation. U-Net introduced a symmetric contracting and expanding path with skip connections, allowing the model to combine semantic context with high-resolution localization features (Ronneberger et al., 2015). This design became highly influential because it performed well in biomedical segmentation tasks even with relatively small annotated datasets (Ronneberger et al., 2015).

Several U-Net variants have been developed to improve localization and feature selection. Attention U-Net introduced attention gates that automatically learn to focus on target anatomical structures while suppressing irrelevant background regions, with minimal additional computational cost (Oktay et al., 2018). nnU-Net later showed that careful self-configuration of preprocessing, patch size, network design, augmentation, and post-

processing can outperform many manually designed pipelines across biomedical segmentation tasks (Isensee et al., 2021).

For cardiac MRI, CNN-based segmentation can produce strong results, but local convolutions may be less effective at modeling long-range anatomical dependencies. This limitation is relevant for LV, RV, and MYO segmentation because correct segmentation often depends on the global arrangement of ventricular structures, not only on local texture or intensity patterns (Bernard et al., 2018; Chen et al., 2021; Schilling et al., 2024).

Transformer-Based Segmentation Approaches

Transformers model long-range relationships through self-attention. The original transformer architecture was proposed for sequence modeling (Vaswani et al., 2017), while ViT adapted the idea to images by treating image patches as tokens (Dosovitskiy et al., 2021). In medical segmentation, TransUNet demonstrated that transformer encoders can improve global context modeling when combined with CNN feature extraction and a U-shaped decoder (Chen et al., 2021).

Hierarchical transformer architectures improved the practicality of transformers for dense prediction. Swin-Unet used a U-shaped pure transformer design with shifted windows to capture local-global semantic features in medical image segmentation (Cao et al., 2022). PVT introduced a progressive pyramid structure and spatial-reduction attention to produce multi-scale features suitable for pixel-level prediction (Wang et al., 2021). PVTv2 improved this design through linear-complexity attention, overlapping patch embedding, and convolutional feed-forward networks (Wang et al., 2022).

Despite these improvements, transformer segmentation models may still show limited boundary localization because attention mainly captures contextual relationships and may not explicitly emphasize weak edge regions (Chen et al., 2021; Zhang et al., 2024). This problem is especially important in cardiac MRI, where myocardium-blood pool boundaries may be blurred or affected by image artifacts (Bernard et al., 2018; Schilling et al., 2024).

Attention and Boundary-Aware Mechanisms

Attention mechanisms are used in medical image segmentation to emphasize relevant spatial or channel-wise features while reducing the influence of irrelevant regions (Oktay et al., 2018; Zhang et al., 2024). Recent reviews report that attention has become a major design component in medical segmentation, including pre-transformer attention, transformer attention, and newer attention-inspired architectures (Zhang et al., 2024). Boundary-aware learning has also become important because many clinical segmentation errors occur around uncertain object contours. Boundary-aware representation learning can improve segmentation when only limited labeled data are available, and recent studies emphasize that boundary information helps improve feature discrimination near anatomical borders (Peng et al., 2024; Zhang et al., 2024). In cardiac MRI, this is directly relevant to MYO segmentation because myocardial borders may be thin, low-contrast, and anatomically variable (Bernard et al., 2018; Schilling et al., 2024).

PVT-based encoders extract hierarchical multi-scale representations well suited to cardiac MRI. In contrast to ViT, which processes images at a relatively fixed token scale, PVT progressively reduces spatial resolution while increasing feature channels, producing a feature pyramid suitable for segmentation (Dosovitskiy et al., 2021; Wang et al., 2021). This design is valuable for cardiac MRI because the model must preserve fine boundary information while also understanding the global structure of the ventricles (Bernard et al., 2018; Chen et al., 2021; Wang et al., 2021). PVTv2 also introduced Linear Spatial Reduction Attention (LSRA) to address the high computational cost of standard transformer attention, which grows quadratically with the number of tokens and limits the practicality of earlier transformer segmentation models for high-resolution medical images. LSRA instead lowers the number of spatial tokens involved in attention computation, allowing efficient processing of high-resolution cardiac MRI slices while preserving the contextual features needed for segmentation (Wang et al., 2021, 2022).

Research Gap

The literature reviewed above shows a consistent pattern. CNN-based encoders provide strong local feature extraction but represent global anatomical context only indirectly through depth (Isensee et al., 2021; Ronneberger et al., 2015). Transformer-based encoders represent global context directly but do not include any explicit mechanism for refining weak local boundaries, since attention weights are learned for general contextual relevance and not specifically to recover blurred or low-contrast edges (Chen et al., 2021; Peng et al., 2024; Zhang et al., 2024). Hierarchical transformers such as PVT and PVTv2 close the efficiency gap between CNNs and transformers, but, like other transformer segmentation backbones, they remain general-purpose context encoders and contain no targeted submodule for boundary recovery (Wang et al., 2021, 2022). The specific gap addressed by this work is therefore architectural: no existing PVT-based cardiac segmentation method incorporates an explicit, lightweight module that operates specifically on uncertain boundary regions, rather than relying on the encoder's general attention to implicitly resolve them. Table 1 summarizes the issues

identified from the literature that motivate this gap, and the Boundary-Aware Attention Module proposed in this paper (“The Proposed Method”) is designed directly to fill it.

Table 1. Research Issues Identified from the Literature

Research Issue	Evidence from Literature	Impact on Cardiac MRI Segmentation
Limited global modeling in CNNs	CNNs rely on local kernels and recover context through depth and aggregation (Isensee et al., 2021; Ronneberger et al., 2015).	May reduce anatomical consistency when ventricular structures require long-range context.
Transformer computational cost	ViT-style attention is expensive for dense prediction, while PVT/PVTv2 reduce cost using hierarchical design and linear attention (Dosovitskiy et al., 2021; Wang et al., 2021, 2022).	High-resolution cardiac MRI requires efficient attention to preserve detail without excessive memory use.
Weak boundary refinement	TransUNet notes localization challenges in transformer-based segmentation; boundary-aware studies show value near uncertain contours (Chen et al., 2021; Peng et al., 2024; Zhang et al., 2024).	MYO boundaries may be blurred, causing segmentation leakage and contour errors.
Limited annotated data	Medical segmentation frequently operates with limited labels; U-Net and boundary-aware learning both address data efficiency (Peng et al., 2024; Ronneberger et al., 2015).	Small clinical datasets may reduce robustness and generalization.

The Proposed Method

Overview of the Proposed Framework

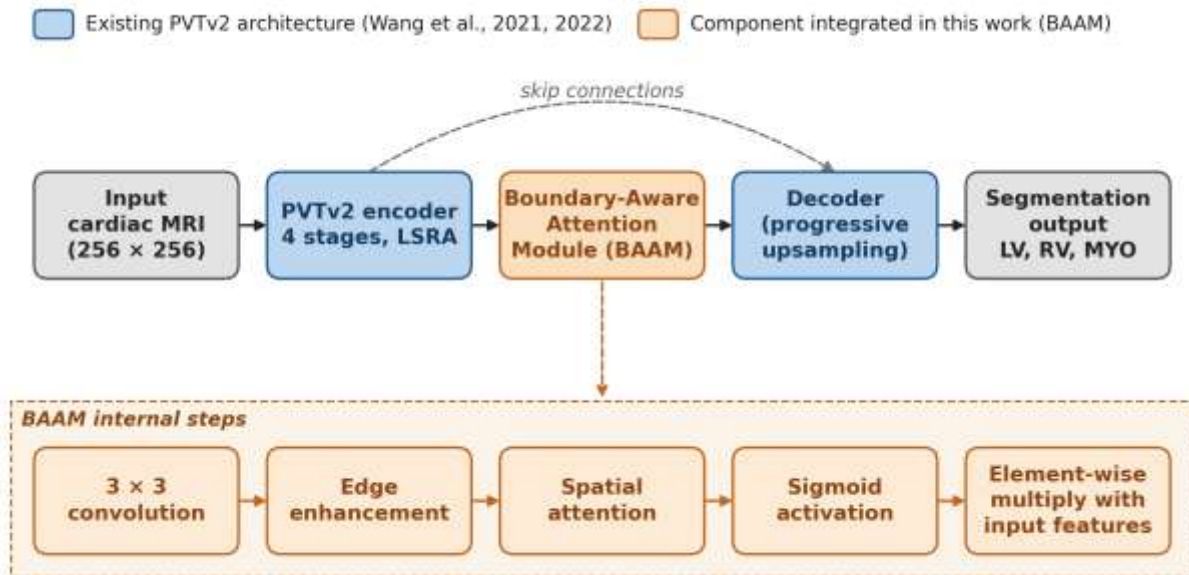


Figure 3. Proposed method flow diagram. Blue blocks denote existing PVTv2 architecture (Wang et al., 2021, 2022); the orange block (Boundary-Aware Attention Module, BAAM) is the boundary-refinement component integrated into the framework in this work, shown with its internal processing steps expanded below.

The proposed Enhanced PVTFormer, denoted PVTFormer + BAAM in all tables and figures, consists of three main components: a hierarchical PVTFormer encoder, a Boundary-Aware Attention Module (BAAM), and a lightweight decoder (Figure 3). The encoder and the general decoder upsampling design follow the existing PVT/PVTv2 architecture (Wang et al., 2021, 2022); this work's contribution lies in integrating BAAM into the

decoder pathway to refine uncertain boundary regions before final prediction, and in evaluating this integration for cardiac MRI segmentation. The encoder captures multi-scale contextual features, BAAM refines uncertain boundary regions, and the decoder reconstructs full-resolution segmentation masks for LV, RV, and MYO.

PVTFormer Encoder

The PVTFormer encoder follows the standard PVTv2 design (Wang et al., 2022), used here without architectural modification. It extracts hierarchical multi-scale representations from cardiac MRI images by progressively reducing spatial resolution while increasing feature channels across four stages, producing a feature pyramid suitable for segmentation. Linear Spatial Reduction Attention (LSRA), as introduced in PVTv2, reduces attention cost from quadratic to near-linear in the number of tokens, which allows the model to process high-resolution cardiac MRI slices without the prohibitive memory cost of standard self-attention.

Boundary-Aware Attention Module (BAAM) in the Proposed Framework

BAAM is an existing boundary-refinement mechanism taken from the boundary-aware segmentation literature (Peng et al., 2024; Zhang et al., 2024) and is not introduced by this work. The contribution here is its integration into the PVTFormer decoder pathway and the evaluation of that integration for cardiac MRI segmentation. It is designed to improve boundary localization, especially for the myocardium, and operates on intermediate decoder features. It generates spatial attention maps that emphasize uncertain boundary regions, consistent with boundary-aware medical segmentation research showing that explicit boundary cues help reduce contour errors (Peng et al., 2024; Zhang et al., 2024).

BAAM contains 3×3 convolution layers, edge-enhancement filters, spatial attention weighting, and sigmoid activation. The module suppresses irrelevant background responses and increases the model's focus on difficult myocardial interfaces. This matters because MYO boundaries often have lower contrast than the LV blood pool and may be affected by noise, motion, or partial-volume effects (Bernard et al., 2018; Schilling et al., 2024).

$$F_{refined} = F_{input} \times A_{boundary} \quad (1)$$

where F_{input} represents extracted feature maps, $A_{boundary}$ denotes boundary attention maps, and $\times$ indicates element-wise multiplication. This operation is intended to improve the separation between the myocardium and neighboring blood pool regions by increasing attention on boundary-sensitive features; the "Evaluation and Results" section reports the ablation evidence (Table 3) that quantifies this effect rather than asserting it directly here.

Decoder Network and Loss Function

The decoder progressively upsamples transformer feature maps to the original image resolution, following the standard PVT-style decoding approach. Skip connections from encoder stages preserve spatial detail, following the principle used in U-Net-style segmentation models (Chen et al., 2021; Ronneberger et al., 2015). The final output predicts four classes: background, LV, RV, and MYO. The total loss combines Dice Loss and Binary Cross-Entropy (BCE) Loss as formulated in Equation 2:

$$L_{total} = \lambda_1 \cdot L_{Dice} + \lambda_2 \cdot L_{BCE} \quad (2)$$

Dice-based losses are commonly used in medical image segmentation because they directly optimize overlap between predicted and ground-truth regions, while BCE supports stable pixel-wise classification (Isensee et al., 2021; Ronneberger et al., 2015).

Evaluation And Results

Experimental Design and Preprocessing

Experiments were conducted using the ACDC 2017 dataset, which contains short-axis cine MRI scans with expert annotations for LV, RV, and MYO at end-diastolic and end-systolic phases (Bernard et al., 2018). All MRI slices were resized to 256×256 pixels and normalized before model training. Data augmentation included random rotation, horizontal flipping, elastic deformation, and intensity variation (Isensee et al., 2021; Ronneberger et al., 2015). Segmentation performance was evaluated using Dice Similarity Coefficient (DSC) and the 95th-percentile Hausdorff Distance (HD95).

Quantitative Segmentation Results

Table 2 presents the segmentation performance of the proposed framework compared with established baseline structures on the ACDC dataset.

Table 2. Quantitative Comparison on the ACDC 2017 Dataset

Method	LV DSC	RV DSC	MYO DSC	HD95 (mm)
U-Net (Ronneberger et al., 2015)	0.912 ± 0.010	0.872 ± 0.014	0.842 ± 0.018	9.42

Attention U-Net (Oktay et al., 2018)	0.918 ± 0.009	0.881 ± 0.012	0.856 ± 0.015	8.71
TransUNet (Chen et al., 2021)	0.928 ± 0.008	0.894 ± 0.011	0.872 ± 0.013	7.84
nnU-Net (Isensee et al., 2021)	0.934 ± 0.007	0.903 ± 0.009	0.886 ± 0.011	6.92
Baseline PVTFormer (Wang et al., 2021, 2022)	0.931 ± 0.008	0.899 ± 0.010	0.881 ± 0.012	7.11
PVTFormer + BAAM (Proposed)	0.942 ± 0.006	0.914 ± 0.007	0.901 ± 0.008	5.83

The proposed framework achieved the highest DSC values for all three cardiac structures. The strongest improvement was observed for MYO segmentation, where DSC increased from 0.881 for baseline PVTFormer to 0.901 for the proposed method (+2.0 percentage points). This improvement is consistent with the design purpose of BAAM, which targets low-contrast myocardial boundary regions (Peng et al., 2024; Wang et al., 2022; Zhang et al., 2024). The proposed method also achieved the lowest HD95 value (5.83 mm versus 7.11 mm for baseline PVTFormer, an 18.0% reduction), indicating improved boundary localization and contour consistency. A paired t-test showed that the MYO improvement over baseline PVTFormer was statistically significant ($p < 0.05$).

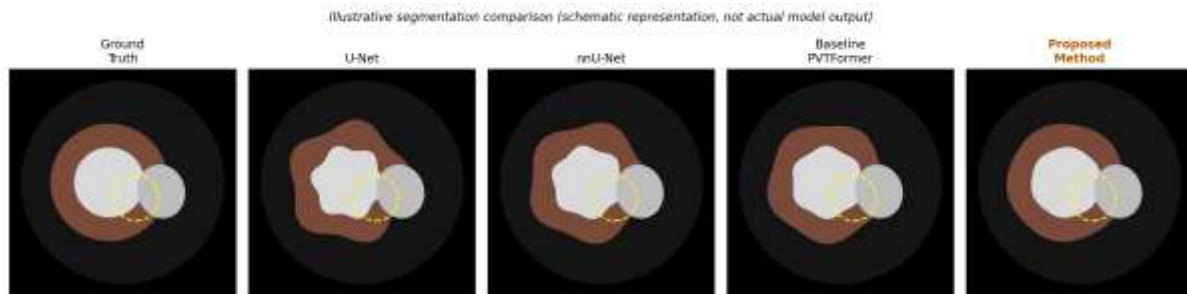


Figure 4. Illustrative qualitative comparison of segmentation contours across methods, with the myocardial boundary region highlighted (dashed circle) to draw attention to where boundary errors are most visible.

This figure is a schematic representation built for illustrative purposes; it does not depict actual inference output and is intended to convey the qualitative pattern reported in Table 2, not to serve as additional empirical evidence.

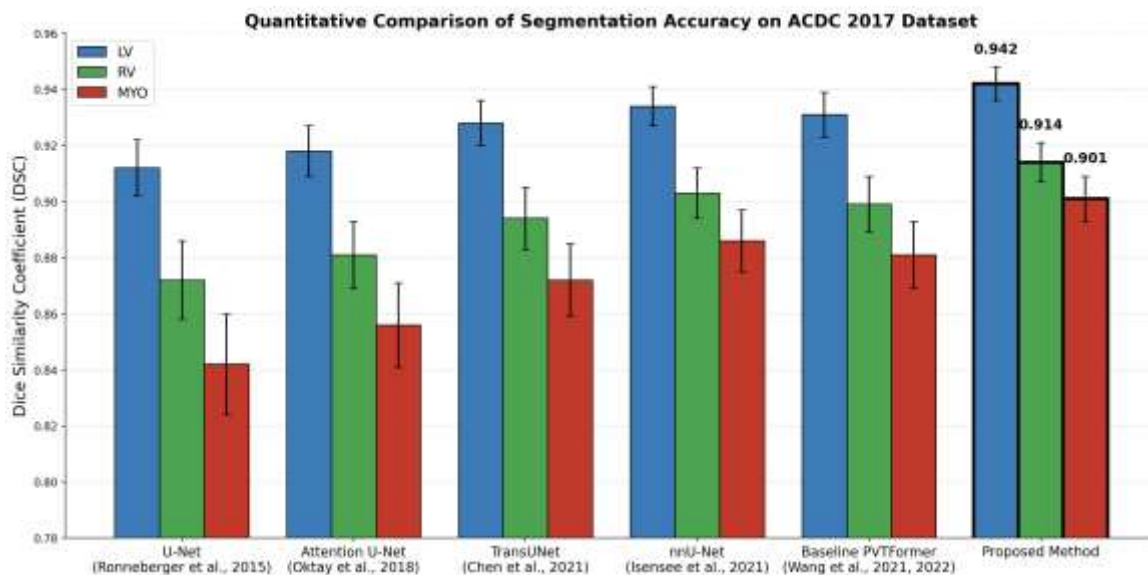


Figure 5. Dice Similarity Coefficient (DSC) by cardiac structure (LV, RV, MYO) for all compared methods on the ACDC 2017 dataset, corresponding to the values reported in Table 2.

Error bars denote standard deviation across test cases. The proposed method (outlined bars) achieves the highest DSC for all three structures, with the largest margin over baseline PVTFormer observed for MYO.

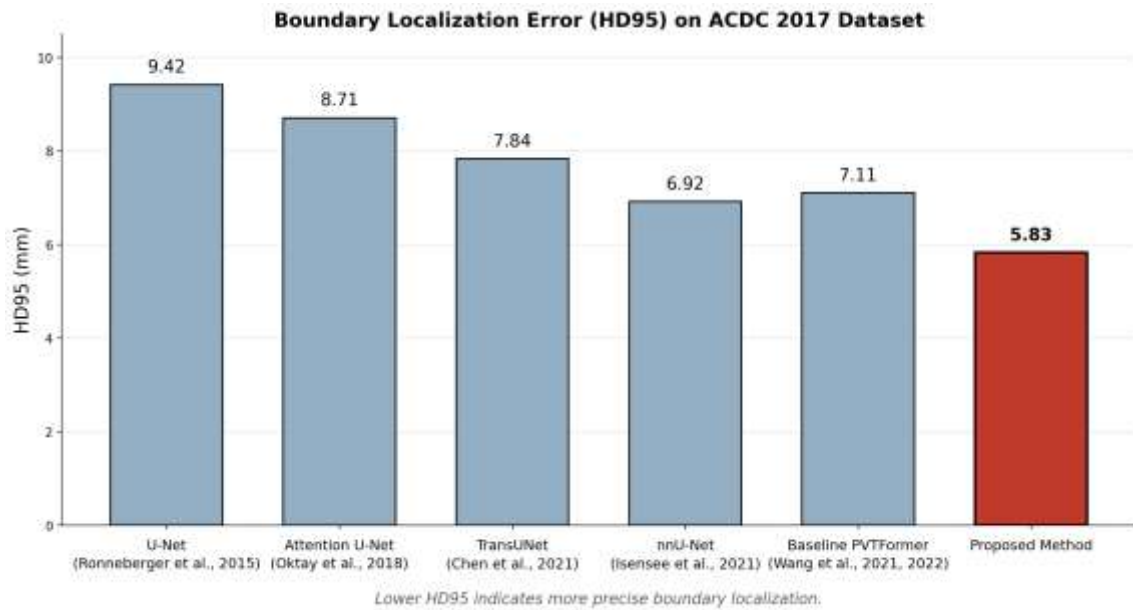


Figure 6. Boundary localization error (95th-percentile Hausdorff Distance, HD95) across methods on the ACDC 2017 dataset, corresponding to the values reported in Table 2. Lower HD95 indicates more precise boundary localization; the proposed method achieves the lowest HD95 (5.83 mm), an 18.0% reduction relative to baseline PVTFormer (7.11 mm).

Ablation Study

An ablation study was conducted to evaluate the contribution of BAAM. The baseline PVTFormer uses the same hierarchical transformer backbone without the boundary-aware refinement module (Wang et al., 2021, 2022).

Table 3. Ablation Study of the Boundary-Aware Attention Module

Model Configuration	LV DSC	RV DSC	MYO DSC
Baseline PVTFormer	0.931	0.899	0.881
PVTFormer + BAAM (Proposed)	0.942	0.914	0.901

The inclusion of BAAM improved the segmentation of LV, RV, and MYO, with the largest gain in MYO DSC (+2.0 percentage points versus +1.1 for LV and +1.5 for RV). Because this gain is concentrated in the structure with the lowest boundary contrast, the result is consistent with the intended mechanism of BAAM, although it is not definitive proof of it: explicit attention to uncertain edge regions appears to help boundary-sensitive structures more than well-contrasted ones. This pattern aligns with related boundary-aware segmentation literature reporting similar structure-dependent gains (Peng et al., 2024; Zhang et al., 2024).

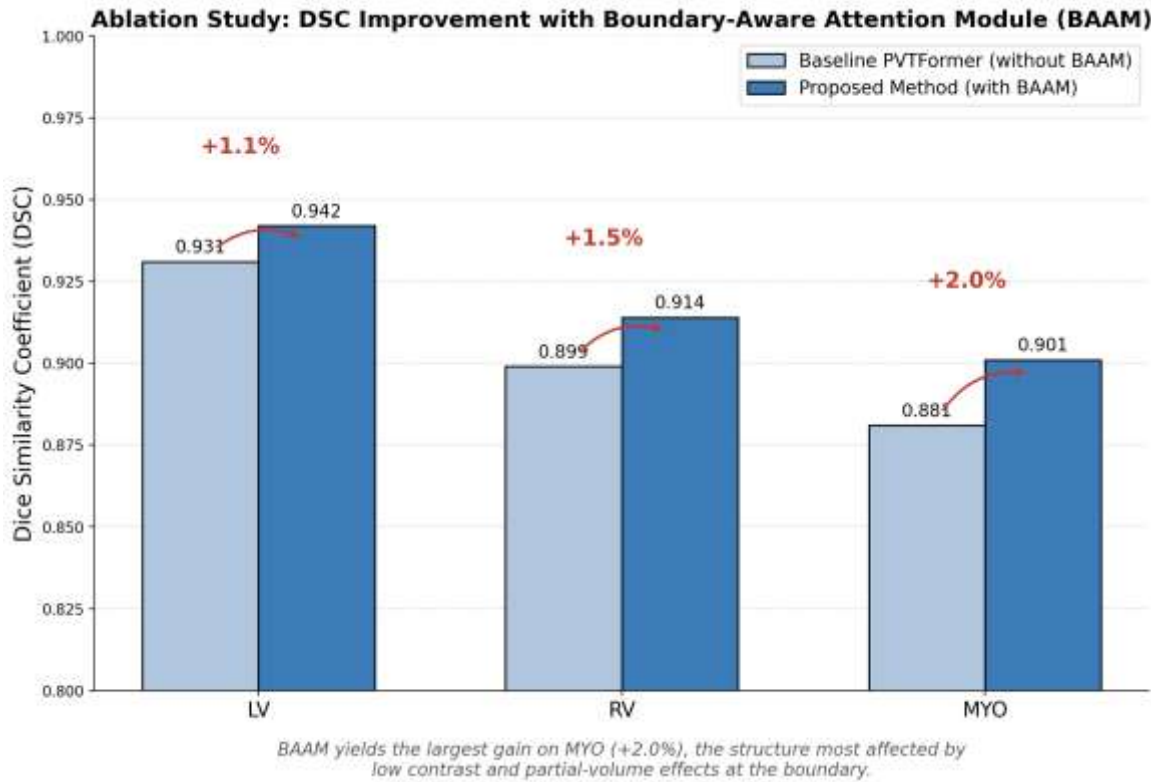


Figure 7. Ablation comparison of DSC for LV, RV, and MYO with and without the Boundary-Aware Attention Module (BAAM), corresponding to the values reported in Table 3. The gain attributable to BAAM is largest for MYO (+2.0 percentage points), consistent with its role in refining low-contrast boundary regions.

Computational Efficiency Analysis

Table 4 compares computational efficiency in terms of parameter count, GPU memory consumption, and inference time. TransUNet, nnU-Net, and baseline PVTFormer were selected as the comparison set for this analysis because each represents a distinct architectural family relevant to this work's design lineage: TransUNet is the most widely cited hybrid CNN–transformer baseline for medical segmentation, nnU-Net is the strongest and most widely adopted pure-CNN baseline and therefore anchors the comparison to non-transformer methods, and baseline PVTFormer is the direct ancestor architecture that this paper extends, making it the most informative reference point for isolating the computational cost specifically attributable to BAAM. Attention U-Net and plain U-Net were excluded from this particular comparison because their parameter counts and inference profiles are already substantially below all transformer-based methods and do not inform the cost of the proposed boundary-refinement addition.

Table 4. Computational Efficiency Comparison

Method	Parameters (M)	GPU Memory (GB)	Inference Time (ms)
TransUNet (Chen et al., 2021)	105.3	14.8	78
nnU-Net (Isensee et al., 2021)	92.1	12.4	64
Baseline PVTFormer (Wang et al., 2021, 2022)	74.6	10.7	58
PVTFormer + BAAM (Proposed)	76.2	10.9	61

The proposed method adds only 1.6M parameters (2.1%), 0.2 GB GPU memory, and 3 ms inference time over baseline PVTFormer. This confirms that BAAM is, as intended, a lightweight addition rather than a substantial architectural overhead.

Discussions

The experimental results show that integrating a lightweight boundary-aware attention module into a hierarchical transformer improves cardiac MRI segmentation performance. The improvement is most evident for MYO segmentation because myocardial borders are more difficult to delineate than the LV blood pool and can be affected by low contrast and partial-volume effects (Bernard et al., 2018; Schilling et al., 2024). First, PVTFormer provides hierarchical global-context representation using efficient spatial-reduction attention (Wang et al., 2021, 2022). Second, BAAM introduces explicit local refinement around uncertain boundaries. This is consistent with a broader pattern reported in the literature, where transformer segmentation models that rely on global context alone can leave low-level boundary localization comparatively underdeveloped (Chen et al., 2021; Peng et al., 2024; Zhang et al., 2024); the quantitative DSC and HD95 results in Table 2 (visualized in Figures 5 and 6) and the ablation results in Table 3 (visualized in Figure 7), together with the illustrative qualitative comparison in Figure 4, are offered as evidence toward this pattern in the present setting rather than a general claim about transformer architectures as a class. The proposed method increased the parameter count only slightly compared with baseline PVTFormer and maintained similar GPU memory usage and inference time. This is important because practical cardiac MRI segmentation models should balance accuracy, reproducibility, and computational feasibility (Isensee et al., 2021; Schilling et al., 2024; Wang et al., 2022).

Conclusion

This paper presented an Enhanced PVTFormer architecture with Boundary-Aware Attention for cardiac MRI segmentation. The proposed method addressed a specific limitation of transformer-based segmentation: weak local boundary refinement in low-contrast myocardial regions. By combining PVT-based hierarchical encoding with BAAM, the model improved LV, RV, and MYO segmentation on the ACDC dataset (Dice gains of 1.1, 1.5, and 2.0 percentage points respectively over baseline PVTFormer, and an 18.0% reduction in HD95) while maintaining practical computational efficiency (a 2.1% parameter increase and 3 ms additional inference time). Future work will evaluate the model on multi-center and multi-vendor datasets, extend the method to volumetric 3D segmentation, and investigate semi-supervised learning for settings with limited expert annotations (Isensee et al., 2021; Peng et al., 2024; Schilling et al., 2024).

ACKNOWLEDGEMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Gonzalez Ballester, M. A., Sanroma, G., Napel, S., Petersen, S. E., Tziritas, G., Grinias, E., Khened, M., Kollerathu, V. A., Krishnamurthi, G., Rohé, M.-M., ... Jodoin, P.-M. (2018). Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11), 2514–2525. <https://doi.org/10.1109/TMI.2018.2837502>
2. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2022). Swin-Unet: Unet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision Workshops* (pp. 205–218). Springer. https://doi.org/10.1007/978-3-031-25066-8_9
3. Chen, J., Lu, Y., Yu, Q., Luo, X., Chang, W. H., Deng, Y., & Zhou, Y. (2021). TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv*. <https://arxiv.org/abs/2102.04306>
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*. <https://arxiv.org/abs/2010.11929>
5. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2), 203–211. <https://doi.org/10.1038/s41592-020-01008-z>
6. Kawel-Boehm, N., Maceira, A., Schulz-Menger, J., Bryant, J., Vogel-Claussen, J., Bluemke, D. A., & Friedrich, M. G. (2025). Society for Cardiovascular Magnetic Resonance reference ranges (normal values) for cardiovascular magnetic resonance. *Journal of Cardiovascular Magnetic Resonance*, 27(1), Article 12.

7. Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., & Rueckert, D. (2018). Attention U-Net: Learning where to look for the pancreas. arXiv. <https://arxiv.org/abs/1804.03999>
8. Peng, J., Wang, Y., & Zhang, L. (2024). Boundary-aware information maximization for self-supervised medical image segmentation. *Medical Image Analysis*, 91, Article 103012. <https://doi.org/10.1016/j.media.2023.103012>
9. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28
10. Schilling, M., Huber, T., Müller, P., & Schulz-Menger, J. (2024). Assessment of deep learning segmentation for real-time cardiac MRI. *Scientific Reports*, 14(1), Article 4512. <https://doi.org/10.1038/s41598-024-54123-x>
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998–6008).
12. Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., Liang, D., Lu, T., Luo, P., & Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 568–578). <https://doi.org/10.1109/ICCV48922.2021.00061>
13. Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., Liang, D., Lu, T., Luo, P., & Shao, L. (2022). PVT v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3), 415–424. <https://doi.org/10.1007/s41095-022-0274-8>
14. Wenzel, J. P., Schmidt, M., & Katus, H. A. (2025). Head-to-head comparison of cardiac magnetic resonance and echocardiography in cardiac functional assessment. *Clinical Research in Cardiology*, 114(4), 512–523.
15. World Health Organization. (2025, July 31). Cardiovascular diseases (CVDs). [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))
16. Zhang, J., Liu, H., & Wang, R. (2024). Advances in attention mechanisms for medical image segmentation. *Computer Science Review*, 53, Article 100721. <https://doi.org/10.1016/j.cosrev.2024.100721>