HALSR-Net: Improving CNN Segmentation of Cardiac Left Ventricle MRI with Hybrid Attention and Latent Space Reconstruction

Mohamed Fakhfakh, Laurent Sarry, Patrick Clarysse

HALSR-Net: Improving CNN Segmentation of Cardiac Left Ventricle MRI with Hybrid Attention and Latent Space Reconstruction

Mohamed Fakhfakh[a], Laurent Sarry[a] and Patrick Clarysse[b]

[a]Université Clermont Auvergne, CHU Clermont-Ferrand, Clermont Auvergne INP, CNRS, Institut Pascal, F-63000, Clermont–Ferrand, France
[b]INSA-Lyon, Université Claude Bernard Lyon 1, CNRS, Inserm, CREATIS UMR 5220, U1294, F-69621, Lyon, France

ARTICLE INFO

Keywords:
Hybrid attention
Multi-scale Fusion
Medical Imaging Segmentation
Deep Learning

ABSTRACT

Accurate cardiac MRI segmentation is vital for detailed cardiac analysis, yet the manual process is labor-intensive and prone to variability. Despite advancements in MRI technology, there remains a significant need for automated methods that can reliably and efficiently segment cardiac structures. This paper introduces HALSR-Net, a novel multi-level segmentation architecture designed to improve the accuracy and reproducibility of cardiac segmentation from Cine-MRI acquisitions, focusing on the left ventricle (LV). The methodology consists of two main phases: first, the extraction of the region of interest (ROI) using a regression model that accurately predicts the location of a bounding box around the LV; second, the semantic segmentation step based on HALSR-Net architecture. This architecture incorporates a Hybrid Attention Pooling Module (HAPM) that merges attention and pooling mechanisms to enhance feature extraction and capture contextual information. Additionally, a reconstruction module leverages latent space features to further improve segmentation accuracy. Experiments conducted on an in-house clinical dataset and two public datasets (ACDC and LVQuan19) demonstrate that HALSR-Net outperforms state-of-the-art architectures, achieving up to 98% accuracy and F1-score for the segmentation of the LV cavity and myocardium. The proposed approach effectively addresses the limitations of existing methods, offering a more accurate and robust solution for cardiac MRI segmentation, thereby likely to improve cardiac function analysis and patient care.

1. Introduction

Magnetic Resonance Imaging (MRI) Lundervold and Lundervold (2019); Grover et al. (2015) is a diagnostic imaging technique able to provide detailed cross-sectional images of the heart, enabling dynamic visualization of cardiac structures. This technology is indispensable in current cardiac imaging for its precise imaging of the ventricles, atria, valves, and vessels. A key measurement in cardiac function evaluation is the ejection fraction (EF), which indicates the percentage of blood ejected from the left LV within each heartbeat D'Elia et al. (2015); Severino et al. (2020). Accurate estimation of the LV volume and EF is crucial for diagnosing various cardiac conditions, such as heart failure, cardiomyopathies, and coronary artery diseases. This estimation is also essential for planning and monitoring treatments to ensure optimal patient care.

Cine MRI, a specific MRI acquisition technique, captures sequences of images throughout the cardiac cycle, creating videos of the heart's movements Ismail et al. (2022). This technique is essential for assessing cardiac function, observing ventricular wall movements, and measuring volumes. Despite the significant advantages cine MRI, manually extracting heart contours from these images is labor-intensive and subject to considerable variability among experts Zhuang (2013) affecting the reliability and reproducibility of measurements. The manual process is time-consuming and can lead to inconsistent results. Therefore, there is a pressing need for more efficient and robust automatic

⋆This work was supported by the LABEX PRIMES (ANR-11-LABX-0063) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR). We are grateful to the Mésocentre Clermont Auvergne University for providing help and computing resources.

*Corresponding author,
M. Fakhfakh : mohamed.fakhfakh.research@gmail.com
**Principal corresponding author,
L. Sarry : laurent.sarry@uca.fr,
P. Clarysse : patrick.clarysse@creatis.insa-lyon.fr,
ORCID:0000-0002-5495-7655 (P. Clarysse)

segmentation methods to improve the accuracy and consistency of cardiac analyses.
Recent advancements in deep learning, particularly convolutional neural networks (CNNs) with encoder-decoder architectures, have revolutionized medical image segmentation Wang et al. (2022); Hesamian et al. (2019); Moeskops et al. (2016); Yu et al. (2021). These architectures enhance segmentation by integrating skip connections that combine low-level details with high-level semantic information, enabling the construction of detailed masks while preserving essential spatial details. However, many encoder-decoder models miss important fine details early in the process because they focus on extracting deeper features, which can affect the accuracy.
In addition to preserving features, deep learning models for segmentation must address structural limitations, especially in the use of convolutions and downsampling layers. These layers, while necessary for extracting contextual information, can compromise spatial precision and increase the model's computational demands Alzubaidi et al. (2021). The upsampling phase also presents challenges, as traditional methods like deconvolutions and unpooling frequently fail to capture the broader context needed for accurately restoring detailed anatomical features Gao et al. (2019). To address these issues, sophisticated network operators are needed to effectively manage both local and global information, integrating seamlessly with existing frameworks to enhance segmentation quality.

To address these issues, this paper proposes a multi-level segmentation architecture that improves both the accuracy and robustness of segmentation results as compared to existing methods through:

- A ROI Extraction Procedure: we introduce a method for extracting the region of interest (ROI), focusing on the left ventricle in MRI images. This step uses a regression architecture to predict diagonal points defining the ROI.
- To enhance segmentation accuracy and robustness, we introduce HALSR-Net, a new architecture designed for analyzing multi-phase and multi-slice cine images.
- HAPM Module (Hybrid Attention Pooling Module): between the encoder and decoder units, we integrate a hybrid pooling module with attention mechanisms, named HAPM. This module merges the advantages of attention and two kinds of pooling methods: Spatial Pyramid Pooling (APP) and Atrous Spatial Pyramid Pooling (ASPP). Executed in parallel, it significantly enhances feature extraction and thus improves the overall network performance by capturing richer and more relevant contextual information.
- In our architecture, a dedicated block for reconstructing an intermediate map from latent space. This map serves as supplementary information to the network's final output. This approach enriches the process by incorporating precise and contextual details, which enhances the reliability and accuracy of the overall result.

The remainder of this paper is organized as follows: section 2 reviews relevant literature on the deep learning approaches for segmentation in general, and cardiac structures segmentation in particular. Section 3 details the developed methodology and describes the architectural framework of the model. Section 4 presents the experimental datasets, along with the segmentation results. Finally, section 5 concludes the paper with a summary of the findings and perspectives for future research.

2. Related work: advances in Cardiac MRI Segmentation Techniques

Cardiac MRI analysis has witnessed significant improvements over the years, evolving from traditional image processing techniques to sophisticated deep learning models. This section reviews the progression of segmentation techniques, highlighting key developments and the impact of various methodologies on the efficiency of cardiac MRI evaluation.

2.1. Conventional Segmentation Techniques

These methods have been widely reviewed in the literature Peng et al. (2016); Chen et al. (2020), with various approaches categorized based on the extent of prior knowledge used during the segmentation processes. Initially, no-prior methods relied solely on the raw image content, using intensity thresholds, edges, and region information to delineate heart structures Wang et al. (2015); Ringenberg et al. (2014). Although straightforward, these techniques often struggle with ill-defined region boundaries in visible image cues. To address these limitations, deformable models such as active contours and level-set methods have been proposed. These models incorporate weak prior information about boundary smoothness. Similarly, graph theoretical models assume connectivity between neighboring pixels, providing

somewhat smooth segmentation results. While these methods are more adaptive than no-prior methods, they still face challenges related to high computational costs and the need for careful initialization Liu et al. (2016); Queirós et al. (2014).

Furthermore, active shape models, appearance models, and atlas-based methods heavily rely on predefined information about the geometry of heart structures, which can be overly constrained by the training set. Although these methods can address segmentation challenges in ill-defined boundary regions, they do so at a high computational cost. Nevertheless, these conventional techniques highlight the need to balance leveraging a-priori knowledge and managing computational complexity in cardiac image segmentation Peng et al. (2016).

2.2. Deep Learning Techniques

The evolution of segmentation techniques took a substantial turn with the advent of deep learning models which can learn from vast amounts of data without requiring explicit shape priors Qureshi et al. (2023); Du et al. (2020). These models represent a major improvement over semi-automatic methods Petitjean and Dacher (2011), that required significant user interventions, and are unsuitable for rapid processing scenarios as in clinical routine.

In medical image segmentation, the U-Net architecture Ronneberger et al. (2015) has had a considerable impact with its encoder-decoder structure and skip connections that retain main spatial details. Built from U-Net, several enhanced variants have been developed. U-Net++ Zhou et al. (2018) integrates nested, dense skip pathways to improve information flow. V-Net Milletari et al. (2016) adapts U-Net for three-dimensional data with residual connections to facilitate easier training. Other prominent architectures include ResUNet Alom et al. (2018) and ResUNet++ Jha et al. (2021), which introduce residual learning and Squeeze-and-Excitation (SE) blocks, and DenseUNet Li et al. (2018), which employs dense connectivity to promote robust feature propagation. SegNet Badrinarayanan et al. (2017) uses pooling indices for upsampling, maintaining precise boundary delineation, while Deeplab Chen et al. (2017) employs atrous convolutions to capture contextual information over larger spatial extents. Another significant development in the field is the nnU-Net Isensee et al. (2021). This architecture does not introduce new network designs but instead optimizes the U-Net for each specific dataset and task by automatically adapting preprocessing, network architecture, training, and post-processing steps. This flexibility reduces the need for manual tuning, ensuring that the model can be easily adapted to various medical image segmentation challenges without extensive manual adjustments.

2.3. Attention Mechanisms in Segmentation

Integrating attention mechanisms into image segmentation has significantly enhanced the ability to selectively prioritize crucial features while disregarding irrelevant background noise, which is essential for tasks such as body structures segmentation. Attention mechanisms dynamically adjust the network's focus, modifying input handling by weighting the importance of different features at various locations within the image.

The concept of attention in neural networks began with spatial transformer networks Jaderberg et al. (2015). This idea was further developed with the integration of attention blocks in U-Net architectures, notably through the Attention U-Net Oktay et al. (2018). This architecture introduces attention gates (AG) to focus on target areas, reducing computational load and enhancing accuracy, especially for delineating complex anatomical structures. Based on these foundations, channel attention mechanisms were introduced in the Squeeze-and-Excitation Network (SE-Net) architecture by Hu et al. Hu et al. (2018). These mechanisms recalibrate network channels to emphasize useful features while suppressing irrelevant ones.

Further improvements have led to the development of mixed attention models. For instance, FocusNet Kaul et al. (2019) integrates both spatial and channel attentions, enhancing feature representation by concurrently addressing spatial and channel-wise relevance. This approach enables more detailed and context-aware analysis. Additionally, the Non-local U-Net Wang et al. (2020) extends attention mechanisms by capturing long-range dependencies within an image. This model improves segmentation accuracy by integrating information across the entire image, thereby considering both local and distant relationships.

2.4. Multi-Scale Feature Handling

Effectively managing varying object scales, especially at different stages of pathology development, is crucial for accurate diagnostic imaging. Standard convolutional operations with fixed-size kernels often fail to capture contextual variations of differently sized objects.

Pyramid pooling has been a key point in enhancing multi-scale feature handling. Introduced by He et al. He et al. (2015) with Spatial Pyramid Pooling (SPP), this approach segments the image into regions of varying resolutions, extracts

features at different scales, and combines them into a comprehensive feature map, enabling neural networks to capture essential details at multiple scales. Residual Multi-Kernel Pooling (RMP) Gu et al. (2019) refines this concept by using multiple pooling kernels, though its upsampling strategy can introduce approximations that may dilute detailed information at higher pooling levels. To reduce detail loss, Atrous Convolution, or dilated convolution, expands the receptive field without increasing parameters or losing resolution. This technique, combined with SPP principles, forms the ASPP module Chen et al. (2017), that enhances segmentation accuracy by recognizing objects across varying scales. ASPP is particularly effective in various medical imaging segmentation tasks Moreno Lopez and Ventura (2018); Lei et al. (2021).

Combining ASPP with Non-local operations further enhances feature extraction by capturing detailed and contextual information across the entire image Yang et al. (2019). Non-local operations evaluate the entire image, which is beneficial for detecting subtle and dispersed anomalies. This integration enriches the feature set with extensive contextual information, preserving fine details and improving the overall accuracy and robustness of medical image analysis.

2.5. Specialized Architectures and Applications

Recent advances in medical image analysis have led to the development of novel architectures incorporating attention mechanisms and multi-scale feature handling. These innovative approaches enhance feature recognition, particularly in complex medical scans, improving diagnostic accuracy.

The ARW-Net architecture Singh et al. (2023) represents a significant improvement, enhancing the traditional U-Net with attention-guided residual connections and deep supervision. This design precisely focuses on relevant features within Magnetic Resonance images, addressing challenges in model generalizability by combining residual blocks in encoders and multiple attention modules in decoders. Similarly, the MA-UNet architecture Cai and Wang (2022) improves upon standard CNNs by integrating attention mechanisms and a multi-scale prediction fusion approach. This model uses attention gates to reduce semantic ambiguity in skip connections between the encoder and decoder networks. However, a notable issue with this approach is that multi-scale fusion at each decoder level can introduce noise, as information loss occurs during the upsampling process needed to reconstruct the feature maps. This simultaneous processing of spatial and channel dependencies, while aggregating features from multiple intermediate layers, aims to utilize global information at various scales. Nevertheless, this can lead to less precise segmentations due to the introduced noise.

Further advancing the field, CFNet Zhan et al. (2023) introduces a U-shaped encoder-decoder architecture enhanced with a multi-view attention mechanism and adaptive fusion strategy. CFNet includes a cross-scale feature fusion method (CFF) and a fusion weight adaptive allocation strategy (FAS), which together address the semantic gap between shallow and deep features. The model effectively extracts features across various scales and employs FAS to balance the fusion of these features adaptively during decoding.

From these advancements, DSGA-Net Sun et al. (2023) features a Deeply Separable Gated Visual Transformer (DSG-ViT) and a Mixed Three-Branch Attention (MTA) module. This architecture enhances the extraction of contextual links among global, local, and channel information, increasing sensitivity to location details and rectifying common issues such as under-delineation and over-delineation of small organs. Additionally, specialized architectures like CFHA-Net Yang et al. (2023) combine a cross-scale fusion strategy and a hybrid attention mechanism, incorporating a Cross-scale Context Fusion (CCF) module and a Triple Hybrid Attention (THA) module to optimize skip connections. This approach enhances long-range dependency and boundary detection in medical image segmentation. For various segmentation tasks, the Weighted ResUNet Xiao et al. (2018) leverages a weighted attention mechanism and skip connection scheme within a U-Net-like architecture, enhancing the handling of small and thin structures and improving discrimination in challenging areas, addressing issues from low contrast and noisy backgrounds. Additionally, Dangi et al. Dangi et al. (2019) proposed a novel regularization method that enhances FCN architectures with a distance map prediction task, improving feature learning without increasing network size.

However, these state-of-the-art methods often face issues related to complexity and computational overhead. Combining multiple attention mechanisms can significantly increase the computational burden, making the models slower and less efficient. Integrating various pooling methods can lead to architectures that are difficult to optimize and may require extensive hyperparameter tuning. The complexity of these models can sometimes result in overfitting, especially when the training data is limited or not sufficiently diverse. Additionally, the increased memory requirements from these integrations can limit the model's scalability and deployment on resource-constrained devices.

2.6. Limitations of Existing Approaches

While both conventional and deep learning-based segmentation techniques have significantly advanced cardiac MRI analysis, they are not without limitations. Conventional methods, such as deformable models and atlas-based approaches, often suffer from high computational costs and the need for careful initialization, making them unsuitable for clinical applications where speed and efficiency are critical. Moreover, these methods rely heavily on prior knowledge, which can lead to over-constrained solutions that fail to generalize well to diverse datasets. Additionally, these approaches may lose critical information in regions with low contrast or complex anatomical variations, further limiting their robustness.

Deep learning models, despite their advantages in automation and accuracy, also face several challenges. Architectures like U-Net and its variants (e.g., U-Net++, ResUNet, and nnU-Net) are limited by their reliance on encoder-decoder structures, which can lead to the loss of fine-grained spatial information during the downsampling process. This compromises the precision of segmentation, particularly for small or complex structures. While attention mechanisms have been introduced to address some of these issues, their combination with multi-scale feature fusion techniques results in a large number of parameters, making models computationally intensive and prone to optimization challenges. This increased complexity can hinder the effectiveness of gradient-based optimization algorithms and amplify the risk of overfitting, especially when training data is limited or not diverse enough.

These limitations underscore the need to find a balanced trade-off between these complementary modules, which have demonstrated high performance individually, but require careful integration to maximize their potential while maintaining computational efficiency and optimization feasibility.

3. Proposed methodology

The primary objective of this research is to enhance the accuracy and robustness of the endocardial and epicardial contours segmentation of the left ventricle from cardiac MRI. We introduce a novel multi-level segmentation architecture that integrates a regression model for ROI extraction and the HALSR-Net architecture for semantic segmentation. This approach aims to achieve more accurate and robust segmentation in multi-phase and multi-slice cine MRI images.

The proposed methodology consists of two main steps, as illustrated in Fig. 1:

- Extraction of the ROI using a Regression Model: the first phase involves the extraction of the ROI for LV from the input MRI image. This step is crucial as it isolates the LV, which is the primary focus for further segmentation. Instead of using a traditional approach for this task, we utilize a regression model. This model predicts the coordinates of the diagonal points of the bounding box surrounding the target region. In the current state of the art, segmentation methods are typically employed for this extraction, as seen in Von Zuben et al. (2023); Abdeltawab et al. (2020). However, regression models offer several advantages. They are generally more computationally efficient, reducing the processing time. By directly predicting the bounding box coordinates, these models can provide more precise localization, especially in cases where anatomical structures have high variability. Additionally, the regression approach simplifies the workflow, making it less prone to errors introduced by more complex procedures.

- Semantic Segmentation using HALSR-Net architecture: once the target region is extracted, the next phase involves the segmentation of the endocardium and epicardium within this isolated area using the proposed HALSR-Net architecture. HALSR-Net is specifically designed to handle the complexities of medical image segmentation. The architecture employs an encoder-decoder structure that effectively captures and processes features at multiple scales. The Hybrid Attention Pooling Module (HAPM) is integrated between the encoder and decoder units, and combines the strengths of both attention mechanisms and pooling methods (SPP and ASPP) executed in parallel. Additionally, a dedicated block for reconstruction uses the latent space to form an intermediate map, which is then used as supplementary information for the network's final output. This incorporation of contextual details enhances the reliability and accuracy of the segmentation results.

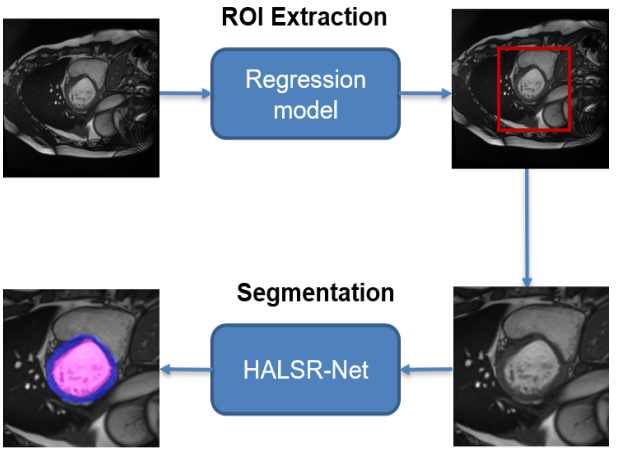


Figure 1: Diagram of the proposed methodology: ROI extraction using a regression model followed by segmentation with HALSR-Net. The regression model isolates the left ventricle and the HALSR-Net performs detailed segmentation of the endocardium and epicardium.

3.1. Regression Model for ROI extraction

The first phase of the proposed methodology focuses on extracting the ROI from the input MRI images, isolating the LV. This phase involves a regression model to predict the coordinates of the bounding box surrounding the ROI, providing precise and efficient LV localization, crucial for subsequent segmentation.

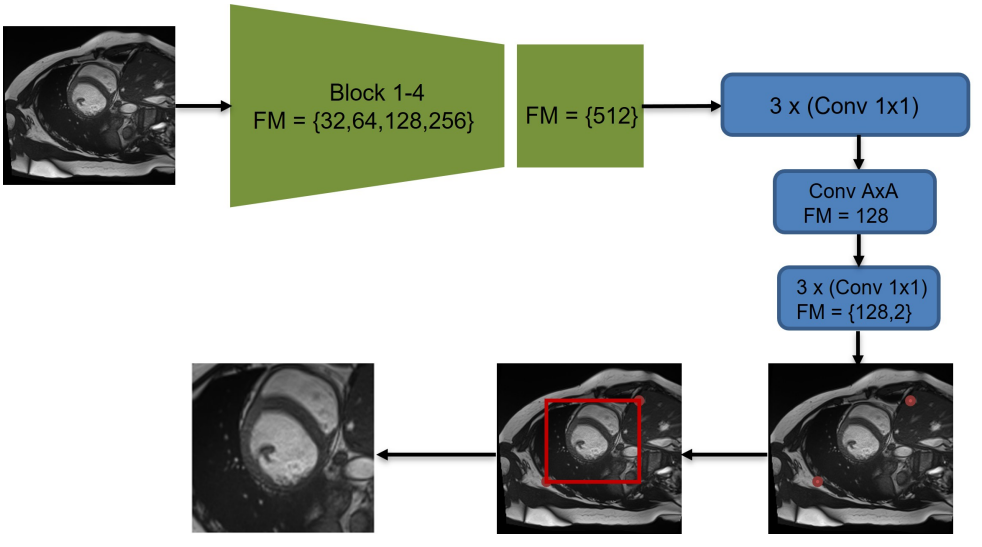


Figure 2: Diagram illustrating the ROI extraction process using the regression model. The model predicts the bounding box coordinates of the ROI within the MRI images.

As illustrated in Fig. 2, the architecture for ROI extraction comprises several convolutional layers organized into an encoder-decoder structure. The encoder consists of four blocks with feature maps of sizes 32, 64, 128, and 256, respectively, designed to progressively extract and compress features from the input image. Following the encoder, a series of fully connected layers is employed to refine the feature representation, up to a feature map of size 512.
In the decoder phase, three convolutional layers (Conv 1×1) are initially used to reduce the dimensionality in the filter space. This is followed by a convolutional layer with a larger kernel size (Conv $A \times A$), where the filter size A is equal to the dimensions of that layer's input, producing feature maps of size 128. This enables the model to capture more

complex spatial relationships, facilitating the extraction of global features. Finally, another set of three convolutional layers (Conv 1×1) further processes these feature maps, resulting in four ouputs for the coordinates defining the ROI bounding box.

The loss function used for training this regression model is the Mean Squared Error (MSE, Eq.(1)), which measures the average squared difference between the predicted and actual bounding box coordinates.

$$MSE = \frac{1}{4}\sum_{i=1}^{4}(y_i - \hat{y}_i)^2, \tag{1}$$

where y_i represents the actual coordinate values and $\hat{y}_i$ represents the predicted coordinate values.

By using this regression approach, the model achieves high accuracy in ROI localization while maintaining computational efficiency. This effectively addresses the challenges posed by the anatomical variability in cardiac MRI images, ensuring that the critical regions are accurately isolated for subsequent segmentation.

3.2. HALSR-Net Architecture

The HALSR-Net architecture employs an encoder-decoder structure to process cine MRI images, efficiently capturing multi-scale features for the segmentation of the endocardium and epicardium.

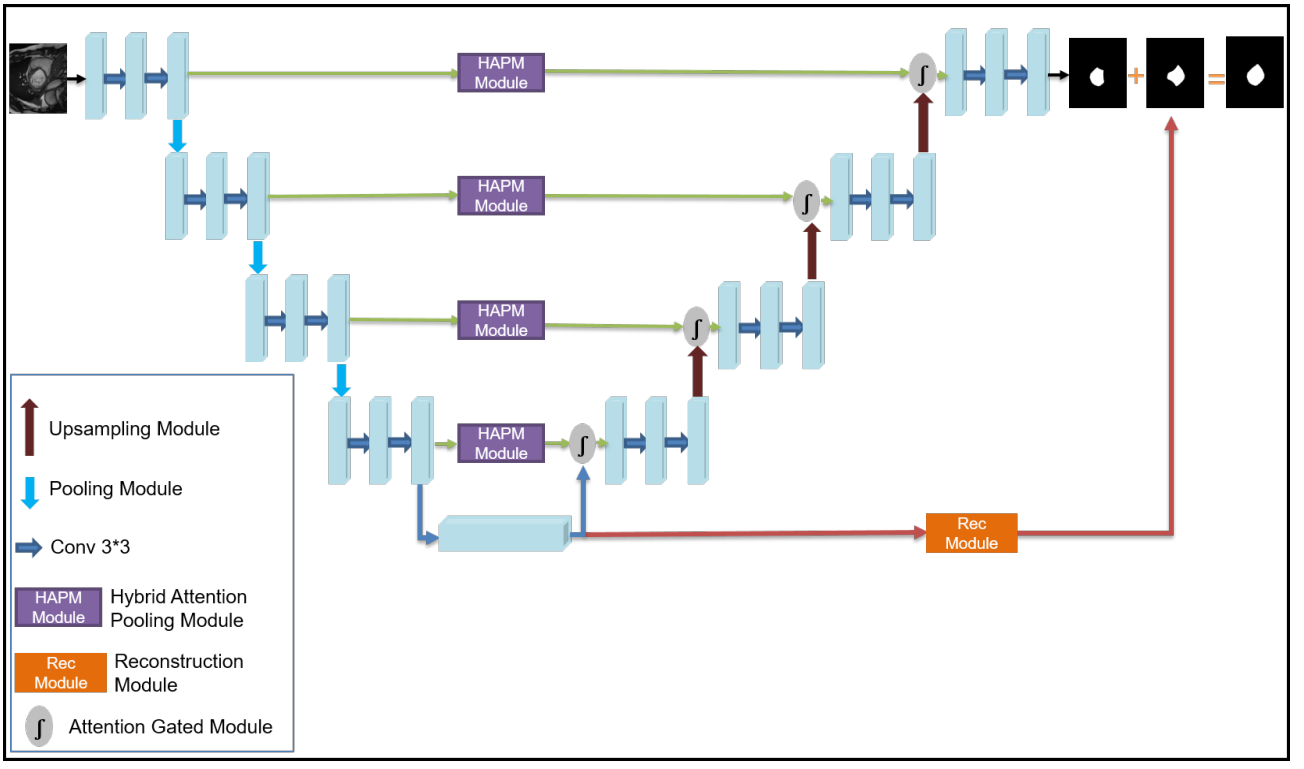


Figure 3: The proposed HALSR-Net architecture for segmenting the endocardium and epicardium. The architecture includes multiple levels of encoding and decoding with integrated Hybrid Attention Pooling Modules (HAPM) and a dedicated reconstruction module to enhance segmentation accuracy.

As depicted in Fig. 3, the encoder consists of five blocks with increasing numbers of filters 64, 128, 256, 512, and 1024. These blocks progressively extract and compress features from the input image, enabling a hierarchical representation of features. Between the encoder and decoder units, the Hybrid Attention Pooling Module (HAPM) is integrated. This module enhances feature extraction and captures richer contextual information. This dual approach allows the model to effectively capture complex spatial relationships and enrich feature representations. Additionally, the architecture's design promotes robust feature learning, thereby alleviating issues such as overfitting.

By combining features at multiple levels, the HAPM module reduces the semantic gap between shallow and deep features, so that both high-level contextual information and fine-grained details are retained. In the decoder, which is structured inversely to the encoder, the feature maps are gradually upsampled to reconstruct the segmented image. Each decoding block incorporates the HAPM module to preserve detailed spatial information and achieve accurate segmentation.

Additionally, HALSR-Net features a dedicated reconstruction module. Unlike other approaches that incorporate reconstruction blocks within each stage of the decoder —potentially leading to information loss and affecting prediction quality— our method uses the latent space to generate an intermediate map. This map serves as supplementary information to the network's final output. The latent space retains the most comprehensive and detailed information, making the reconstruction process both precise and contextually rich. This approach contrasts with methods such as Cai and Wang (2022), where reconstruction at each decoder block can degrade prediction quality due to cumulative information loss. By focusing reconstruction efforts on the latent space, HALSR-Net preserves crucial details, thereby enhancing the reliability and accuracy of the segmentation results. At the final stage, we combine the network's output feature map with the reconstruction map through element-wise addition, which further refines the segmentation by integrating both sources of information.

3.2.1. Attention Gated Module

In the decoder part of HALSR-Net, the attention gate mechanism is employed to improve feature refinement by selectively emphasizing relevant regions, such as the endocardium and epicardium, while suppressing background or non-target elements. As shown in Fig. 4, this mechanism allows the network to focus on the most informative features, enhancing the accuracy of segmentation. The improvement in feature maps is reflected in a clearer distinction between target and non-target areas, progressively refining segmentation quality as features pass through each layer of the decoder.

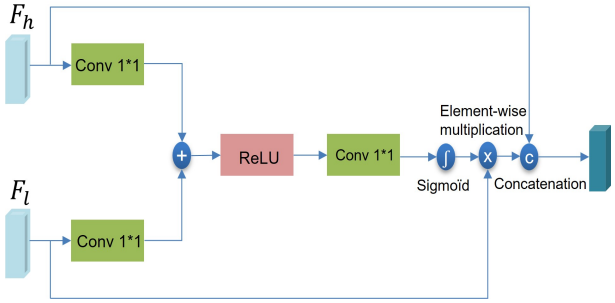


Figure 4: Network framework of the attention gated module, showing the combination of convolutions, activations, and attention mechanisms to enhance feature refinement.

Specifically, the feature map F_h of the current layer of the decoder and the feature map F_l of the corresponding layer in the encoder are used as two branch inputs to the attention gated mechanism. Initially, F_l is adjusted in resolution by upsampling, denoted by (UP(F_l)), followed by two parallel 1×1 convolutions applied to F_h and F_l to obtain F_H and F_L respectively :

$$F_H = \text{Conv}_{1\times1}(F_h), \tag{2}$$

$$F_L = \text{Conv}_{1\times1}(\text{UP}(F_l)). \tag{3}$$

These two features are added point by point to obtain the intermediate feature F_{int} followed by a ReLU activation. Another 1×1 convolution and a sigmoid activation are applied to produce the feature weight:

$$F_\sigma = \sigma(\text{Conv}_{1\times1}(\text{ReLU}(F_H + F_L))). \tag{4}$$

Finally, the output F_{out} is obtained by multiplying F_l by the calculated feature weight and concatenating it with F_h:

$$F_{out} = \text{Concat}(F_\sigma \times F_l, \text{Conv}_{1\times1}(F_h)). \tag{5}$$

This attention gate mechanism ensures that the features contributing to the target region are emphasized, enhancing the accuracy of the segmentation.

3.2.2. Hybrid Attention Pooling Module

As illustrated in Figure 5, the Hybrid Attention Pooling Module in the HALSR-Net architecture significantly enhances feature extraction by integrating advanced techniques such as Atrous Spatial Pyramid Pooling (ASPP), Spatial Pyramid Pooling (SPP), Channel Attention (CA), and Spatial Attention (SA).

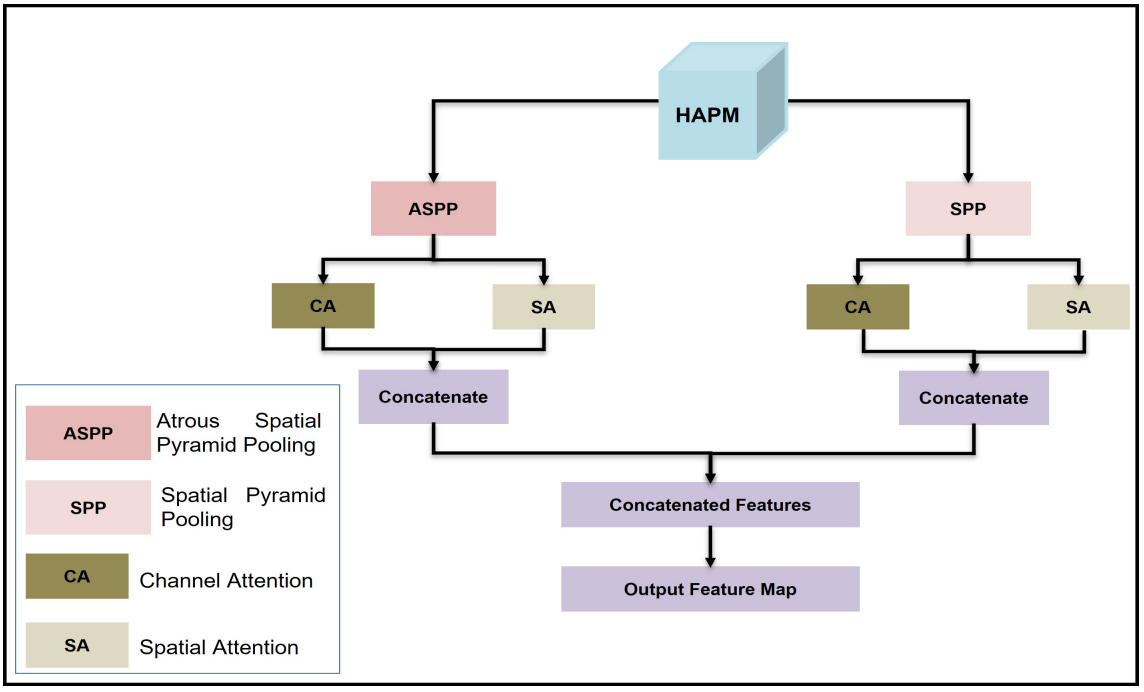


Figure 5: Diagram of the Hybrid Attention Pooling Module (HAPM) in the HALSR-Net architecture. The module integrates Atrous Spatial Pyramid Pooling (ASPP), Spatial Pyramid Pooling (SPP), Channel Attention (CA), and Spatial Attention (SA) to enhance feature extraction and contextual information.

ASPP uses atrous convolution to capture multi-scale information, enabling a larger receptive field without significantly increasing the number of parameters. This helps in capturing spatial context at various resolutions, which is essential for segmenting complex structures. SPP, on the other hand, divides the input feature map into multiple sub-regions and pools information from each, effectively capturing spatial hierarchies. This improves the model's ability to recognize objects at different scales and locations within the image while keeping important spatial information.
The attention mechanisms used in HAPM also include methods to re-weight the importance of different channels within the feature map, allowing the network to prioritize more informative channels and suppress less useful ones. This enhances the feature representation by emphasizing the most relevant features. Meanwhile, spatial attention highlights critical spatial locations in the feature map, enabling the network to focus on specific regions that are crucial for the task, leading to more precise and reliable segmentation results.
By combining the outputs of these techniques in a parallel configuration, the HAPM creates a comprehensive feature representation. This integration captures both global and local contextual information as well as spatial and channel-wise dependencies, leading to more robust and accurate segmentation results.

3.2.3. Reconstruction Module

The Reconstruction Module in our proposed architecture is crucial for enhancing the accuracy and reliability of the segmentation results. This module leverages the latent space to generate an intermediate map, which serves as supplementary information to the network's final output. By focusing the reconstruction efforts on the latent space,

the module ensures that the most comprehensive and detailed information is retained, significantly improving the segmentation quality.

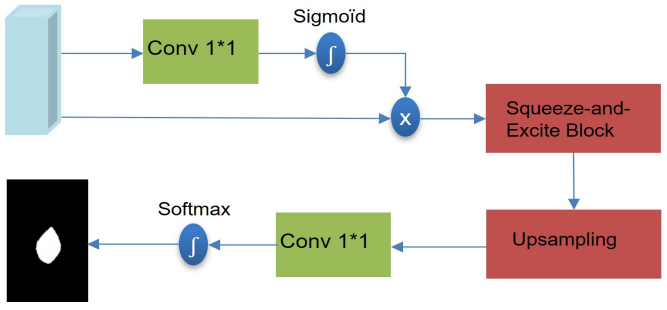


Figure 6: Diagram of the Reconstruction Module, showcasing the use of attention mechanisms, Squeeze-and-Excite blocks, and upsampling to enhance the feature map for accurate segmentation.

As depicted in Fig. 6, the Reconstruction Module begins with the application of an attention mechanism to the intermediate layer. This attention mechanism uses a 1×1 convolution followed by a sigmoid activation to generate an attention map. This map is then multiplied with the intermediate layer to enhance the relevant features, ensuring that important spatial details are highlighted.

Following the attention mechanism, the Squeeze-and-Excite (SE) block is applied to the intermediate layer. The SE block enhances feature representation by explicitly modeling the inter-dependencies between channels. The SE block operates in two main steps: *squeeze* and *excite*.

The *squeeze* step in the SE block involves global average pooling to compress the input feature map into a channel descriptor, capturing global spatial information. In the *excite* step, this descriptor passes through fully connected layers with ReLU and sigmoid activations to learn and scale channel-wise dependencies. The resulting feature map is scaled by multiplying with the attended intermediate layer, emphasizing informative channels and suppressing less useful ones. The SE-enhanced feature map is then upsampled using bilinear interpolation to align with the final output size, preserving resolution and details. A 1×1 convolution with softmax activation converts this map into a precise, contextually rich probability map over the classes of interest. The SE block significantly enhances the reconstruction process by focusing on critical features, ensuring high-quality feature representations crucial for accurate segmentation.

To optimize the segmentation performance of HALSR-Net, the loss function (Eq. (8)) is a combination of Binary Cross-Entropy (BCE) (Eq.(6)) and Dice Loss (DL) (Eq. (7)). DL measures the overlap between the predicted and ground truth masks, while BCE measures the binary classification error at each pixel:

$$\begin{aligned} &\mathrm{BCE}(y_{\mathrm{true}}, y_{\mathrm{pred}}) = \\ &-\frac{1}{N}\sum_{i=1}^{N}[y^i_{\mathrm{true}}\log(y^i_{\mathrm{pred}}) + (1-y^i_{\mathrm{true}})\log(1-y^i_{\mathrm{pred}})], \end{aligned} \tag{6}$$

$$\mathrm{DL}(y_{\mathrm{true}}, y_{\mathrm{pred}}) = 1 - \frac{2\sum y_{\mathrm{true}} y_{\mathrm{pred}} + \epsilon}{\sum y_{\mathrm{true}} + \sum y_{\mathrm{pred}} + \epsilon}, \tag{7}$$

where ϵ is a small constant to avoid division by zero. The combined BCE-DL loss is then defined as:

$$\begin{aligned} &\mathrm{BCE\text{-}DL}(y_{\mathrm{true}}, y_{\mathrm{pred}}) = \\ &\alpha\mathrm{BCE}(y_{\mathrm{true}}, y_{\mathrm{pred}}) + \beta\mathrm{DL}(y_{\mathrm{true}}, y_{\mathrm{pred}}), \end{aligned} \tag{8}$$

where α and β are the weights assigned to the BCE and DL components, respectively. This combined loss function balances pixel-wise accuracy and overlap, leading to more accurate and robust segmentation results.

Table 1
Details of the datasets used with training, validation, and test sets repartition.

Dataset	Train set	Validation set	Test set
In-house	2268	750	223
ACDC	1402	460	1062
LVQuan19	789	250	122

4. Experiment results

In this section, we present the results of our experiments conducted on three different cardiac MRI datasets: one private in-house clinical dataset and two public datasets (see section 4.1). We compare our method against eight state-of-the-art architectures that are widely recognized for their performance: U-Net Ronneberger et al. (2015), Attention Gates Oktay et al. (2018), Non-local U-Net Wang et al. (2020), UNet++ Zhou et al. (2018), SegNet Badrinarayanan et al. (2017), U-SegNet Dangi et al. (2019), ResUNet Alom et al. (2018), ResUNet++ Jha et al. (2021), and 2D nnU-Net Isensee et al. (2021). Of note, we were not able to include some of the most recent architectures in our tests due to the unavailability of their source codes and the complexity of their implementations.

4.1. Datasets

For the construction, training, and validation of our models, three different cardiac MRI datasets were used as shown in Table 1. The in-house dataset consists of cine images of 30 subjects. Images of the endocardium and epicardium at end-diastole (ED) and end-systole (ES) were manually delineated, resulting in a total of 3241 images. 2268 images were used for model training, while 750 images were used for validation, and 223 images for testing.
The second dataset was originally created for the Automated Cardiac Diagnosis Challenge (ACDC) Bernard et al. (2018). It comprises cine scans of 100 subjects across five different disease groups. The images have varying fields of view (FOV) and acquisition matrices, with reconstructed pixel sizes ranging from 0.72×0.72 mm^2 to 1.92×1.92 mm^2. Ground truth segmentations for the left ventricle, myocardium, and right ventricle (RV) are available for ED and ES. This resulted in a total of 2924 delineated 2D short-axis (SA) images. 1402 images were used for model training, while 460 images were used for validation, and 1062 images for testing.
The third dataset was obtained from the Left Ventricular Quantification Challenge held at STACOM'19 (LVQuan19 [1]). This dataset contains mid-cavity SA slices for the complete cardiac cycle of 56 patients, with a wide variety of image sizes, pixel sizes, and image appearances. Ground truth segmentations of the LV and myocardium, represented as images, and ground truth values for 11 cardiac measures (LV and myocardial area, three LV dimensions, and six regional wall thicknesses) are provided for all 20 time points, comprising a total of 1161 delineated 2D SA images. 789 images were used for model training, 250 for validation, and 122 for testing.

4.2. Evaluation metrics

To measure the performance of the proposed method as well as compare it with state-of-the-art and competing methods, we used the following segmentation evaluation metrics:

Dice Coefficient:

The Dice Coefficient (DC) measures the overlap between two sets and is defined as:

$$DC = \frac{2|A \cap B|}{|A| + |B|}. \tag{9}$$

where A is the set of predicted pixels and B is the set of ground truth pixels.

Hausdorff Distance

The Hausdorff Distance (HD) quantifies how far two subsets of a metric space are from each other. It is defined as:

$$HD(A, B) = \max\left\{ \sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(b, a) \right\}. \tag{10}$$

[1] https://lvquan19.github.io/

Table 2
Parameter Settings for the experiments.

Parameters		Value
Adam optimizer	Learning rate	10^{-3}
	β_1	0.9
	β_2	0.999
	ϵ	10^{-8}
Batch size		32
Dropout		0.30 to 0.50

where $d(a, b)$ is the Euclidean distance between points a and b.

Average Symmetric Surface Distance The Average Symmetric Surface Distance (ASSD) is the average of the shortest distances from points on the contour/surface of one set to the contour/surface of another set. It is defined as:

$$ASSD(A, B) = \frac{1}{2}\left(\frac{1}{|A|}\sum_{a\in A}\min_{b\in B} d(a, b) + \frac{1}{|B|}\sum_{b\in B}\min_{a\in A} d(b, a)\right). \tag{11}$$

where $d(a, b)$ is the Euclidean distance between points a and b.

4.3. Parameter settings

The HALSR network model proposed in this paper is implemented using the Keras and TensorFlow frameworks. Model training and evaluation are performed on a compute node with 128 cores and a NVIDIA™ H100 Tensor Core GPU 96GB.

Training of the model uses early stopping. During the training process of both the proposed and competing networks, we use the ADAM optimizer with a learning rate set to 10^{-3}, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for moment estimates and $\epsilon = 10^{-8}$ for numerical stability. These parameters were chosen based on prior experimentation, as these values offered a stable convergence for the models under different datasets and training conditions. The dropout rate was determined through cross-validation, ranging between $p = 0.30$ and $p = 0.50$, based on its ability to prevent overfitting while maintaining a sufficient model flexibility. The specific parameter settings are shown in Table 2. These choices, in conjunction with early stopping, allow for an efficient training process and prevent overfitting, which is crucial for achieving high generalization performance.

The weights of BCE and DL in the combined loss BCE-DL are $\alpha = 0.5$ and $\beta = 0.5$ respectively. These values were chosen to balance pixel-wise classification accuracy and segmentation overlap, ensuring stable convergence across different datasets and training conditions.

4.4. Performance of the Regression Model for ROI Extraction

To evaluate the regression-based ROI extraction model, we conducted a detailed analysis of its performance across the three datasets. The results are summarized in Table 3.

The regression model demonstrated robust localization performance across all datasets, achieving a small distance between predicted and true ROI points. This distance is quantified using the Root Mean Squared Error (RMSE), which represents the average deviation between the predicted and ground truth ROI locations, defined by the bottom-left and top-right points. A lower RMSE indicates higher localization accuracy. The values remain below 0.7 mm across all datasets, confirming the model's ability to generalize across different anatomical structures. The corresponding Standard Deviation (SD) further reflects the stability and reliability of the model's predictions, as lower SD values indicate more consistent performance across samples.

To establish the ground truth ROIs, we computed the bounding box encompassing the reference segmented structure and extracted its two extremal points (bottom-left and top-right corners) to ensure consistency with the definition above. These points serve as the reference ROIs for comparison with the predicted ones. This methodology guarantees a uniform evaluation across datasets while preserving anatomical constraints.

However, in extreme cases, prediction errors can occur. The outlier rate reaches approximately 4%, as reported in Table

Table 3
Evaluation of the regression model for ROI extraction on in-house, ACDC, and LVQuant19 datasets with Root Mean Squared Error (RMSE), corresponding Standard Deviation (SD), and outlier rate.

Dataset	RMSE ± SD (mm)	Outlier Rate (%)
In-house	0.50 ± 0.3	3.8
ACDC	0.61 ± 0.4	4.1
LVQuant19	0.65 ± 0.5	4.2

Table 4
Experiment 1: Accuracy (Acc) and Loss obtained for train, validation and test sets on the in-house dataset.

CNNs	Years	Acc_{train}	$Loss_{train}$	Acc_{val}	$Loss_{val}$	Acc_{test}	$Loss_{test}$
U-Net	2015	0.95	0.08	0.93	0.08	0.92	0.06
Attention Gates	2018	0.95	0.06	0.95	0.06	0.93	0.09
Non-local U-Net	2020	0.90	0.09	0.89	0.13	0.84	0.29
UNet++	2018	0.89	0.10	0.87	0.12	0.85	0.23
SegNet	2017	0.95	0.06	0.94	0.09	0.91	0.11
U-SegNet	2019	0.95	0.06	0.94	0.06	0.91	0.11
ResUNet	2018	0.90	0.10	0.89	0.11	0.87	0.15
ResUNet++	2021	0.94	0.10	0.92	0.12	0.89	0.16
nnU-Net	2021	0.94	0.07	0.94	0.06	0.93	0.09
HALSR-Net	-	**0.96**	**0.04**	**0.95**	**0.04**	**0.95**	**0.06**

3. In instances where the predicted ROI is not visually satisfactory, manual correction is applied to ensure accurate segmentation.

4.5. Experiment 1: in-house dataset

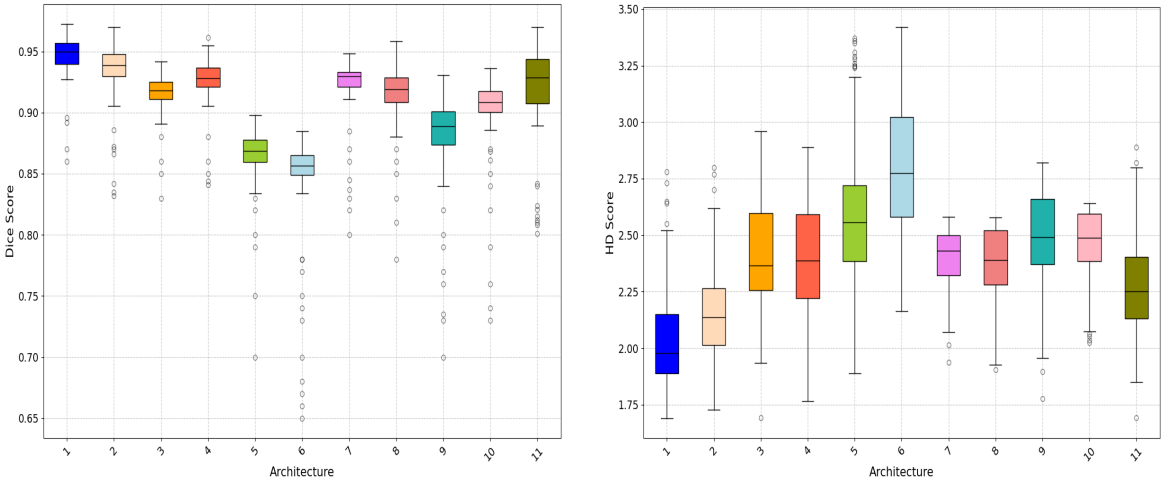
In this experiment, we evaluated the performance of our proposed HALSR-Net on an in-house dataset, which includes cine MRI images with manually delineated endocardium and epicardium at ED and ES. The dataset was split into training, validation, and test sets with 2268, 750, and 223 images, respectively. The results in Table 4 demonstrate that HALSR-Net achieved a training accuracy of 0.96 and a validation accuracy of 0.95, with low training and validation losses of 0.04. The test accuracy of 0.95 and test loss of 0.06 further confirm the robustness of our model. In comparison, models such as ResUNet++, SegNet, and UNet++ showed higher test losses and lower test accuracies.

The boxplot analysis of the DC, HD, and ASSD metrics (Figs. 7(a), 7(b), and 7(c)) provides deeper insights into model performance variability. The highest median DC of 0.95 with minimal variability, achieved by HALSR-Net (with ROI), indicates precise and consistent segmentation performance, demonstrating the model's stability in accurately capturing the target regions across different samples. A high Dice score implies a close match between the predicted segmentation and the ground truth, crucial for accurately identifying the target regions. In contrast, models such as Non-local U-Net and UNet++, which integrate non-local operations and dense skip connections, showed greater variability with median Dice scores of 0.87 and 0.86 respectively, suggesting less consistent performance. Among existing semantic segmentation architectures, nnU-Net achieves the highest median Dice score, confirming its effectiveness in learning robust feature representations and achieving a precise match with the ground truth. This superior performance can be attributed to its self-adjusting architecture, which optimizes hyperparameters and network configurations based on the dataset, as well as its advanced preprocessing and data augmentation strategies, which enhance its generalization capabilities and adaptability to inter-patient variations. Furthermore, its optimized inference pipeline ensures efficient and consistent segmentation, reducing structural inconsistencies observed in other architectures.

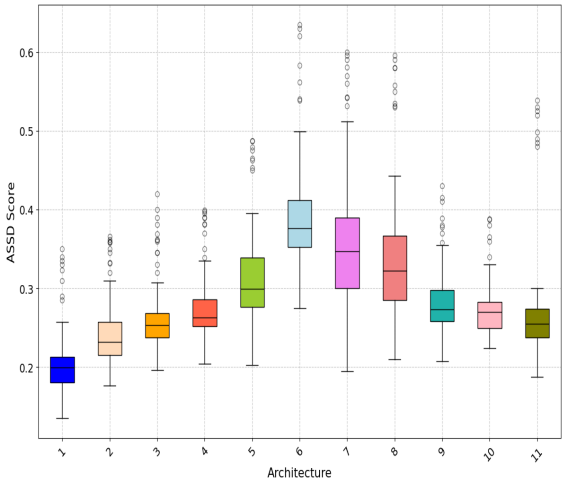
Moreover, the lowest median HD of around 2.0 pixels of HALSR-Net reflects the model's precision in boundary delineation, crucial for accurately identifying cardiac structure edges and reducing boundary errors. Competing architectures displayed higher median HD values, indicating less precise boundary detection. Specifically, the median HD for Non-local U-Net and UNet++ were higher. A low HD is particularly important for tasks where boundary accuracy is paramount, such as in detailed anatomical studies or pre-surgical planning. For semantic architectures, nnU-Net also demonstrates a relatively low HD compared to other methods in this category, further reinforcing its

ability to capture structural boundaries with high accuracy and reduce segmentation inconsistencies. In addition, the smallest median ASSD observed was almost 0.2 pixels for HALSR-Net, emphasizing the model's ability to maintain minimal contour/surface distance errors between predicted and ground truth segmentations. Larger ASSD values in other models suggest greater contour errors, which could impact the accuracy and reliability of segmentation in clinical practice. Although nnU-Net stands out as a very effective semantic segmentation model, our approach demonstrates superior accuracy, particularly in terms of contour precision and robustness. Unlike nnU-Net, which benefits from extensive preprocessing, 5-fold cross-validation, and an optimized inference pipeline, our method, leveraging the Hybrid Attention Pooling Module (HAPM) and a latent space-based reconstruction block, achieves high segmentation accuracy without requiring additional preprocessing or post-processing steps.

To further investigate the impact of the ROI extraction module, we implemented two versions of our model: HALSR-Net (with ROI) and HALSR-Net (without ROI). The results show that both versions exhibit homogeneous performance, with comparable median values across all metrics. However, the model without ROI demonstrates slightly higher variability. This increased variability can be attributed to the lack of ROI constraints, which may introduce additional



(a): Dice (b): HD



(c): ASSD

Figure 7: Boxplots of metrics for the in-house dataset: (a) Dice score, (b) Hausdorff distance (HD, pixels), (c) Average Asymmetric Surface Distance (ASSD, pixels). The methods using preliminary ROI detection are marked with (*): 1. HALSR-Net (with ROI), 2. HALSR-Net (without ROI), 3. U-Net, 4. Att. gates*, 5. Non-local U-Net, 6. UNet++, 7. SegNet, 8. U-SegNet*, 9. ResUNet, 10. ResUNet++*, 11. nnU-Net.

Table 5
Experiment 2: Accuracy (Acc), and Loss obtained for train, validation and test sets on the ACDC dataset.

CNNs	Years	Acc_{train}	$Loss_{train}$	Acc_{val}	$Loss_{val}$	Acc_{test}	$Loss_{test}$
U-Net	2015	0.98	0.02	0.98	0.03	0.98	0.03
Attention Gates	2018	0.98	0.09	0.97	0.15	0.97	0.11
Non-local U-Net	2020	0.98	0.05	0.97	0.07	0.97	0.09
UNet++	2018	0.93	0.05	0.92	0.08	0.91	0.08
SegNet	2017	0.98	0.12	0.97	0.15	0.96	0.17
U-SegNet	2019	0.98	0.08	0.97	0.08	0.97	0.08
ResUNet	2018	0.95	0.14	0.95	0.16	0.94	0.18
ResUNet++	2021	0.97	0.05	0.97	0.07	0.96	0.08
nnU-Net	2021	0.98	0.05	0.98	0.05	0.96	0.07
HALSR-Net	-	**0.98**	**0.02**	**0.98**	**0.03**	**0.98**	**0.03**

background information, making segmentation more sensitive to spatial inconsistencies. The ROI module effectively refines the focus on the relevant cardiac structures, reducing unnecessary feature activations and improving boundary delineation stability.
Regarding the competing methods, it is important to note that some of them also incorporate ROI extraction techniques. For instance, models such as Attention Gates, U-SegNet, and ResUNet++ utilize ROI-based mechanisms to enhance their segmentation performance (as quoted in Fig. 7). This confirms the relevance of ROI-based strategies in improving segmentation accuracy and robustness.

Qualitative analysis of the visual segmentation outcomes further supports the quantitative findings, as demonstrated in Fig. 8. Our architecture accurately delineates the LV cavity (red) and myocardium (blue) with minimal errors, effectively avoiding segmentation of non-targeted elements such as papillary muscles. This level of precision is challenging for other models. For instance, ResUNet++ occasionally included papillary muscles into the myocardium segmentation, leading to inaccuracies. SegNet, while maintaining boundary details reasonably well, struggled with accuracy, particularly in complex regions.
In comparison, models like U-Net and UNet++ do not achieve the same level of detail and accuracy. U-Net, while influential, does not utilize attention mechanisms and shows inconsistencies in capturing the full extent of the myocardium, resulting in partial segmentations. While the attention gates model improves focus on relevant regions, it still misses fine details and struggles with complex anatomical structures, resulting in less precise segmentation. The Non-local U-Net, designed to capture long-range dependencies, also fell short in effectively segmenting complex regions, often including unwanted elements and reflecting difficulty in distinguishing between similar textures and boundaries in cardiac MRI images. Additionally, U-SegNet and SegNet, which employ pooling indices and distance map regularization, show reasonable boundary preservation but lack the fine accuracy of detail seen in our architecture. These models are less able to manage the variability in cardiac structures, leading to occasional segmentation errors.
Notably, nnU-Net, recognized for its adaptability and strong generalization capabilities across various medical imaging tasks, achieves results that are comparable to those of our proposed method. Its self-configuring pipeline effectively optimizes preprocessing, architecture selection, and post-processing, allowing for competitive segmentation performance. However, despite its adaptability, nnU-Net exhibits limitations in capturing finer anatomical details. In contrast, our approach leverages a HAPM to enhance feature refinement and a latent space-based reconstruction block to improve structural consistency. These components allow the model to focus on clinically relevant features while preserving local and global coherence, leading to more precise delineation of cardiac structures. By explicitly integrating these mechanisms, our method surpasses traditional architectures in capturing fine-grained anatomical information without relying on extensive preprocessing or post-processing steps. This highlights the advantage of combining attention-driven feature extraction with structural refinement, ensuring anatomically consistent and high-fidelity segmentations.

4.6. Experiment 2: ACDC dataset

In this experiment, we evaluated the performance of our model on the public ACDC dataset, which includes cine MRI images with manually delineated endocardium and epicardium at ED and ES. The results are presented in Table 5. Our model demonstrated similar performance with a training and validation accuracy achieving up to 0.98.

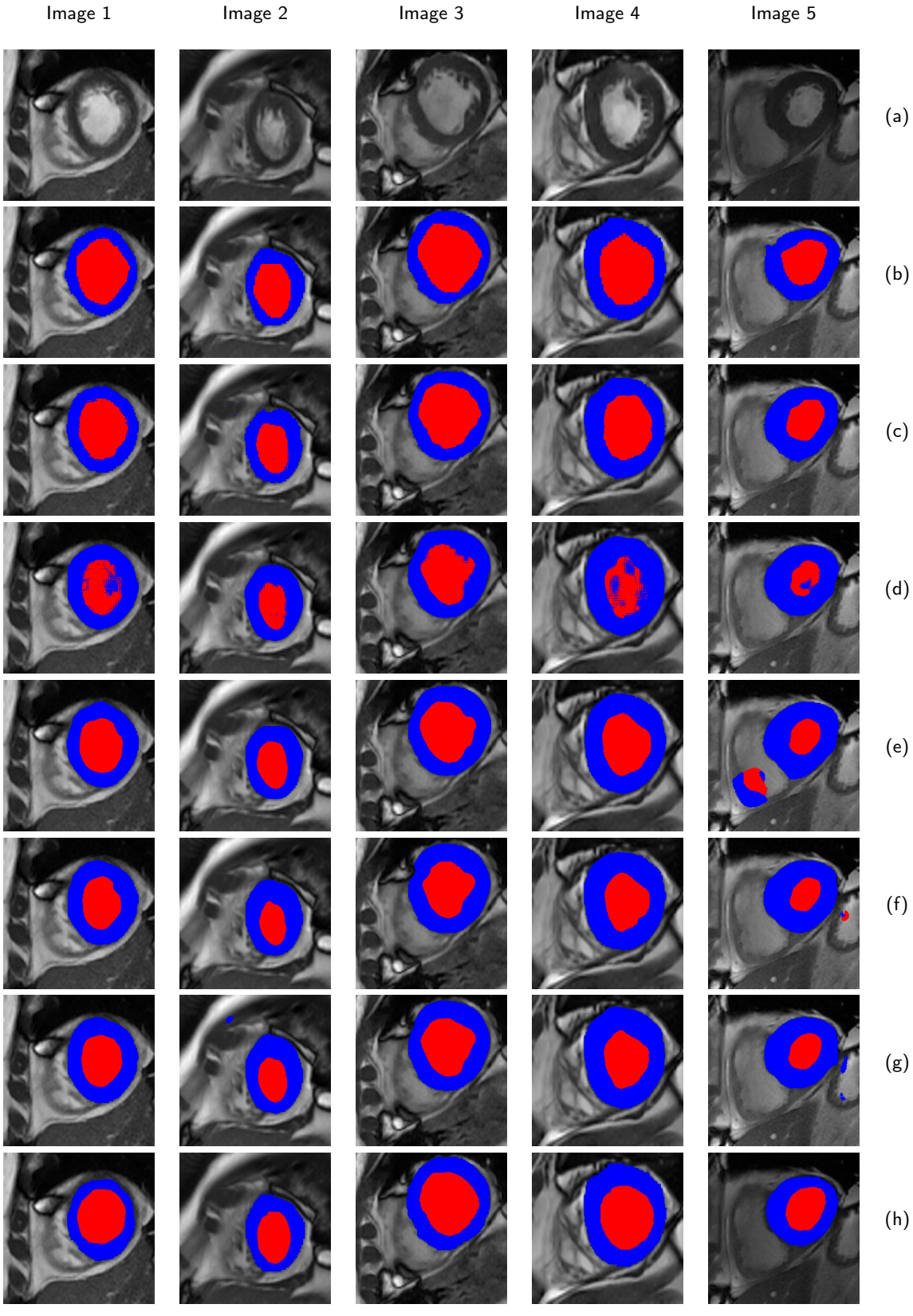
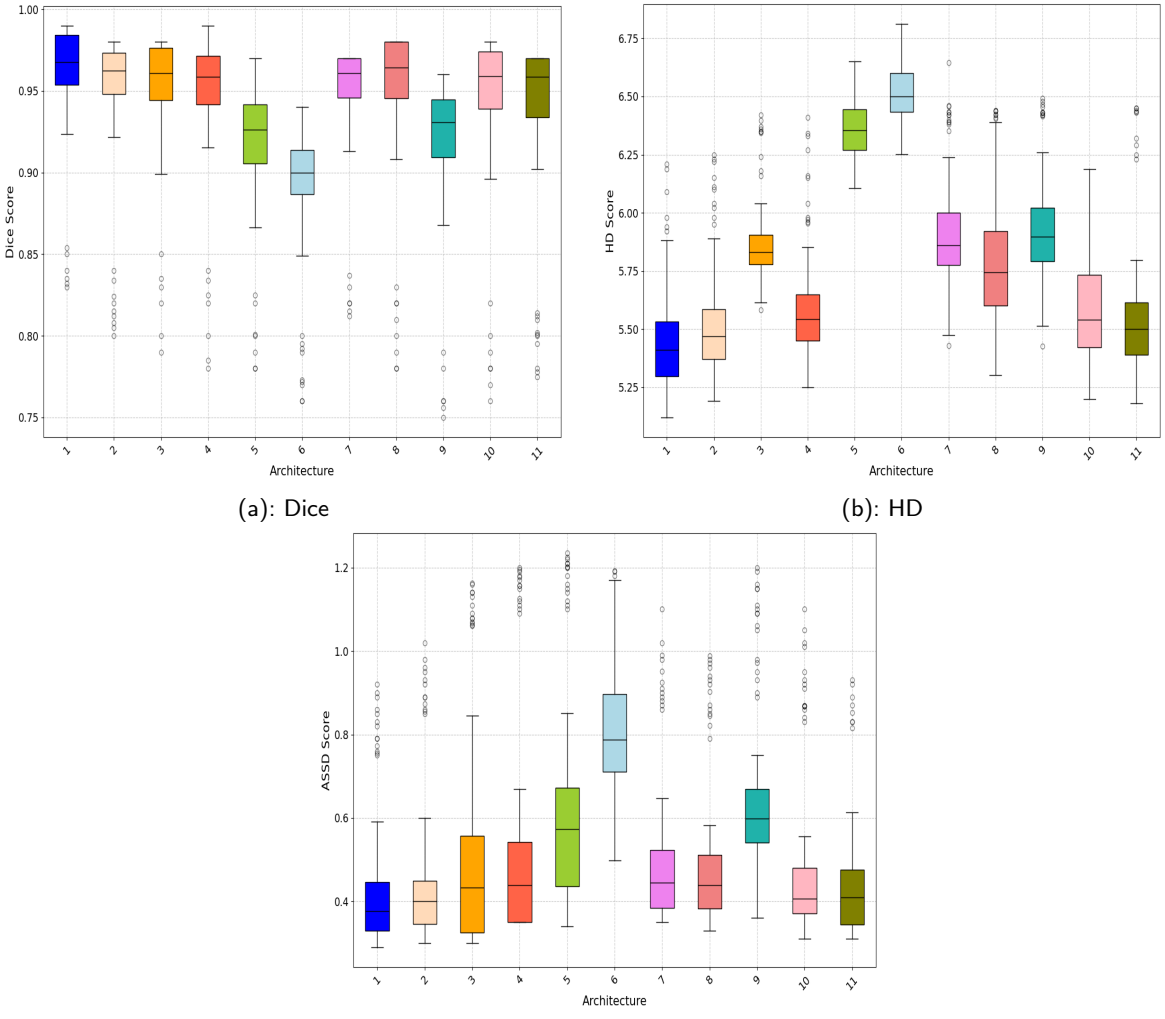


Figure 8: Segmentation of LV cavity (red) and Myocardium (blue) : visual outcomes demonstrated for the in-house dataset compared to the most efficient competing architectures: (a) original image, (b) ground truth, (c) Proposed HALSR-Net, (d) ResUNet++, (e) segNet, (f) U-Net, (g) Attention Gates, (h) nnU-Net.

Table 6
Comparative Performance Analysis on the ACDC dataset for our method and recent approaches.

CNNs	Years	Dice	HD	ASSD
CFNet Zhan et al. (2023)	2021	0.940	6.71	-
DSGA-Net Sun et al. (2023)	2023	0.968	-	-
ARW-Net Singh et al. (2023)	2023	0.967	5.65	-
HALSR-Net	-	**0.978**	**5.11**	**0.36**

In the competing architectures, UNet++ integrates nested, dense skip pathways that enhance information flow across the network. However, despite these improvements, it achieved a lower median DC around 0.90 and higher HD (Fig. 9). This suggests that while UNet++ improves on the basic U-Net by facilitating better information flow, it struggles with precise boundary detection and maintaining contour accuracy in complex regions, possibly due to overfitting or difficulties in capturing fine details. Similarly, SegNet uses pooling indices for upsampling, which helps



(a): Dice (b): HD

(c): ASSD

Figure 9: Boxplots of metrics for the ACDC dataset: (a) Dice score, (b) HD (pixels), (c) ASSD (pixels). The methods using preliminary ROI detection are marked with (*): HALSR-Net (with ROI), 2. HALSR-Net (without ROI), 3. U-Net, 4. Att. gates*, 5. Non-local U-Net, 6. UNet++, 7. SegNet, 8. U-SegNet*, 9. ResUNet, 10. ResUNet++*, 11. nnU-Net.

maintain boundary delineation. Despite this advantage, SegNet showed higher variability in segmentation accuracy and precision. The reliance on pooling indices can lead to less detailed reconstructions in regions with high anatomical variability, resulting in a higher Average ASSD. The pooling and unpooling operations, although efficient, may fail to preserve the fine-grained details necessary for precise medical image segmentation.
Furthermore, nnU-Net emerges as the most robust purely semantic approach, dynamically adapting to dataset characteristics through automated preprocessing and architectural tuning. These capabilities allow it to achieve superior Dice scores and relatively lower boundary errors, making it a strong baseline for medical image segmentation.
Our results show that HALSR-Net achieves robust performance in both configurations (with and without ROI), demonstrating the adaptability of its architecture to different input conditions. The high median Dice score and low HD and ASSD variability in both cases highlight the model's ability to extract and process multi-scale contextual features effectively, regardless of the presence of an explicit ROI constraint. This suggests that HALSR-Net's hybrid attention pooling mechanism, combined with its multi-scale feature representation, allows the model to focus on relevant structures dynamically, ensuring accurate segmentation even in the absence of a ROI module.
The slightly increased variability observed in the metrics for the version without ROI is primarily due to a broader receptive field that includes more background information. However, this does not compromise the overall segmentation accuracy, as the architecture is inherently designed to adapt to spatial inconsistencies and refine structural details.

Likewise, ResUNet++ incorporates residual learning and SE blocks to enhance feature extraction. Although it showed improvements over simpler models, its performance was less consistent, particularly with complex anatomical structures. While the residual connections help in learning deeper features, they may not fully capture the intricate details required for precise segmentation, leading to occasional inaccuracies in boundary delineation and higher error rates.
Non-local U-Net captures long-range dependencies within the image, theoretically improving global context understanding. However, it had a lower median Dice score of around 0.93 and a higher HD, indicating that while the model is effective at integrating global information, it may overgeneralize, leading to less precise boundary delineations in specific regions. The attention gates model focuses on relevant regions using attention mechanisms, which improves focus on target areas but still had difficulties in complex regions, resulting in less precise segmentation. The median Dice score and HD were lower, reflecting challenges in maintaining detailed accuracy across varied anatomical structures. Balancing global context with local detail is a challenge that HALSR-Net manages more effectively through its hybrid attention mechanisms.
While ResUNet++, Non-local U-Net, and Attention Gates incorporate feature recalibration, long-range dependencies, and attention mechanisms to improve segmentation, their performance remains inconsistent in capturing fine details. Compared to these models, nnU-Net offers a more balanced integration of such strategies, leading to more robust and adaptive performance. However, our results indicate that HALSR-Net, leveraging its hybrid attention pooling mechanism, surpasses purely semantic approaches by ensuring greater structural consistency and boundary precision without relying on extensive preprocessing or post-processing.

Recent architectures such as CFNet Zhan et al. (2023), DSGA-Net Sun et al. (2023), and ARW-Net Singh et al. (2023) showed competitive performance with a slight inferiority of approximately 1% to HALSR-Net with Dice score and 0.5mm with HD metrics (Table 6). CFNet focuses on cross-feature learning, which aids in understanding inter-slice relationships but may miss out on intra-slice details. The complexity of integrating multiple attention mechanisms, as seen in DSGA-Net and ARW-Net, can introduce challenges in balancing these mechanisms effectively, leading to difficulties in achieving optimal model performance. This integration, while theoretically enhancing feature representation, may lead to an overly complex model that struggles with the fine detail required in medical imaging.

Qualitative analysis of the visual segmentation outcomes in Fig. 10 supports the quantitative findings. Our architecture accurately delineates the LV and myocardium with minimal errors, avoiding segmentation of non-target elements. In contrast, ResUNet++ and SegNet struggled with accuracy in complex regions, highlighting the clear advantage of our model in handling intricate anatomical details with greater accuracy and consistency.

4.7. Experiment 3: LVQuan19 dataset

In this experiment, we evaluated the performance of HALSR-Net on the LVQuan19 dataset. The results are presented in Table 7. Despite the dataset's poorer resolution and image quality, we found similar performance to our

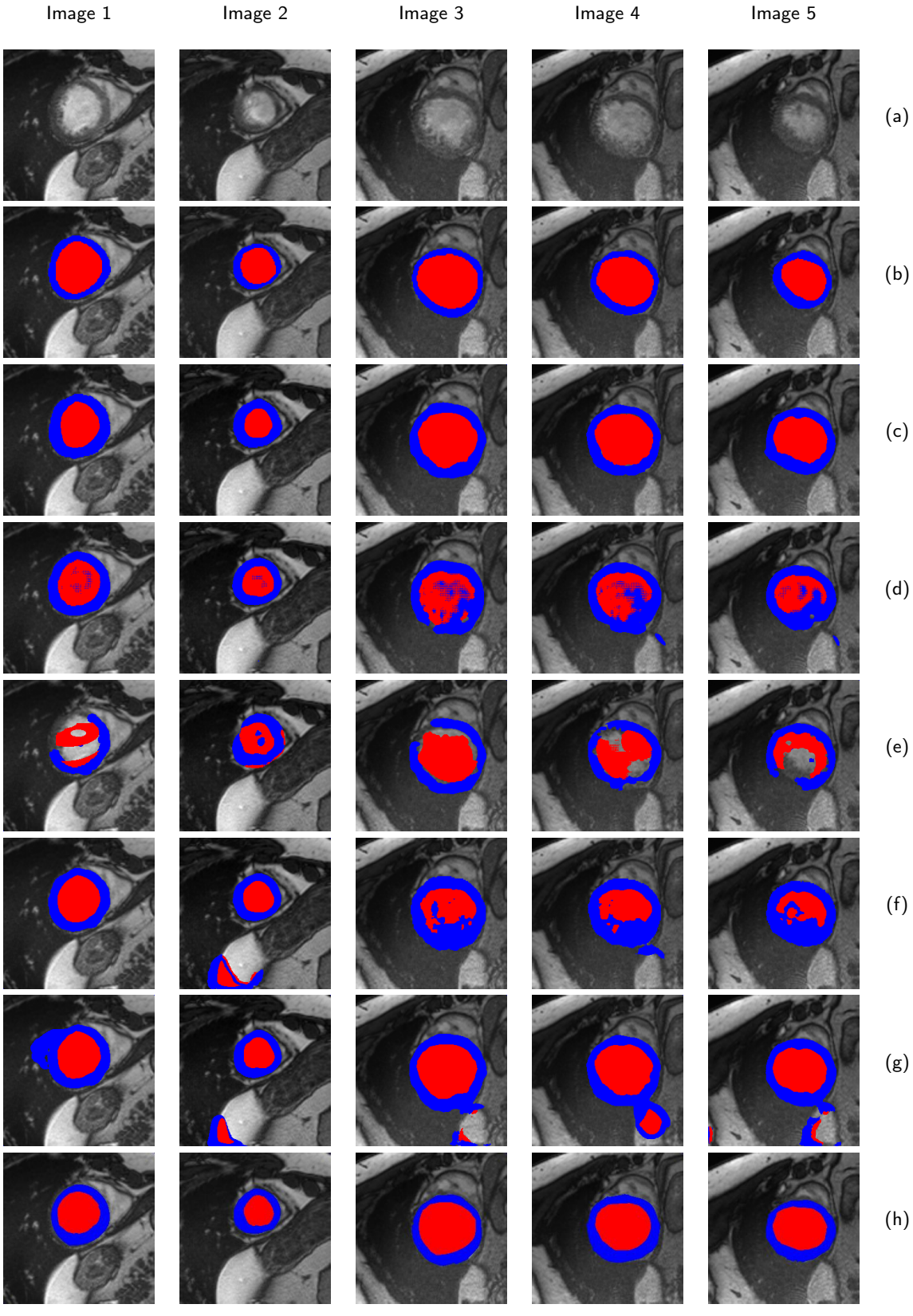


Figure 10: Segmentation of LV cavity (red) and myocardium (blue) : visual outcomes demonstrated for the ACDC dataset Using the most performant competing architectures: (a) original image, (b) ground truth, (c) Proposed HALSR-Net, (d) ResUNet++, (e) segNet, (f) U-Net, (g) Attention Gates, (h) nnU-Net.

Table 7
Experiment 3: Accuracy (Acc) and loss obtained for train, validation and test on the LVQuan19 dataset.

CNNs	Years	Acc_{train}	$Loss_{train}$	Acc_{val}	$Loss_{val}$	Acc_{test}	$Loss_{test}$
U-Net	2015	0.97	0.10	0.96	0.12	0.92	0.18
Attention Gates	2018	0.97	0.08	0.96	0.10	0.93	0.16
Non-local U-Net	2020	0.95	0.09	0.94	0.11	0.90	0.17
UNet++	2018	0.92	0.10	0.92	0.15	0.89	0.21
SegNet	2017	0.90	0.17	0.89	0.19	0.88	0.25
U-SegNet	2019	0.90	0.18	0.89	0.18	0.89	0.24
ResUNet	2018	0.94	0.16	0.93	0.15	0.92	0.23
ResUNet++	2021	0.97	0.08	0.96	0.10	0.93	0.15
nnU-Net	2021	0.97	0.07	0.97	0.08	0.94	0.10
HALSR-Net	-	**0.98**	**0.04**	**0.98**	**0.05**	**0.96**	**0.06**

Table 8
Cross-Dataset Evaluation Results.

CNNs	In-house			ACDC			LVQuan19		
	Dice	HD	ASSD	Dice	HD	ASSD	Dice	HD	ASSD
HALSR-Net trained on IH	**0.95**	**2.00**	**0.20**	0.94	5.55	0.38	0.89	3.45	0.40
HALSR-Net trained on ACDC	0.93	2.22	0.22	**0.97**	**5.40**	**0.36**	0.90	3.50	0.42
HALSR-Net trained on LVQuan19	0.92	2.30	0.25	0.94	5.70	0.39	**0.92**	**3.30**	**0.37**

previous experiments, with an overall accuracy of 0.98 on this challenging dataset. This indicates efficient learning and minimal prediction errors, with HALSR-Net outperforming other models such as U-Net, Attention Gates, and ResUNet++.

Fig. 11 shows that HALSR-Net maintains high segmentation accuracy, precise boundary delineation, and minimal contour distance errors across varying imaging conditions, whether or not the ROI module is used. The model's architecture is inherently designed to adapt to different levels of image quality and anatomical variability, ensuring consistent performance even in challenging cases such as those found in the LVQuan19 dataset. This robustness can be attributed to its ability to extract and integrate multi-scale spatial features effectively, allowing it to focus on the most relevant structures despite differences in input quality.

In contrast, other models exhibited lower performance, struggling to generalize to the variability in the dataset. This suggests that these methods are more sensitive to changes in resolution and contrast, whereas HALSR-Net, through its hybrid attention and feature selection mechanisms, dynamically adapts to different imaging conditions, preserving segmentation accuracy across diverse clinical scenarios. The low variability observed in its results further reinforces its stability and efficiency, demonstrating its capacity to handle real-world challenges in medical image segmentation.

The visual segmentation outcomes in Fig. 12 reinforce the quantitative findings. The proposed method accurately delineates the LV and myocardium with minimal errors, effectively avoiding the inclusion of non-target elements. In contrast, other models faced difficulties in maintaining accuracy, particularly in complex regions, leading to segmentation inaccuracies. The results from the LVQuan19 dataset affirm the robustness and reliability of the proposed segmentation approach despite the dataset's low contrast and overall poor image quality, thus proving its applicability in various clinical scenarios.

4.8. Cross-Dataset Evaluation

To assess the generalizability of the proposed model, we trained it on one dataset and tested it on the other two. The results are summarized in Table 8 with performance evaluated using the three previous metrics.

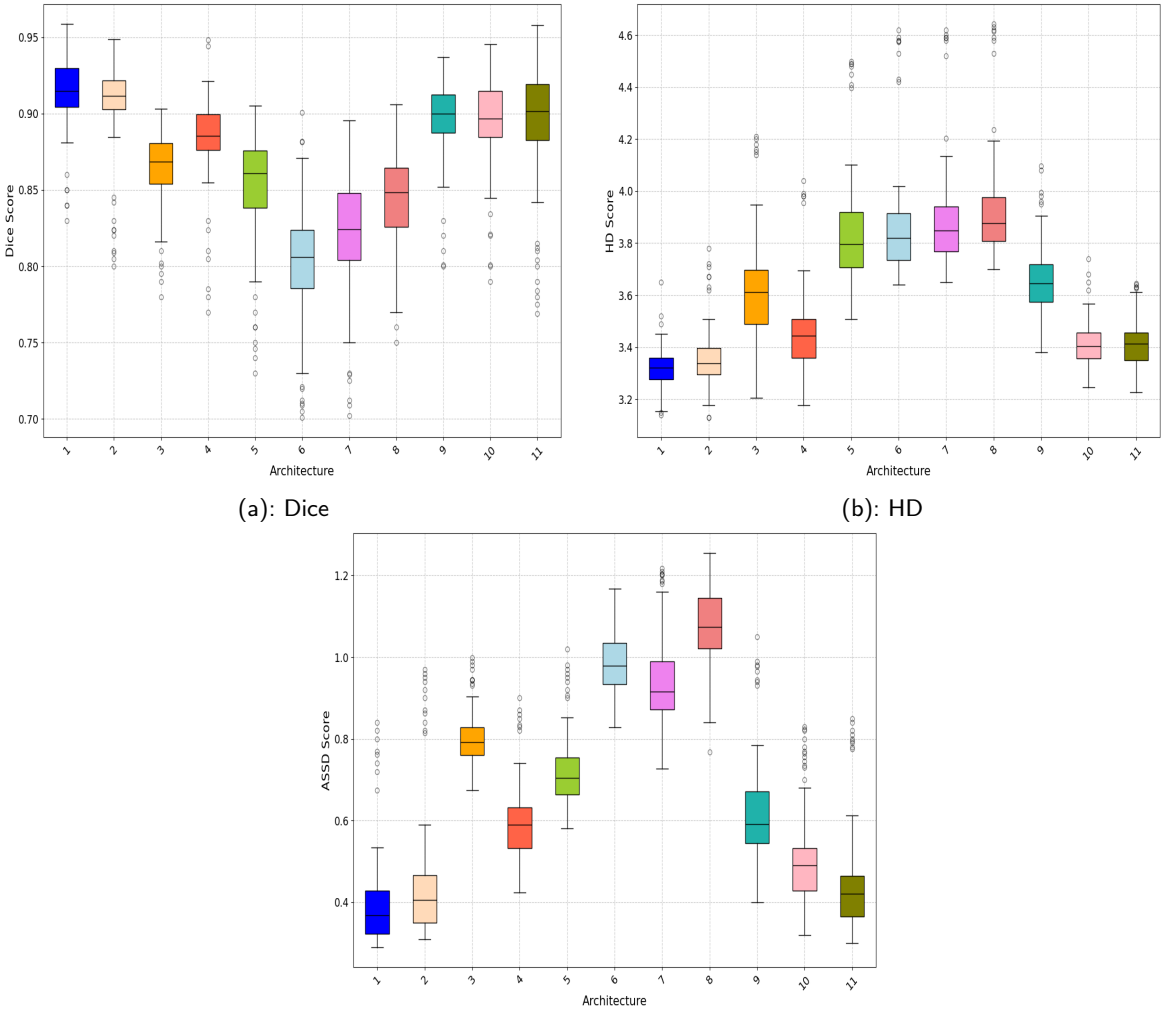
They demonstrate the stability and generalization capacity of HALSR-Net across different datasets. Despite inevitable variations due to domain shifts, the model maintains competitive performance when trained on one dataset and tested on another, with only moderate deviations in segmentation accuracy. This suggests that the learned features in HALSR-Net capture robust anatomical structures that generalize well across different imaging conditions.

Unlike many deep learning models that suffer from significant performance drops when applied to unseen data, HALSR-Net shows a consistent and stable performance across datasets. This is particularly relevant in medical imaging, where models are often required to adapt to different acquisition protocols, anatomical variations, and imaging

system features. The stability observed across datasets reinforces the effectiveness of the model's feature extraction strategy, which enables it to preserve segmentation accuracy even when trained on a different dataset.
Additionally, the results suggest that leveraging pre-trained weights from one dataset to another does not drastically degrade performance, making HALSR-Net a suitable candidate for scenarios with limited labeled data. This generalization ability reduces the need for extensive re-training and facilitates model adaptation to new datasets without requiring significant modifications. Overall, these findings highlight the model's robustness and its potential for clinical deployment across diverse imaging conditions.

4.9. Discussion

The experiments conducted on the in-house clinical dataset and two public datasets (ACDC and LVQuan19) showcase the enhanced performance of the proposed HALSR-Net architecture. Our model consistently outperformed other state-of-the-art models, including U-Net, attention gates, ResUNet++, nnU-Net and recent architectures like CFNet, DSGA-Net, and ARW-Net, by achieving better accuracy and lower loss values. Specifically, the model attained a training accuracy of up to 0.98 and maintained low training and validation losses values across all datasets, indicating



(a): Dice (b): HD

(c): ASSD

Figure 11: Boxplots of metrics for the LVQuan19 dataset: (a) Dice score, (b) HD (pixels), (c) ASSD (pixels). The methods using preliminary ROI detection are marked with (*): HALSR-Net (with ROI), 2. HALSR-Net (without ROI), 3. U-Net, 4. Att. gates*, 5. Non-local U-Net, 6. UNet++, 7. SegNet, 8. U-SegNet*, 9. ResUNet, 10. ResUNet++*, 11. nnU-Net.

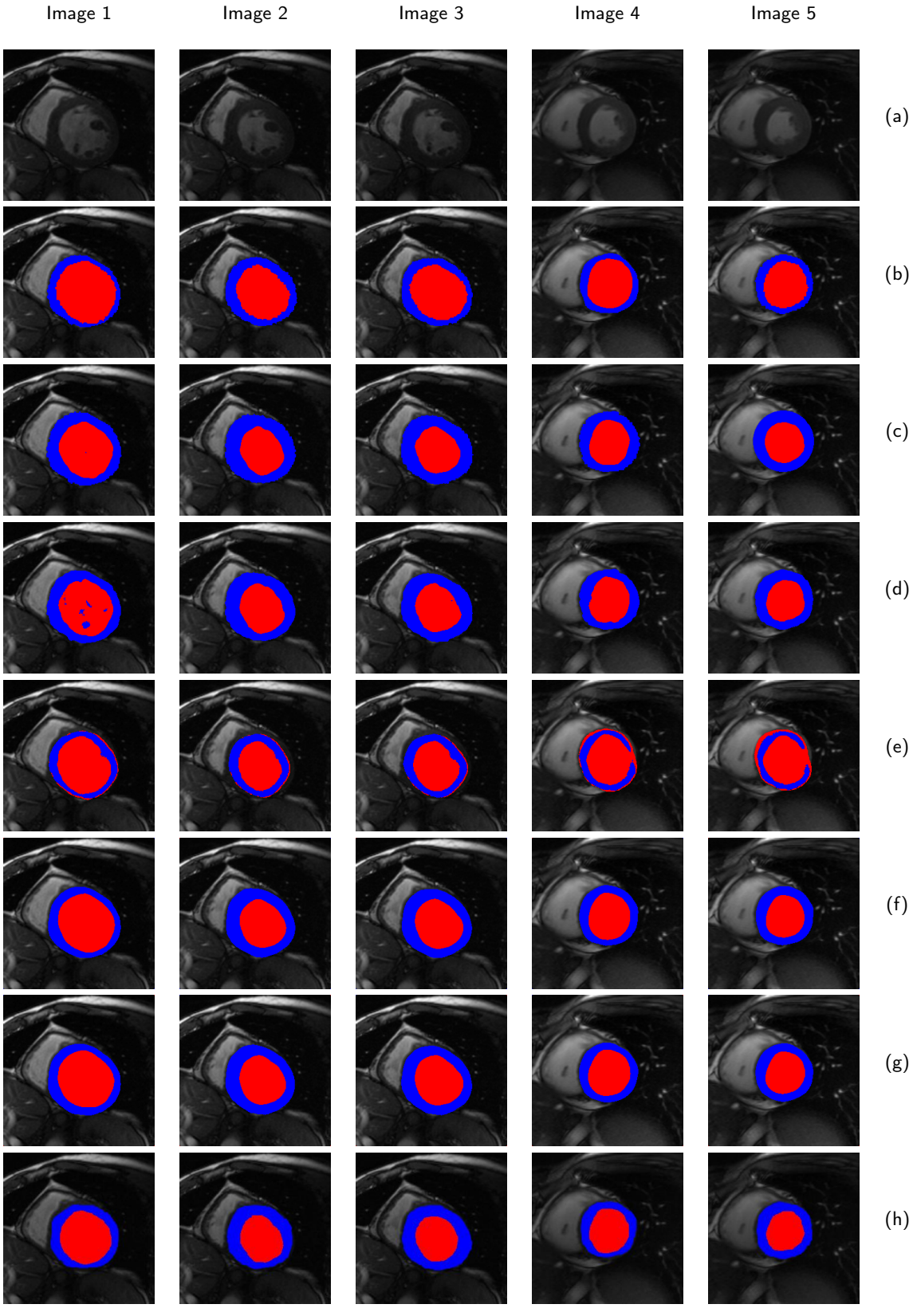


Figure 12: Segmentation of LV cavity (red) and myocardium (blue) : visual outcomes demonstrated for the LVQuan19 dataset Using the most performant competing architectures: (a) original image, (b) ground truth, (c) Proposed HALSR-Net, (d) ResUNet++, (e) segNet, (f) U-Net, (g) Attention Gates, (h) nnU-Net.

Table 9
Parameters (M, in millions) and Total Training Time by Architecture and Dataset.

Model	Parameters (M)	In-house (s/epoch)	ACDC (s/epoch)	LVQuan19 (s/epoch)
HALSR-Net	10.7	13.58	13.13	12.18
U-Net	8.8	12.00	11.48	10.42
UNet++	9.8	12.45	12.05	11.46
Attention Gates	11.3	14.38	14.04	13.11
Non-local U-Net	12.2	15.85	15.40	14.65
ResUNet++	9.4	13.22	12.72	11.88
nnU-Net	105.9	79.99	76.52	69.43

efficient learning and minimal prediction errors.

Analyzing the results from the boxplots of Dice Coefficient, Hausdorff Distance, and Average Symmetric Surface Distance metrics (Figs. 7, 9, 11), we observe that our architecture consistently achieved the highest median DC, the lowest median HD, and median ASSD across all datasets. These findings indicate that our model not only performs in segmentation accuracy but also maintains global stability and coherence of the results. The high median DC suggests a significant overlap between the predicted and expert segmentations, while the low median HD and ASSD values indicate minimal local boundary errors and precise global contour localization. This comprehensive performance analysis confirms that our model delivers reliable and accurate segmentation results across various datasets.

The model's accuracy in boundary delineation and minimal contour distance errors are highlighted by the lowest HD and ASSD values. The minimal variability in these metrics indicates that our proposed architecture is highly reliable, maintaining high accuracy and low error rates across samples. Compared to other state-of-the-art methods like Non-local nnU-Net, UNet++, and Attention Gates, our model consistently demonstrates superior performance with reduced variability, making it a robust choice for clinical applications requiring high precision and reliability.

A significant aspect of the performance is the reduced number of outliers compared to competing models. The boxplots in Figs. 7, and especially 9 and 11, show significantly fewer outliers for our approach, with around 9% of cases classified as outliers, compared to about 14% for competing models. This reduced number of outliers indicates more consistent performance and fewer instances of large errors, highlighting the model's stability and robustness. The presence of fewer outliers means that our architecture can handle a wider variety of cases more effectively, avoiding significant errors commonly observed in competing models. This is particularly important in clinical settings where reliable and consistent performance is essential for accurate diagnosis and treatment planning.

The superior performance of our model can be attributed to its design. The HAPM Module enhances feature extraction by combining attention and pooling mechanisms, capturing both broad and detailed contextual information. The reconstruction module ensures detailed and contextually rich feature representations, further improving segmentation accuracy. In contrast, models lacking such advanced mechanisms struggle with greater variability and higher median errors, indicating that complexity alone does not necessarily translate to better generalization or accuracy in segmentation tasks.

The added complexity in models with multiple attention mechanisms, such as Non-local U-Net, often leads to slower convergence during training and an increased risk of overfitting. This is due to the expanded parameter space introduced by these mechanisms, which complicates the gradient descent optimization process. The optimizer may struggle to navigate the high-dimensional parameter space efficiently, resulting in slower convergence rates and a higher likelihood of getting trapped in suboptimal regions Sun et al. (2019). Moreover, the increased model capacity can cause the model to memorize training data rather than generalizing well to unseen data.

In contrast, our model achieves a balance between complexity and efficiency. By incorporating the Hybrid Attention Pooling Module, it effectively captures both global and local contextual information without excessively increasing the number of parameters.

This balance is further highlighted in Table 9, which compares the number of parameters and training times per epoch for HALSR-Net and other state-of-the-art models across the three datasets (in-house, ACDC, LVQuan19).

The computational efficiency analysis reveals that HALSR-Net, with 10.7M parameters, maintains competitive training times compared to simpler models like U-Net and UNet++, while outperforming more complex architectures like Attention Gates and Non-local U-Net in terms of training speed. These results demonstrate that HALSR-Net strikes

a balance between advanced mechanisms, such as the hybrid attention and reconstruction modules, and computational efficiency, making it practical for real-world clinical applications.
This computational balance ensures that HALSR-Net remains efficient and avoids challenges often associated with overly complex architectures, such as slower convergence rates and higher memory consumption, compared with Non-local U-Net for instance. By optimizing the integration of attention mechanisms and reconstruction modules, HALSR-Net achieves a design that is both efficient and effective, allowing it to handle diverse datasets with reduced computational overhead.
Although nnU-Net has emerged as a reference architecture for medical image segmentation due to its strong adaptability and segmentation quality, its main drawback remains its high computational cost. The significantly longer training times of nnU-Net, compared to other evaluated architectures, may limit its applicability in scenarios requiring fast and lightweight solutions.

The streamlined design of HALSR-Net reduces the risk of overfitting and facilitates a more efficient optimization process, allowing the model to converge more quickly and reliably to an optimal solution. The reconstruction module further enhances the model's ability to generalize across different datasets, leading to more stable and robust performance without the drawbacks associated with overly complex attention mechanisms.

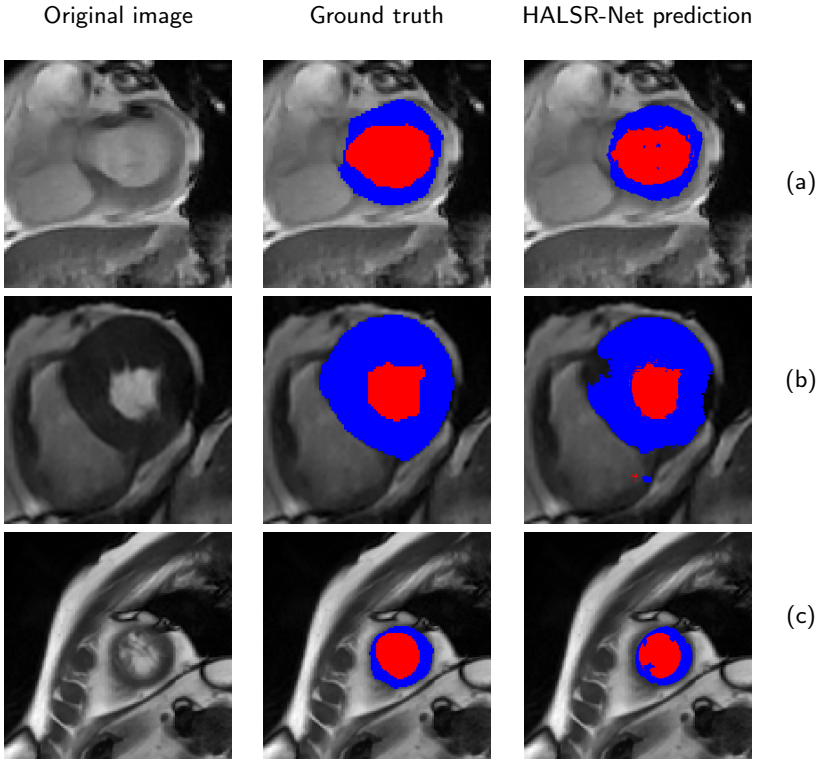


Figure 13: Examples of non optimal segmentations with our HALSR-Net architecture on in-house image samples: (a) basal level, (b) medial level and (c) apical level.

Despite its superior performance, our model still presents some challenges within the tested databases. Fig. 13 illustrates examples of erroneous segmentations from the in-house and ACDC datasets. The outlier rate across the entire dataset is approximately 8.35%, indicating cases where the model struggles to accurately delineate the cardiac structures. These errors, although less frequent and severe compared to competing models, highlight the difficulties in accurately capturing complex anatomical structures, particularly in challenging regions such as the apical and basal slice levels. These regions often present lower contrast and more variability in shape, making it harder for the model to consistently delineate the endocardium and epicardium.
Fig. 13 shows that discrepancies may occur at the junction with right ventricle for medial slices, at the pillars for

apical slices, and as shape deformations for basal slices. Despite these minor errors, the overall results remain largely unaffected, and the performance of our model still outperforms many competing architectures. In contrast, competing models often exhibit more severe errors, including completely empty segmentations in some cases, which severely affect the global accuracy of the results.
Accurate cardiac MRI segmentation is crucial for clinicians to diagnose and treat cardiac conditions effectively. Delineation of the endocardium and epicardium has a direct impact onto the measurements quality of cardiac volumes and ejection fractions, which are essential for assessing heart function. Our model's ability to provide reliable and consistent segmentation results can significantly enhance cardiac analysis and therefore patient care.

As a future direction, we aim to extend our method to 3D cardiac segmentation, which would allow for a more comprehensive analysis of the entire heart. This extension is crucial for capturing the full 3D anatomical shape of the heart, enabling the model to perform segmentation tasks more coherently across all regions. Additionally, this 3D extension would facilitate the extraction of valuable LV functional metrics, such as End-Diastolic Volume, End-Systolic Volume, and Ejection Fraction, which are essential for assessing cardiac function. These metrics are particularly important for evaluating heart health and are commonly used in clinical practice to monitor the progression of various heart diseases. However, generalization to 3D would not be immediate, as cine MRI data are not truly 3D due to the acquisition of each slice on independent apneas. As a result, the slices present geometric shifts that are sometimes significant, which would need to be jointly estimated to produce high-performance segmentation.
Moreover, integrating geometrical shape constraints during the segmentation process could improve the model's ability to delineate the endocardium and epicardium more accurately. These constraints would enforce realistic anatomical shapes and help the model avoid unrealistic segmentations, thus improving robustness.

5. Conclusion

This paper introduced HALSR-Net, a novel multi-level segmentation deep-learning architecture designed to enhance the efficiency and accuracy of cardiac cine-MRI segmentation of the left ventricular myocardium. The combination of a regression model for ROI extraction with HALSR-Net architecture for semantic segmentation effectively addressed the limitations of existing techniques. Key components, the HAPM and a sophisticated reconstruction module, significantly improved feature extraction and segmentation accuracy in the three studied databases.
On these databases, the experiments showed that the proposed method outperforms state-of-the-art architectures, achieving up to 98% Dice at testing for one of the datasets. The model also demonstrated robustness across the different datasets, maintaining high performance even with lower-quality images. In conclusion, this approach appears to provide a reliable and efficient solution for cardiac MRI segmentation, with the potential to further improve cardiac image analysis.

Code Availability Statement

The source code used in this study is publicly available at: GitHub repository. It provides all the necessary scripts and resources to reproduce the results presented in this work.

References

Abdeltawab, H., Khalifa, F., Taher, F., Alghamdi, N. S., Ghazal, M., Beache, G., Mohamed, T., Keynton, R., and El-Baz, A. (2020). A deep learning-based approach for automatic segmentation and quantification of the left ventricle from cardiac cine mr images. Computerized medical imaging and graphics, 81:pp.101717.

Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M., and Asari, V. K. (2018). Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. arXiv preprint arXiv:1802.06955, pages pp.1–12.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., and Farhan, L. (2021). Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. Journal of big Data, 8:pp.1–74.

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12):pp.2481–2495.

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Ballester, M. A. G., et al. (2018). Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging, 37(11):pp.2514–2525.

Cai, Y. and Wang, Y. (2022). Ma-unet: An improved version of unet based on multi-scale and attention mechanism for medical image segmentation. 12167:pp.205–211.

Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W., and Rueckert, D. (2020). Deep learning for cardiac image segmentation: a review. Frontiers in cardiovascular medicine, 7:pp.1–33.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4):pp.834–848.

Dangi, S., Linte, C. A., and Yaniv, Z. (2019). A distance map regularized cnn for cardiac cine mr image segmentation. Medical Physics, 46(12):pp.5637–5651.

D'Elia, E., Vaduganathan, M., Gori, M., Gavazzi, A., Butler, J., and Senni, M. (2015). Role of biomarkers in cardiac structure phenotyping in heart failure with preserved ejection fraction: critical appraisal and practical use. European journal of heart failure, 17(12):pp.1231–1239.

Du, G., Cao, X., Liang, J., Chen, X., and Zhan, Y. (2020). Medical image segmentation based on u-net: A review. Journal of Imaging Science & Technology, 64(2):pp.1–47.

Gao, H., Yuan, H., Wang, Z., and Ji, S. (2019). Pixel transposed convolutional networks. IEEE transactions on pattern analysis and machine intelligence, 42(5):pp.1218–1227.

Grover, V. P., Tognarelli, J. M., Crossey, M. M., Cox, I. J., Taylor-Robinson, S. D., and McPhail, M. J. (2015). Magnetic resonance imaging: principles and techniques: lessons for clinicians. Journal of clinical and experimental hepatology, 5(3):pp.246–255.

Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., Zhang, T., Gao, S., and Liu, J. (2019). Ce-net: Context encoder network for 2d medical image segmentation. IEEE Transactions on Medical Imaging, 38(10):pp.2281–2292.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 37(9):pp.1904–1916.

Hesamian, M. H., Jia, W., He, X., and Kennedy, P. (2019). Deep learning techniques for medical image segmentation: achievements and challenges. Journal of digital imaging, 32:pp.582–596.

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. pages pp.7132–7141.

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. (2021). nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods, 18(2):pp.203–211.

Ismail, T. F., Strugnell, W., Coletti, C., Božić-Iven, M., Weingaertner, S., Hammernik, K., Correia, T., and Kuestner, T. (2022). Cardiac MR: from theory to practice. Frontiers in cardiovascular medicine, 9:pp.826283.

Jaderberg, M., Simonyan, K., Zisserman, A., et al. (2015). Spatial transformer networks. Advances in neural information processing systems, 28:pp.1–9.

Jha, D., Smedsrud, P. H., Johansen, D., De Lange, T., Johansen, H. D., Halvorsen, P., and Riegler, M. A. (2021). A comprehensive study on colorectal polyp segmentation with resunet++, conditional random field and test-time augmentation. IEEE journal of biomedical and health informatics, 25(6):pp.2029–2040.

Kaul, C., Manandhar, S., and Pears, N. (2019). Focusnet: An attention-based fully convolutional network for medical image segmentation. pages pp.455–458.

Lei, T., Wang, R., Zhang, Y., Wan, Y., Liu, C., and Nandi, A. K. (2021). Defed-net: Deformable encoder-decoder network for liver and liver tumor segmentation. IEEE Transactions on Radiation and Plasma Medical Sciences, 6(1):pp.68–78.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., and Heng, P.-A. (2018). H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE transactions on medical imaging, 37(12):pp.2663–2674.

Liu, Y., Captur, G., Moon, J. C., Guo, S., Yang, X., Zhang, S., and Li, C. (2016). Distance regularized two level sets for segmentation of left and right ventricles from cine-mri. Magnetic resonance imaging, 34(5):pp.699–706.

Lundervold, A. S. and Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on mri. Zeitschrift für Medizinische Physik, 29(2):pp.102–127.

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. pages pp.565–571.

Moeskops, P., Wolterink, J. M., Van Der Velden, B. H., Gilhuijs, K. G., Leiner, T., Viergever, M. A., and Išgum, I. (2016). Deep learning for multi-task medical image segmentation in multiple modalities. pages pp.478–486.

Moreno Lopez, M. and Ventura, J. (2018). Dilated convolutions for brain tumor segmentation in mri scans. pages pp.253–262.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., et al. (2018). Attention u-net: Learning where to look for the pancreas. pages pp.1–10.

Peng, P., Lekadir, K., Gooya, A., Shao, L., Petersen, S. E., and Frangi, A. F. (2016). A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging. Magnetic Resonance Materials in Physics, Biology and Medicine, 29:pp.155–195.

Petitjean, C. and Dacher, J.-N. (2011). A review of segmentation methods in short axis cardiac mr images. Medical Image Analysis, 15(2):pp.169–184.

Queirós, S., Barbosa, D., Heyde, B., Morais, P., Vilaça, J. L., Friboulet, D., Bernard, O., and D'hooge, J. (2014). Fast automatic myocardial segmentation in 4d cine cmr datasets. Medical Image Analysis, 18(7):pp.1115–1131.

Qureshi, I., Yan, J., Abbas, Q., Shaheed, K., Riaz, A. B., Wahid, A., Khan, M. W. J., and Szczuko, P. (2023). Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. Information Fusion, 90:pp.316–352.

Ringenberg, J., Deo, M., Devabhaktuni, V., Berenfeld, O., Boyers, P., and Gold, J. (2014). Fast, accurate, and fully automatic segmentation of the right ventricle in short-axis cardiac mri. Computerized Medical Imaging and Graphics, 38(3):pp.190–201.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. pages pp.234–241.

Severino, P., Maestrini, V., Mariani, M. V., Birtolo, L. I., Scarpati, R., Mancone, M., and Fedele, F. (2020). Structural and myocardial dysfunction in heart failure beyond ejection fraction. Heart failure reviews, 25:pp.9–17.

Singh, K. R., Sharma, A., and Singh, G. K. (2023). Attention-guided residual w-net for supervised cardiac magnetic resonance imaging segmentation. Biomedical Signal Processing and Control, 86:pp.105177.

Sun, J., Zhao, J., Wu, X., Tang, C., Wang, S., and Zhang, Y. (2023). Dsga-net: Deeply separable gated transformer and attention strategy for medical image segmentation network. Journal of King Saud University-Computer and Information Sciences, 35(5):pp.101553.

Sun, S., Cao, Z., Zhu, H., and Zhao, J. (2019). A survey of optimization methods from a machine learning perspective. IEEE transactions on cybernetics, 50(8):pp.3668–3681.

Von Zuben, A., Perotti, L. E., and Viana, F. A. (2023). Anatomically-guided deep learning for left ventricle geometry generation with uncertainty quantification based on short-axis mr images. Engineering Applications of Artificial Intelligence, 121:pp.106012.

Wang, L., Pei, M., Codella, N. C., Kochar, M., Weinsaft, J. W., Li, J., Prince, M. R., Wang, Y., et al. (2015). Left ventricle: fully automated segmentation based on spatiotemporal continuity and myocardium information in cine cardiac magnetic resonance imaging (lv-fast). BioMed research international, 2015(1):pp.367583.

Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., and Nandi, A. K. (2022). Medical image segmentation using deep learning: A survey. IET image processing, 16(5):pp.1243–1267.

Wang, Z., Zou, N., Shen, D., and Ji, S. (2020). Non-local u-nets for biomedical image segmentation. 34(04):pp.6315–6322.

Xiao, X., Lian, S., Luo, Z., and Li, S. (2018). Weighted res-unet for high-quality retina vessel segmentation. pages pp.327–331.

Yang, L., Song, Q., Wang, Z., and Jiang, M. (2019). Parsing r-cnn for instance-level human analysis. pages pp.364–373.

Yang, L., Zhai, C., Liu, Y., and Yu, H. (2023). Cfha-net: A polyp segmentation method with cross-scale fusion strategy and hybrid attention. Computers in Biology and Medicine, 164:pp.107301.

Yu, H., Yang, L. T., Zhang, Q., Armstrong, D., and Deen, M. J. (2021). Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives. Neurocomputing, 444:pp.92–110.

Zhan, B., Song, E., Liu, H., Gong, Z., Ma, G., and Hung, C.-C. (2023). Cfnet: A medical image segmentation method using the multi-view attention mechanism and adaptive fusion strategy. Biomedical Signal Processing and Control, 79:pp.104112.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. pages pp.3–11.

Zhuang, X. (2013). Challenges and methodologies of fully automatic whole heart segmentation: a review. Journal of healthcare engineering, 4(3):pp.371–407.