Hybrid 3D U-Net and Attention Mechanisms for Whole Heart Segmentation from CT Images

Anusha Kotte
Department of CSE, Jawaharlal Nehru Technological University, Hyderabad, India
anusha.jntuh@gmail.com (corresponding author)

V. Kamakshi Prasad
Department of CSE, Jawaharlal Nehru Technological University, Hyderabad, India
kamakshiprasad@jntuh.ac.in

Received: 2 January 2025 | Revised: 27 January 2025 and 4 February 2025 | Accepted: 6 February 2025

 DOI: https://doi.org/10.48084/etasr.10115

ABSTRACT

Accurate delineation of heart structures from multimodal images is crucial for the treatment and investigation of different cardiovascular diseases. Automated whole-heart segmentation remains a challenging task due to its complex structure and imbalances in sample data. Convolutional Neural Networks (CNNs) are popular due to their efficiency in segmenting medical images. However, they often struggle to capture long-range dependencies and lack the precision needed for complex anatomic structures such as the heart. To overcome these limitations, this study presents a hybrid 3D U-Net framework that utilizes residual connections with attention mechanisms to improve feature learning and localization of cardiac structures. Residual connections stabilize training in deeper networks and attention blocks focus on relevant regions, refining segmentation quality. This network focuses on relevant regions and uses attention blocks to enhance quality. The proposed architecture was evaluated on 40 volumetric CT images of the Multi-Modality Whole Heart Segmentation (MM-WHS) dataset, achieving an average dice score of 85%. These results demonstrate the effectiveness and high accuracy of the proposed method for delineating cardiac substructures, offering potential clinical utility for automated cardiac analysis.

Keywords-deep learning; cardiac CT; attention mechanisms, whole-heart segmentation

I. INTRODUCTION

Cardiovascular disease has been recognized as a significant contributor to mortality worldwide in recent years. The Golden Burden of Disease statistics in 2021 reported that 19.4 million people died from various cardiovascular diseases such as rheumatic heart disorders, ischemic heart disease, and ischemic stroke [1]. In recent years advanced imaging techniques have advanced to enhance the diagnosis and examination of cardiovascular conditions, including Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Positron Emission Tomography (PET). Among these modalities, CT provides precise anatomical details on heart structures [2]. Delineating heart structures from CT scans is critical to measuring pathological and morphological changes. Manual segmentation of cardiac structures is challenging and requires meticulous study. In asymptomatic individuals, automatic segmentation of heart images can help prevent various heart diseases, such as cardiomyopathy, heart stroke, coronary artery disease, and others.

In the past ten years, advances in artificial intelligence within the realm of computer vision, especially through Convolutional Neural Networks (CNNs), have shown impressive levels of efficiency. CNNs have been employed extensively in different automated medical image segmentation tasks [3, 4]. In [3], CNNs were used in medical imaging, greatly improving the process. The U-Net design integrates skip connections within an encoder-decoder framework to facilitate the transmission of low-level spatial information. Alongside architectural advancements, the Dice Similarity Coefficient (DSC) loss improves the effectiveness of deep learning methods in addressing the data imbalance problem across background and foreground voxels [4]. In [5], Principal Component Analysis (PCA) and feature selection techniques were integrated to improve coronary heart disease prediction by addressing high-dimensional data challenges.

Various deep-learning methods utilizing CNNs have recently been proposed for cardiac imaging [6]. Researchers have developed a 2.5D multislice network [7, 8] aimed at cardiac ventricle segmentation. However, 2.5D methods, which utilize numerous 2D sliced images, do not fully exploit the inherent 3D spatial characteristics. Researchers have explored automatic whole-heart segmentation approaches using standard segmentation networks such as FCN and U-Net. In [9], a pyramid local attention module was proposed to enhance the features by gathering pertinent information from compact and limited surrounding regions. In [10], a Fine-grained Calibrated Double-Attention Convolutional Network (FCDA-Net) was

proposed to effectively distinguish between the endocardium and the epicardium in ventricular MRIs. The FCDA-Net was built on top of the U-Net. It has an encoder-decoder framework and a dual grouped-attention module consisting of a fine calibration Spatial Attention Module (fcSAM) and a fine calibration Channel Attention Module (fcCAM). In [11], a stack attention technique was introduced to evaluate adjacent image slices to extract pertinent characteristics, using an intended image as a reference. To address the semantic segmentation difficulty, a novel Stack Attention U-Net (SAUN) was proposed and trained, which combined stack attention with the classic U-Net design. The cross-entropy and Dice coefficients were combined as a loss function to train SAUN. In [12], the Multi-Attention Efficient Feature Fusion Network (MAEF-Net) was used for automatic left ventricle segmentation. Afterward, the Left Ventricular Ejection Fraction (LVEF) was calculated by automatically identifying the End-Diastolic Frames (EDFs) and End-Systolic Frames (ESFs) overall cardiac cycles. The MAEF-Net technique includes a multi-attention mechanism to efficiently collect heartbeat properties while decreasing noise. It also uses spatial pyramid feature fusion and deep supervision to increase the effectiveness of feature extraction.

In [13], a new deep network with a U-shaped design was proposed for accurate skin lesion segmentation. This network uses two small attention modules, Adaptive Channel-Context-Aware Pyramid Attention (ACCAPA) and Global Feature Fusion (GFF). Utilizing channel information, contextual cues, and global structural data in real-time, the ACCAPA module dynamically learns and predicts lesion area features. In [14], the RIANet efficiently reused parameters to encode intricate representative characteristics by integrating a recurrent feedback framework known as the clique block, which merges forward and backward connections across layers of uniform resolution. A plug-and-play Interleaved Attention (IA) block manages data flow into the decoding phase by efficiently integrating multilevel contextual information. In [15], a novel deep network was used to effectively capture extensive features through a channel attention mechanism, incorporating a Dual-Path Feature Extraction Module (DP-FEM). A High- and Low-Level Feature Fusion Module (HL-FFM) was developed, and spatial attention was used to selectively combine low-level spatial details with high-level semantic information from features. The coarse segmentation model in [16] used the left ventricle myocardial structure as a shape prior. The fine segmentation model used a pixel-wise attention mechanism and an auto-weighted supervision model to find and pull out important pathological structures from multi-sequence CMR data. In [17], the CAB module was incorporated into the 3D U-Net framework, resulting in a new model known as CAB U-Net that improved gradient flow within the network and effectively leveraged low-resolution feature information.

Different state-of-the-art methods have been proposed for medical image segmentation specifically on cardiac structures. These approaches incorporated the classic U-Net architecture along with attention mechanisms to enhance feature extraction and segmentation performance. Table I presents a brief summary of these works.

TABLE I. SUMMARY OF CURRENT TECHNIQUES IN MEDICAL IMAGE SEGMENTATION

Study	Techniques used	Strengths	Limitations
Pyramid Local Attention [9]	A pyramid local attention module enhances feature extraction by collecting pertinent information from compact regions.	Improves performance in compact regions.	Limited by the focus on small regions, might not generalize well for large or complex structures such as the heart.
FCDA-Net [10]	Fine-grained calibrated double-attention convolutional network combining spatial and channel attention.	Effectively distinguishes between endocardium and epicardium in MRI.	Primarily focused on MRI data, may not generalize to other imaging modalities such as CTs.
Stack Attention U-Net (SAUN) [11]	Combines stack attention with U-Net to extract characteristics from adjacent image slices.	Effective for extracting slice-wise information in sequential images.	May struggle with handling non-sequential structures, such as the heart's 3D nature.
MAEF-Net [12]	Multi-Attention Efficient Feature Fusion Network that combines spatial pyramid feature fusion and deep supervision.	Multi-Attention Efficient Feature Fusion Network that combines spatial pyramid feature fusion and deep supervision.	Focused on the left ventricle, limiting applicability to whole heart segmentation tasks.
ACCAPA and GFF [13]	Adaptive channel-context-aware pyramid attention module and global feature fusion for skin lesion segmentation.	Dynamically learns and predicts lesion area features in real-time.	Not designed for 3D volumetric medical images, which are essential for whole-heart segmentation.
RIANet [14]	Utilizes recurrent feedback mechanisms and interleaved attention blocks.	Efficient in encoding complex features with dynamic feedback.	Computationally expensive and not optimized for large volumetric medical images such as cardiac CT.
Dual-Path Feature Extraction (DP-FEM) [15]	Dual-path feature extraction using channel attention mechanisms.	Effective at capturing broad features through attention.	High computational cost and may not adequately handle class imbalances in complex cardiac segmentation.
Coarse-to-Fine Segmentation [16]	Uses coarse segmentation with shape priors and fine segmentation with pixel-wise attention mechanisms.	Provides good accuracy in segmenting key cardiac structures.	Not optimized for handling class imbalances in large datasets.
CAB U-Net [17]	Incorporates the CAB module into 3D U-Net, enhancing gradient flow and low-resolution feature usage.	Good for gradient flow and multi-sequence segmentation tasks.	May still struggle with complex structures and large volumetric data.

These methods have several limitations in terms of generalization and performance in clinical settings. Techniques such as U-Net and its variants fail to segment complex structures, such as the heart, and they are computationally expensive. To address challenges in whole heart segmentation, this study proposes a hybrid 3D U-Net framework that consists of residual connections and attention mechanisms. Residual blocks are used to address the vanishing gradient problem and

perform efficient feature learning. Attention mechanisms are used to highlight the relevant regions and suppress irrelevant features. This framework is designed to handle complex structures within the heart and class imbalances in cardiac segmentation tasks. The main contributions of this study are:

- Develops a hybrid 3D U-Net framework using residual connections and attention mechanisms for efficient feature learning and accurate segmentation of cardiac structures.
- Evaluates comprehensively the proposed method on the MM-WHS dataset, demonstrating a good performance with an average Dice score of 85%, outperforming baseline methods.
- Uses advanced data augmentation techniques to address data variability and class imbalances.

II. METHODOLOGY

This research utilized a hybrid 3D U-Net model to improve the accuracy in identifying the heart structures from CT images. The results observed on public datasets, including MM-WHs, were favorable. Figure 1 illustrates the proposed architecture for a hybrid 3D U-Net model incorporating attention mechanisms. The proposed hybrid U-Net model is designed for tasks that require precise segmentation or localization within 3D images, such as medical image analysis. The proposed hybrid 3D U-net model uses residual learning, attention mechanisms, and normalization techniques. This architecture enhances the feature extraction and segmentation accuracy.

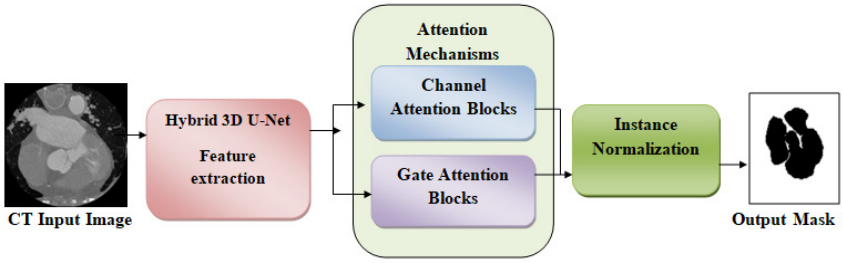

Fig. 1. Overall architecture of the proposed hybrid 3D U-Net framework.

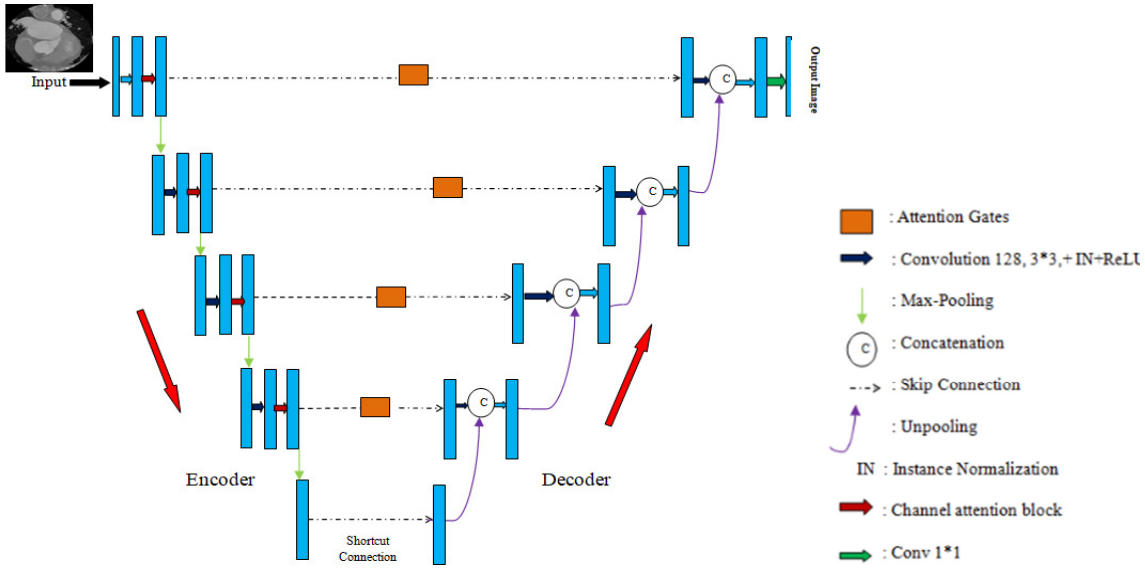

Fig. 2. 3D hybrid U-Net architecture.

A. 3D Hybrid U-Net Model

This model, a variation of the U-Net design, seeks to precisely segregate cardiac components. The model accepts a 3D volume with dimensions of 128×128×128×1. The downsampling comprises a sequence of residual blocks using progressively larger filter sizes (16, 32, 64, 128), followed by max pooling with a pool size of 2×2×2. The final downsampling block employs a residual block with 256 filters to extract the most abstract information. This approach reduces spatial dimensions while increasing feature complexity. The upsampling path reconstructs spatial dimensions by incorporating attention mechanisms and concatenating features from the corresponding downsampling blocks. The system continuously applies residual learning to enhance segmentation. Figure 2 illustrates the framework of the 3D hybrid U-Net model.

B. Attention Mechanisms

Attention mechanisms improve model performance by allowing the network to focus on relevant regions and features. The proposed model combines two types of attention mechanisms, channel attention and attention gate mechanisms, to emphasize important features and refine skip connections, which play a critical role in enhancing it overall performance.

1) Channel Attention Blocks

Channel-based attention mechanisms in CNNs generate channel attention maps using inter-channel relationships among features. Each feature map functions as a feature detector, focusing on the significant aspects of the input image. A single channel attention feature is produced by combining the outputs from the max pooling and average pooling path. Figure 3 shows details on the channel attention block.

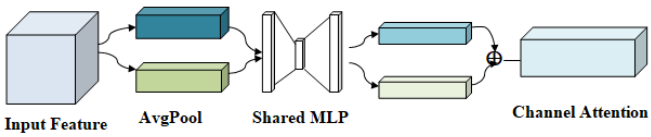

Fig. 3. Channel attention block diagram.

- Assume that the input feature map is $X \in R^{H\times W\times D\times C}$ where H, W, and D denote the spatial dimensions, specifically height, width, and depth, while C denotes the total number of channels.
- These input feature maps represent anatomical structures of the heart. Global average pooling averages the feature values along the spatial dimensions, resulting in a tensor shape of $1\times1\times1\times C$. Max pooling takes the maximum value of the feature map across the spatial dimensions and yields the shape $1\times1\times1\times C$. Two aggregated feature vectors are produced as:

$$AvgPool\ (X) = \frac{1}{H\times W\times D}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{D} X\ (i,j,k,:) \quad (1)$$

$$MaxPool\ (X) = max_{i,j,k}\ X(i,j,k,:) \quad (2)$$

- After pooling, $AvgPool(X)$ and $MaxPool(X)$ are passed through two parallel Conv3D operations to generate the attention maps. The dimensionality is first reduced by a factor of r (e.g., $r = 8$) and then expanded back:

$$Avg_{Out} = \sigma(W_2 \cdot ReLU(W_1 \cdot AvgPool(X))) \quad (3)$$

$$Max_{Out} = \sigma(W_4 \cdot ReLU(W_3 \cdot MaxPool(X))) \quad (4)$$

where W_1 and $W_3 \in R^{1^x1^x1^x1^xc\frac{C}{r}}$ are weights for the first convolutional layer and W_2 and $W_4 \in R^{1^x1^x1^x\frac{C_x}{r}C}$ are weights for the second convolutional layer. σ denotes the sigmoid activation function. Finally, the pooled features are refined by learning non-linear interchannel dependencies. This approach focuses on more relevant features specific to heart structures and avoids less important information. To create a single-channel attention map, the two outputs are joined using element-wise addition:

$$ChannelAttention\ (X) = Avg_{Out} + Max_{Out} \quad (5)$$

- This attention map highlights the important channels in the feature map by combining global average and max pooling paths. For segmentation, this helps the model to focus more effectively on relevant structures of the heart by emphasizing the features. Finally, this attention map is multiplied with the original input:

$$X_{Out} = X * ChannelAttention(X) \quad (6)$$

- The obtained feature map X_{Out} is attention enhanced, optimized to focus on meaningful channels for whole heart segmentation, and then it is passed to subsequent layers of the segmentation network (U-Net) to predict the accurate segmentation masks for different heart structures.

$$X_{Out} = X \times (\sigma(W_2 \cdot ReLU(W_1 \cdot AvgPool(X))) + \sigma(W_4 \cdot ReLU(W_3.\ MaxPool(X)))) \quad (7)$$

2) Attention Gate Mechanism

Attention gate mechanisms are useful to suppress irrelevant areas and highlight important regions in the feature maps. This improves the integration of high- and low-level features. This attention gate mechanism works by comparing two inputs: the skip connection feature map and a gating signal. The skip connection feature map contains high-resolution spatial information on the heart structures, denoted as $X \in R^{H\times W\times D\times C_x}$, where H, W, and D represent the spatial dimensions, and C_x is the number of channels. Another feature map i.e., gating signal from a deeper layer represents the global context of the whole heart image, denoted as $G \in R^{H_g\times W_g\times C_g}$ where H_g and W_g represent spatial dimensions and C_g represents the count for the gating signal. This attention gate helps to focus on more meaningful features by comparing the local features in X with the global context in G. Figure 4 shows a detailed block diagram of the attention gate.

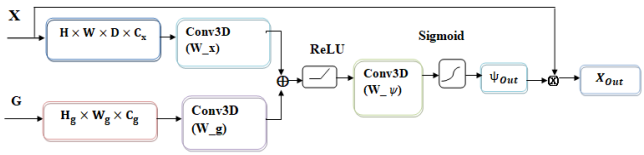

Fig. 4. Attention gate block diagram.

First, both X and G pass through separate Conv3D layers to transform their channels. The operations are as follows:

$$g_1 = W_g.G \quad (8)$$

$$x_1 = W_x.X \quad (9)$$

where $W_g \in R^{1\times1\times1\times C_g}$ is a Conv3D layer applied to the gating signal G, and $W_x \in R^{1\times1\times1\times C_x\times C_x}$ is applied to the input feature map X. This allows the network to extract meaningful features and maintain consistent dimensions. Next, the two transformed features are combined using element-wise addition, followed by a ReLU activation. So g_1 and x_1 fuse local information generated from the skip connection with the global context, which is generated from the gating signal. The relevant features

for segmentation are highlighted by applying the ReLU activation function.

$$\psi = ReLU(g_1 + x_1) \quad (10)$$

A single-channel attention map is created using another Conv3D layer, followed by a sigmoid activation to constrain the values between 0 and 1.

$$\psi_{Out} = \sigma(W_{\psi}.\psi) \quad (11)$$

where $W_{\psi} \in R^{1\times1\times1\times c_x\times1}$ is the weight of the convolutional layer. ψ_{Out} represents the attention map that highlights the regions of the feature map and suppresses the irrelevant background and noise. It modulates the feature map by assigning scores to each voxel. The resulting attention map ψ_{Out} is multiplied element-wise with the original input feature map X.

$$X_{Out} = X * \psi_{Out} \quad (12)$$

This attention mechanism produces the resulting feature map X_{Out} that retains the spatial information from the actual feature map. This model identifies the most critical heart structures, such as the left ventricle, myocardium, and others, during the segmentation process and improves the accuracy in delineating the boundaries and structures within the heart. The following algorithm describes how the proposed hybrid 3D U-Net segments cardiac structures.

```
Algorithm: Hybrid 3D U-Net with attention mechanisms
Input: MM-WHS Dataset.
Output: Delineation of cardiac structures
1. Consider 3D CT images of the heart and
   normalize the input images
2. Perform data augmentation (random
   rotations, flips, scaling, and
   translations) to increase diversity.
3. Apply 3D convolutions with residual
   connections for feature extraction.
4. Attention Mechanism (Channel
   Attention):
   For input feature map X (size
   H × W × D × C):
     Apply global average pooling and max
     pooling
     Generate the channel attention map
     Apply Conv3D with weights W1, W2 to
     generate attention maps
     Combine the outputs from AvgPool and
     MaxPool
5. Apply attention gate mechanism
   Refine features by multiplying the
   attention map with the input feature
   map
6. For each level in upsampling,
     apply deconvolution (to upsample the
     features
7. Concatenate features from
   corresponding encoder block
8. Continue until the final segmentation
   mask is generated with Conv3D
9. Compute evaluation metrics
10. Final segmentation results and
    performance metrics
```

C. Residual Blocks

In deep learning architectures, residual block connections are important to address complex tasks such as image segmentation. In the proposed model, each residual block has two convolutional layers, each using 3×3×3 kernels. These layers are designed to extract fine-grained features to delineate intricate cardiac structures such as ventricles, atria, and myocardium. Computational complexity is manageable by using these layers as they learn from fine-grained features. Instance normalization is applied to stabilize the learning process and to normalize the feature maps. After each convolution layer, a ReLU is used to enable the network to learn complex patterns associated with the anatomical structures of the heart. To ensure compatibility with the output of the two convolution layers, a 1×1×1 convolution is used to modify the dimensions of the input features. This alignment allows the addition of a shortcut connection, which establishes a direct path for gradients to propagate through the network. This shortcut connection effectively mitigates the problem of vanishing gradients by enabling the training of deeper networks. This ability is helpful for whole-heart segmentation, where the model accurately differentiates structures that vary significantly in size and shape. Residual connections enhance the model's ability to achieve accurate and robust segmentation of the heart, even in the presence of challenges such as class imbalances and anatomical variability.

III. EXPERIMENTAL RESULTS

This study used the MM-WHS dataset [18-21], which consists of 20 labeled and 40 unlabeled cardiac CT images. In the WHS analysis, seven substructures are important: 1) Left Ventricle (LV) cavity, 2) The Right Ventricular chamber (RV), 3) The Left Atrial chamber (LA) 4) The chamber of the Right Atrium (RA), 5) The myocardial tissue of the left ventricle (LV-myo), 6) The trunk of the ascending aorta (AA), 7) The trunk of the Pulmonary Artery (PA).

A. Implementation Details

The model was trained on these volumetric medical images, where each volume is associated with its corresponding segmentation mask. The input to the model is a 3D CT image (size: depth, height, width, and channels) and the output is a multiclass segmentation mask with one hot encoded labels for different structures of the heart. To mitigate the constraint of GPU memory, the image was scaled to 128×128×128 voxels for model input. Different data augmentation techniques, including random flipping (horizontal and vertical), random rotations in increments of 90°, and adjustments to brightness and contrast, were applied to both input images and segmentation masks.

The proposed model was trained with the cross-entropy loss, which is suitable for multiclass segmentation problems.

Class weights were also calculated and applied to handle class imbalances in the dataset. Ten-fold cross-validation was performed, wherein each fold, data allocation involved 90% assigned to the training set and the other 10% assigned to the validation set. This entire process was executed ten times, each iteration using a distinct subset designated as a validation set, ensuring that all data are used for both training and validation purposes. The proposed model was implemented using Tensorflow and Keras.

TABLE II. COMPARISON OF DICE COEFFICIENT IN %

Model	Comparison of Dice Coefficient in %							
	MLV	LA	LV	RA	RV	AA	PUA	Avg. DSC
Two-Stage U-net [22]	0.756	0.821	0.865	0.667	0.771	0.756	0.723	0.793
Our Model	0.853117	0.876	0.849	0.884	0.861	0.834	0.805	0.858

The performance of the proposed hybrid U-Net model for whole heart segmentation was evaluated and compared with the Two-Stage U-Net model [22]. The proposed model demonstrated a significant improvement over the Two-Stage U-Net across most metrics, particularly in average DSC, where it achieved an average of 0.858 compared to 0.793 for the Two-Stage U-Net. The performance metrics considered included the Dice Similarity Coefficient (DSC) [23] for multiple cardiac structures, including the LV, LA, RA, RV, and AA, along with the overall average DSC. These findings indicate that the proposed method is a promising advance in the field of cardiac segmentation, meriting further exploration and validation in larger datasets and varied clinical scenarios. DSC was calculated using:

$$DSC = \frac{2|A \cap B|}{|A|+|B|} \quad (1)$$

All experiments were carried out using an NVIDIA Quadro RTX 5000 GPU with driver version 552.86 and CUDA version 12.4. The initial learning rate was 0.0001 with batch size 2. A total of 50 training epochs and the Adam optimizer addressed the gradient descent problem. Figures 5 and 6 show the training and validation loss and accuracy.

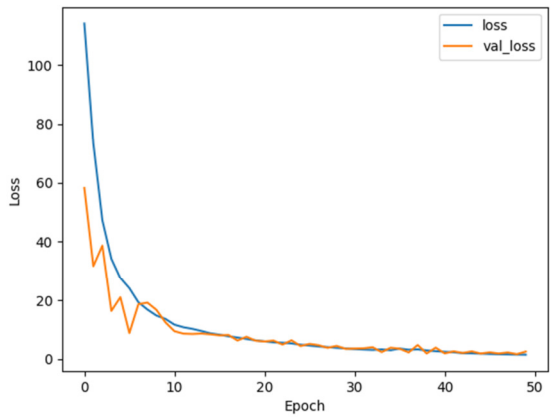

Fig. 5. Training and validation loss.

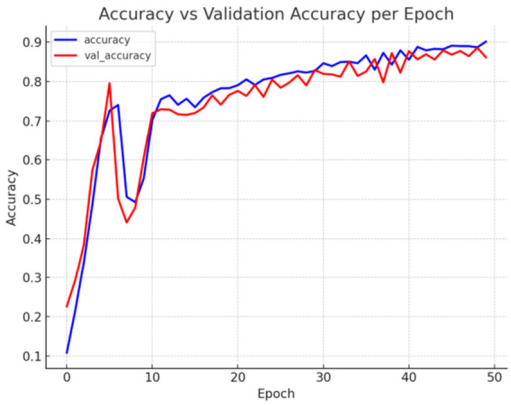

Fig. 6. Training and validation accuracy.

Figures 5 and 6 provide insights into model performance and convergence. The minimum gaps between the training and validation curves in both plots indicate that the model is well-regularized and achieved good performance on unseen data with limited overfitting.

IV. CONCLUSION

This study presented a hybrid 3D U-Net architecture for the challenging task of whole heart segmentation from cardiac CT images. The proposed architecture leverages advanced techniques, such as residual learning and channel attention mechanisms, to create channel attention features. Residual learning ensures that the model effectively captures complex patterns using the vanishing gradient problem. Attention gates are incorporated to further enhance the model's ability to avoid irrelevant regions of the images and to identify the relevant regions in feature maps. These gates selectively highlight the salient features for segmentation, refine the focus of the network, and improve accuracy. The proposed model achieved an average dice score of 85% when combined with the data augmentation techniques, showcasing its robustness in handling diverse variations in the dataset. It also demonstrated a significant improvement in segmenting critical structures within the heart, such as the left ventricle (LV) and the left atrium (LA), achieving dice scores of 88.4% and 87.6%, respectively, compared to 86.5% and 82.1% in [22]. The integration of attention-gate mechanisms played a pivotal role in refining the fusion of multiscale features, allowing for a more precise segmentation of intricate cardiac structures and improved segmentation accuracy. This advancement has significant potential for clinical applications, as precise segmentation is crucial for diagnosis and decision-making for further treatment in cardiology.

Future work will focus on extending the proposed method to multi-modality datasets, allowing the integration of complementary information from different imaging modalities such as MRI and echocardiography. Additionally, real-time applications in clinical settings will be explored, along with addressing segmentation challenges in imbalanced datasets. This could involve integrating advanced domain adaptation and uncertainty estimation techniques to generalize the model and enhance its reliability in clinical settings.

REFERENCES

[1] "Cardiovascular diseases - Level 2 cause | Institute for Health Metrics and Evaluation." https://www.healthdata.org/research-analysis/diseases-injuries-risks/factsheets/2021-cardiovascular-diseases-level-2-disease.

[2] S. Park and M. Chung, "Cardiac segmentation on CT Images through shape-aware contour attentions," *Computers in Biology and Medicine*, vol. 147, Aug. 2022, Art. no. 105782, https://doi.org/10.1016/j.compbiomed.2022.105782.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015, pp. 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.

[4] F. Milletari, N. Navab, and S. A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, Stanford, CA, USA, Oct. 2016, pp. 565–571, https://doi.org/10.1109/3DV.2016.79.

[5] M. J. J. Ghrabat *et al.*, "Utilizing Machine Learning for the Early Detection of Coronary Heart Disease," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 17363–17375, Oct. 2024, https://doi.org/10.48084/etasr.8171.

[6] E. Gibson *et al.*, "Automatic Multi-Organ Segmentation on Abdominal CT With Dense V-Networks," *IEEE Transactions on Medical Imaging*, vol. 37, no. 8, pp. 1822–1834, Dec. 2018, https://doi.org/10.1109/TMI.2018.2806309.

[7] O. Oktay *et al.*, "Attention U-Net: Learning Where to Look for the Pancreas." arXiv, May 20, 2018, https://doi.org/10.48550/arXiv.1804.03999.

[8] M. Chung, J. Lee, S. Park, C. E. Lee, J. Lee, and Y. G. Shin, "Liver segmentation in abdominal CT images via auto-context neural network and self-supervised contour attention," *Artificial Intelligence in Medicine*, vol. 113, Mar. 2021, Art. no. 102023, https://doi.org/10.1016/j.artmed.2021.102023.

[9] F. Liu, K. Wang, D. Liu, X. Yang, and J. Tian, "Deep pyramid local attention neural network for cardiac structure segmentation in two-dimensional echocardiography," *Medical Image Analysis*, vol. 67, Jan. 2021, Art. no. 101873, https://doi.org/10.1016/j.media.2020.101873.

[10] F. Liu, K. Wang, D. Liu, X. Yang, and J. Tian, "Deep pyramid local attention neural network for cardiac structure segmentation in two-dimensional echocardiography," *Medical Image Analysis*, vol. 67, Jan. 2021, Art. no. 101873, https://doi.org/10.1016/j.media.2020.101873.

[11] X. Sun, P. Garg, S. Plein, and R. J. van der Geest, "SAUN: Stack attention U-Net for left ventricle segmentation from cardiac cine magnetic resonance imaging," *Medical Physics*, vol. 48, no. 4, pp. 1750–1763, 2021, https://doi.org/10.1002/mp.14752.

[12] Y. Zeng *et al.*, "MAEF-Net: Multi-attention efficient feature fusion network for left ventricular segmentation and quantitative analysis in two-dimensional echocardiography," *Ultrasonics*, vol. 127, Jan. 2023, Art. no. 106855, https://doi.org/10.1016/j.ultras.2022.106855.

[13] W. Zhang, F. Lu, W. Zhao, Y. Hu, H. Su, and M. Yuan, "ACCPG-Net: A skin lesion segmentation network with Adaptive Channel-Context-Aware Pyramid Attention and Global Feature Fusion," *Computers in Biology and Medicine*, vol. 154, Mar. 2023, Art. no. 106580, https://doi.org/10.1016/j.compbiomed.2023.106580.

[14] Q. Tong *et al.*, "RIANet: Recurrent interleaved attention network for cardiac MRI segmentation," *Computers in Biology and Medicine*, vol. 109, pp. 290–302, Jun. 2019, https://doi.org/10.1016/j.compbiomed.2019.04.042.

[15] L. Guo *et al.*, "Dual attention enhancement feature fusion network for segmentation and quantitative analysis of paediatric echocardiography," *Medical Image Analysis*, vol. 71, Jul. 2021, Art. no. 102042, https://doi.org/10.1016/j.media.2021.102042.

[16] K. N. Wang *et al.*, "AWSnet: An auto-weighted supervision attention network for myocardial scar and edema segmentation in multi-sequence cardiac magnetic resonance images," *Medical Image Analysis*, vol. 77, Apr. 2022, Art. no. 102362, https://doi.org/10.1016/j.media.2022.102362.

[17] X. Ding, Y. Peng, C. Shen, and T. Zeng, "CAB U-Net: An end-to-end category attention boosting algorithm for segmentation," *Computerized Medical Imaging and Graphics*, vol. 84, Sep. 2020, Art. no. 101764, https://doi.org/10.1016/j.compmedimag.2020.101764.

[18] S. Gao, H. Zhou, Y. Gao, and X. Zhuang, "BayeSeg: Bayesian modeling for medical image segmentation with interpretable generalizability," *Medical Image Analysis*, vol. 89, Oct. 2023, Art. no. 102889, https://doi.org/10.1016/j.media.2023.102889.

[19] X. Zhuang, "Multivariate Mixture Model for Myocardial Segmentation Combining Multi-Source Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 12, pp. 2933–2946, Sep. 2019, https://doi.org/10.1109/TPAMI.2018.2869576.

[20] X. Luo and X. Zhuang, "*X*-Metric: An N-Dimensional Information-Theoretic Framework for Groupwise Registration and Deep Combined Computing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 9206–9224, Jul. 2023, https://doi.org/10.1109/TPAMI.2022.3225418.

[21] F. Wu and X. Zhuang, "Minimizing Estimated Risks on Unlabeled Data: A New Formulation for Semi-Supervised Medical Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 6021–6036, Feb. 2023, https://doi.org/10.1109/TPAMI.2022.3215186.

[22] T. Liu, Y. Tian, S. Zhao, X. Huang, and Q. Wang, "Automatic Whole Heart Segmentation Using a Two-Stage U-Net Framework and an Adaptive Threshold Window," *IEEE Access*, vol. 7, pp. 83628–83636, 2019, https://doi.org/10.1109/ACCESS.2019.2923318.

[23] L. R. Dice, "Measures of the Amount of Ecologic Association Between Species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945, https://doi.org/10.2307/1932409.