University of Science and Technology of Ha Noi

Machine Learning in Medicine

Final Project Report

Multiview Cardiac Structure Segmentation

Student Name	Student ID
Le Linh Long	22BI13262
Nguyen Tuan Khai	22BI13202
Nguyen Quang Huy	22BI13195
Pham Thai Son	22BI13397
Nguyen Hai Dang	22BI13073

Lecturer:

Assoc. Prof. Tran Giang Son

Submission Date: 31/03/2025

1 Introduction

In the field of medicine, one of the most powerful and highly utilized tools is magnetic resonance imaging (MRI) which offers high spatiotemporal for the assessment of various anatomical structures of the body, allowing doctors and medical practitioners to gain critical information for diagnostics and therapeutic applications. However, the accurate segmentation and identification of these structures for downstream application is often very tedious and labor-intensive, and subject to inter-observer variability. This aspect therefore has prompted many researchers to look for automated approaches to this task, leveraging deep learning architectures such as U-Net, to great effects. This solution has demonstrated substantial promise in alleviating these limitations.

With recent advances in the field, we have been able to understand more thoroughly the complementary strengths of 2D and 3D architectures. While 2D U-Nets benefit from higher in-plane resolution and reduced computational cost making it viable for smaller computing systems, 3D U-Nets can capture critical information for understanding structural coherence and changes across these slices, such as the volumetric context between them. This led many researchers to theorize that the potential harmonization between these two methods' predictions can lead to a unified segmentation method that can outperform the constituents, and effectively integrating them have become a prominent direction for exploration.

In this work, we focus exclusively on the segmentation stage. In addition, we implement a region of interest cropping method to proportionally enhance the target object while reducing computational cost, which when combined with our preprocessing pipeline and model enhancements which we will expound upon, led to improved results shown through our increased metrics. This underscores the improvement of combining both views towards anatomy segmentation.

2 Dataset

We conducted our experiments using the benchmark dataset Automated Cardiac Diagnosis Challenge (ACDC) [1], which contains short-axis cine-MRI scans of 150 patients acquired from two clinical sites using 1.5T and 3.0T MRI scanners. The dataset includes five equally distributed classes of patients: normal (NOR), prior myocardial infarction (MINF), dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM) and abnormal right ventricle (ARV). This suite of patient scans are then split into train and test sets of size 100 and 50 patients respectively, and each patient folder contains:

- A sequence of 2D DICOM slices and 3D volums for each time frame across the cardiac cycle.
- Ground truth segmentation masks for the end-diastolic (ED) and end-systolic (ES) phases, annotated for three anatomical structures: left ventricular cavity (LVC), left ventricular myocardium (LVM) and right ventricular cavity (RVC).
- An info.cfg file containing the patient-level metadata such as weight, heigh, and diagnosis category.
- A 4D sequence representing the full cardiac cycle.

For this study, we focused only on the ED and ES frames, which represent key clinical moments of maximum and minimum ventricular volume and thus are relevant to our segmentation task.

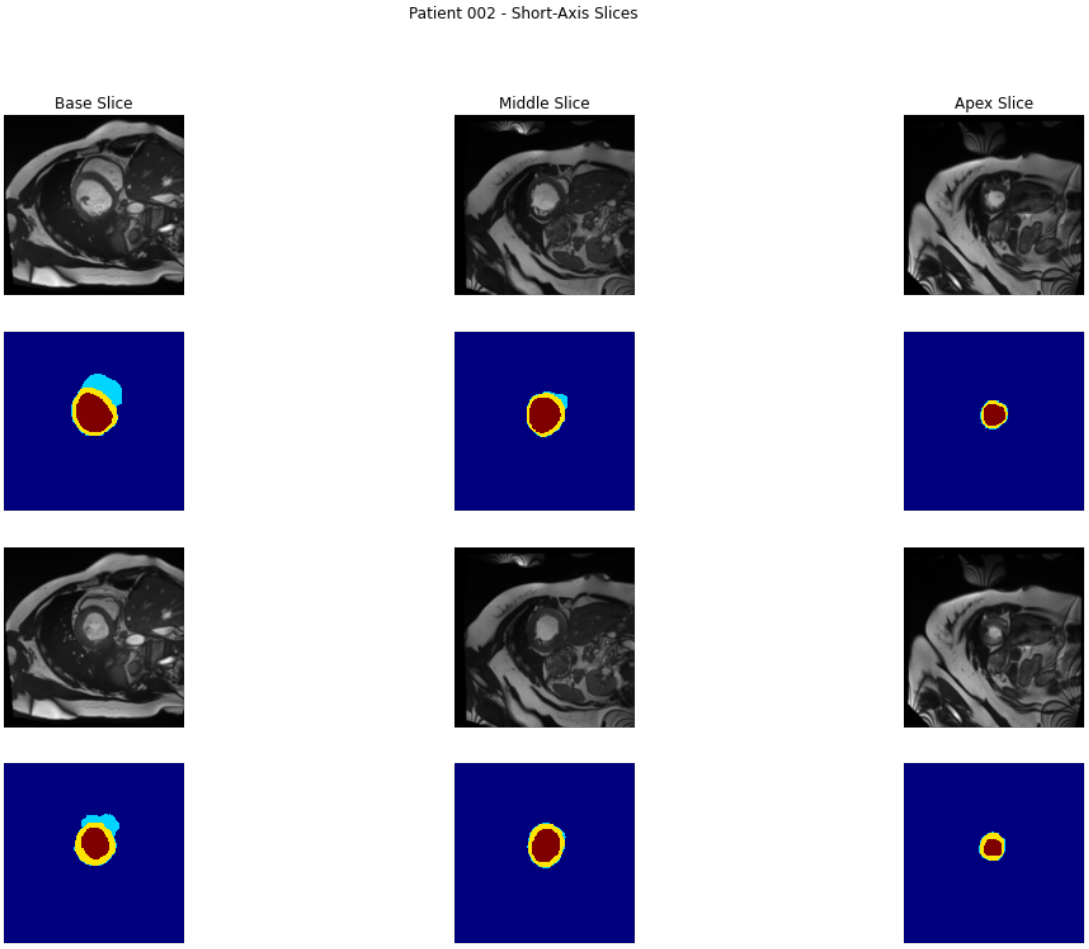

Figure 1: Sample patient MRI scan, including both ED and ES phases and their respective ground truths.

3 Method

In this section, we illustrate our method for segmenting the left, right ventricular cavity and the left ventricular myocardium.

3.1 Preprocessing

For 2D input, to guarantee that every image has the same resolution and relative scale, bilinear interpolation is first used to resize each slice to 352 × 352 pixels. After that, a 224 × 224 pixel center crop is used to highlight the area of interest while preserving anatomical integrity. We use the z-score normalization method to normalize each image's intensity values in order to guarantee consistency in intensity across images. Ground truth masks are transformed into a one-hot encoded format, for ease of learning and to aid in computation of our custom loss function during training. In addition, several data augmentation methods are used to improve model generalization. Among these are random flipping in both the horizontal and vertical directions, which increases resistance to changes in orientation. Furthermore, to dynamically modify image contrast and replicate changes in intensity distribution seen in real-world situations, random gamma correction is applied with a random factor ranging from 0.9 to 1.1.

Preparing our data for 3D input, we first standardize spatial dimensions by resizing each scan to (num_layers, 352, 352) with adaptive cropping/padding techniques. The central region of interest is then extracted while maintaining volumetric integrity by applying a center crop of (10, 224, 224). To guarantee uniformity in the distribution of intensity, we also follow up with Z-score normalization, applied independently to every volume. Segmentation masks are only cropped and resized to match their matching images. Augmentation is used at the scan level for volumetric data, which involves introducing variability in anatomical orientation through random flipping along spatial axes (X, Y). Random elastic deformations also replicate realistic

anatomical variations brought on by scanning artifacts or patient movement. In order to adjust the intensity distribution while maintaining anatomical structures, random gamma adjustment is also included.

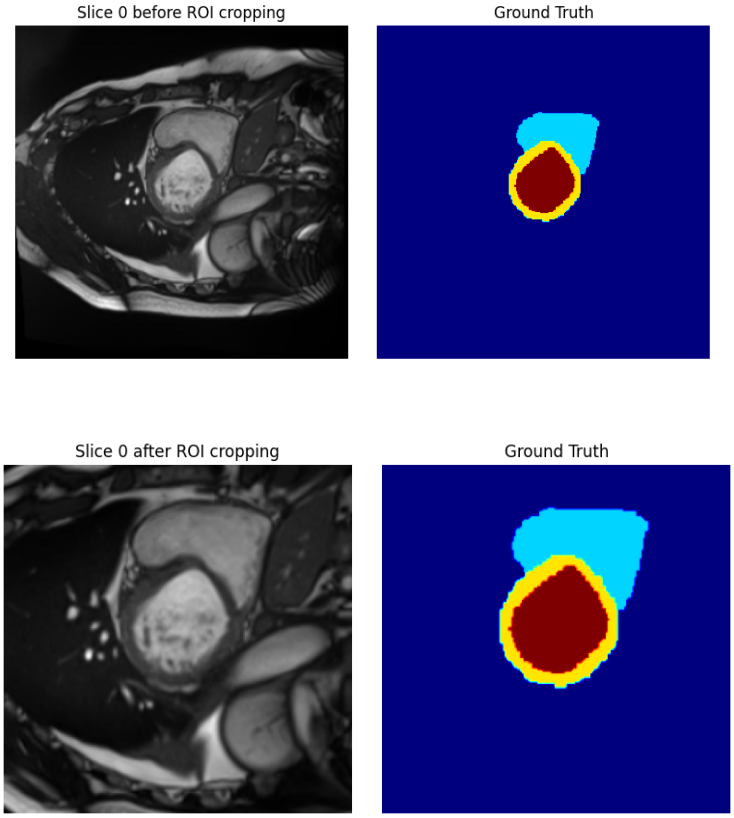

Figure 2: Before and After ROI cropping

3.2 Loss function

As described in Section 2, we are mainly interested in segmenting the 3 regions LVC, LVM, RVC. Thus, we utilized the dice loss function, which is a region-based loss, for both models. It is more suitable for segmenting medical-related regions compared to pixel-based loss function such as Cross Entropy. Mathematically, the loss function is calculated as:

$$L = 1 - \frac{2 * |y_{true} \cap y_{pred}|}{|y_{true}| + |y_{pred}|}$$

where the y_{pred} is the predicted output of the model, y_{true} is the ground truth component.

3.3 2D segmentation

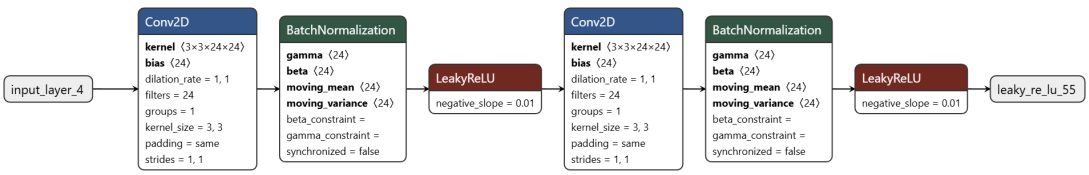

Figure 3: Encoder Block

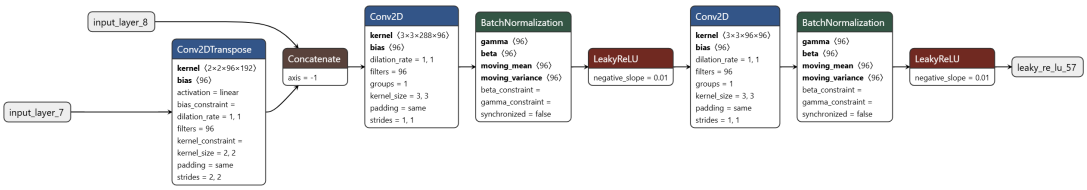

Figure 4: Decoder Block

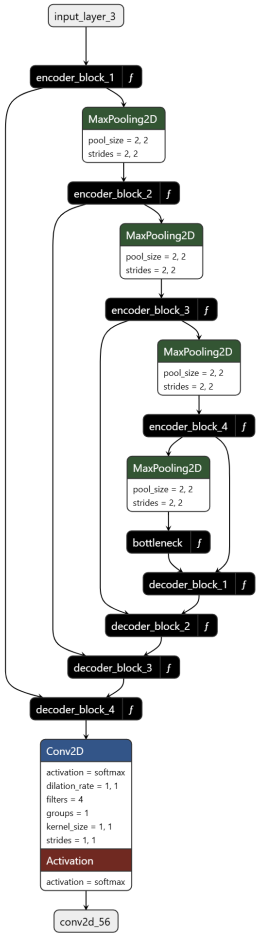

Figure 5: 2D U-Net Architecture

For our 2D U-Net architecture, we built upon the original U-Net framework by [3], with several modifications ensure feature representation with such a small input size. Our encoder convolution blocks are each comprised of two consecutive 2D convolution layers each using He normal initialization, followed by BatchNormalization and LeakyReLU activations. The number of filters we use is relatively small, starting at 24 and progressively doubling through each of the four layers. Each convolution block in this phase is followed by a MaxPooling2D layer with kernel 2x2 and stride 2x2.

The decoder mirrors the encoder, but uses Conv2DTranspose layers for upsampling. Correspondingly to the encoder, each upsampling operation redoubles that spatial resolution while simultaneously halving the number of filters. Skip connections are implemented by concatenating upsampled features with its respective encoder features, preventing spatial information loss from the downsampling steps. These two phases are linked through a bottleneck at the deepest layer, consisting of a single convolution block with 384 filters, serving as the most compressed and semantically rich representation of the input.

At the output layer is a 1x1 convolution with four filters, followed by a softmax activation to produce pixel-wise class probability distribution prediction.

After the preprocessing pipeline, we constructed a modular data loading framework to prepare for training. For each patient, corresponding processed image and ground truth mask

pairs are loaded from .nii.gz volumes, and each volume was normalized to zero mean and unit variance. Subsequently, the 2D slices along the axial plane were extracted, and the masks converted to one-hot encoded categorical format for our three main classes. Our training set was further divided into training and validation sets, where we take the last 4 patient samples of each class for our validation data with which we can monitor our model's performance and robustness. This is true also of our 3D U-Net training session.

The 2D segmentation model was compile with Adam optimizer, using an initial learning rate of 0.001. We employ a custom Dice loss function that calculates Dice loss for each individual class, so that we can directly penalize segmentation misalignments during training. Lastly, we assisted along with the model's optimization process using two callbacks: ModelCheckpoint which saves the best model based of validation Dice loss, and ReduceLROnPlateau with a factor of 0.8 and patience of 10 epochs where the validation loss does not improve. The training was conducted for a total of 170 epochs.

3.4 3D segmentation

For 3D segmentation, we first one hot the ground truth to obtain a tensor of size (3,10,224, 224), i.e., each channel is a binary 3D image. By doing this, we could convert the original multiclass segmentation to binary segmentation.

We utilized a modified version of 3D U-Net [4] proposed in [2]. Figure 6 shows the structure of the model. It is relatively similar for 2D U-Net except that we use 3D convolution layers instead of 2D. The Conv Block consists of 2 convolution layers with kernel of size 3x3x3, stride and padding of 1, followed by batch normalization, leaky ReLU and max pooling. It is important to note the depth remain the same through out the model because we only downsample the height and width. The upsampling is done by using 3D transposed convolution layer with kernel of size 2x2x1. Unlike the traditional U-Net, in the synthesis phase, the upsampled output in the third level is summed with output in the second layer and so on. This allow the model to amplify the pixel that has high value, that is more more likely to be in one of the 3 classes. The process can be summarised by the following equation:

$$output_1 = output_1 + TransposedConv(TransposedConv(ouput_3) + output_2)$$

After the summation, the output go through a 1x1x1 3d convolution layer to learn smaller details and is finalized by a sigmoid activation function to obtain the probabilities:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

The model is trained for 180 epochs, using dice loss and Adam optimizer algorithm with learning rate = 0.0001. The learning rate decreases by a factor of 0.1 if there is no improvement in the dice score coefficient in the validation set in the last 10 epochs.

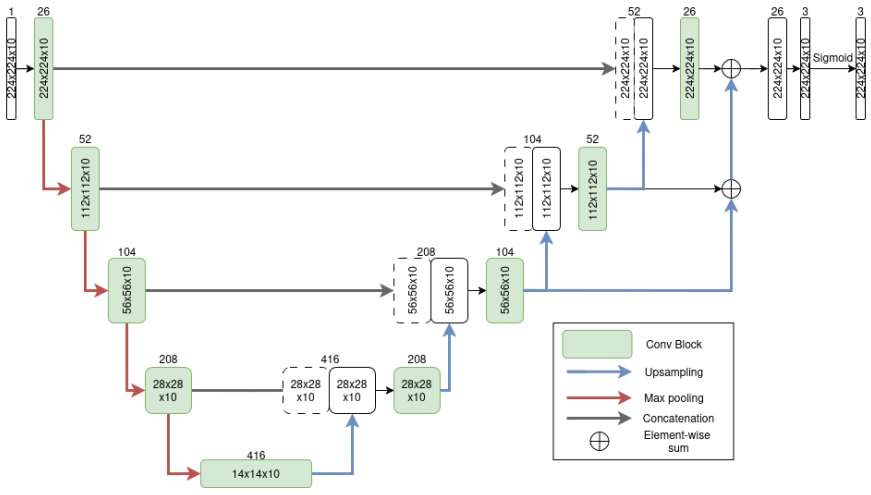

Figure 6: Modified 3D U-Net

4 Results

To evaluate the pipeline, we used Dice Score Coefficient, measuring the similarity between two binary masks. It is calculated by:

$$D = \frac{2 * |y_{true} \cap y_{pred}|}{|y_{true}| + |y_{pred}|}$$

Table 2 compares the results of our method with others. Table 1 compares the dice score between 2D and 3D U-Net.

	RVC	LVM	LVC
2D U-Net	**0.9024**	**0.8818**	**0.9529**
3D U-Net	0.866	0.805	0.905

Table 1: Comparison between different models' output

	RVC	LVM	LVC
[2]	0.923	0.911	0.95
[5]	0.83	0.79	0.9
[6]	0.89	0.85	0.92
[7]	0.78	N/A	0.87
[8]	0.89	0.87	0.92
Ours	0.9024	0.8818	0.9529

Table 2: Comparison to related work

5 Conclusion

In this project, we explored the integration of 2D and 3D U-Net architectures for cardiac structure segmentation. By implementing robust preprocessing techniques we were able to achieve stellar results, with 2D U-Net performing slightly better across all classes. Future work should focus on the potential of combining spatial detail and volumetric context to achieve more accurate and efficient predictions.

References

[1] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, et al., *Deep Learning Techniques for Automatic MRI Cardiac Multi-structures Segmentation and Diagnosis: Is the Problem Solved?*, IEEE Transactions on Medical Imaging, vol. 37, no. 11, pp. 2514–2525, Nov. 2018. doi:10.1109/TMI.2018.2837502

[2] F. Isensee, P. F. Jaeger, P. M. Full, I. Wolf, S. Engelhardt, and K. H. Maier-Hein, "Automatic Cardiac Disease Assessment on cine-MRI via Time-Series Segmentation and Domain Specific Features," in Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges, vol. 10663, M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard, Eds., in Lecture Notes in Computer Science, vol. 10663. , Cham: Springer International Publishing, 2018, pp. 120–129. doi: 10.1007/978-3-319-75541-0_13.

[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)*, vol. 9351, pp. 234–241, Springer, 2015.

[4] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," Jun. 21, 2016, arXiv: arXiv:1606.06650. doi: 10.48550/arXiv.1606.06650.

[5] G. Tziritas and E. Grinias, "Fast fully-automatic localization of left ventricle and myocardium in MRI using MRF model optimization, substructures tracking and B-spline smoothing," in *Proc. STACOM-MICCAI*, LNCS, vol. 10663, 2017, pp. 91–100.

[6] M.-M. Rohe, M. Sermesant, and X. Pennec, "Automatic multi-atlas segmentation of myocardium with SVF-Net," in *Proc. STACOM-MICCAI*, LNCS, vol. 10663, 2017, pp. 170–177.

[7] X. Yang, C. Bian, L. Yu, D. Ni, and P.-A. Heng, "Class-balanced deep neural network for automatic ventricular structure segmentation," in *Proc. STACOM-MICCAI*, LNCS, vol. 10663, 2017, pp. 152–160.

[8] J. Patravali, S. Jain, and S. Chilamkurthy, "2d-3d fully convolutional neural networks for cardiac mr segmentation," in Proc. STACOMMICCAI, LNCS, volume 10663, 2017, pp. 130–139.