Biomedical Physics & Engineering Express

OPEN ACCESS

RECEIVED
7 February 2024

REVISED
8 August 2024

ACCEPTED FOR PUBLICATION
30 August 2024

PUBLISHED
13 September 2024

PAPER

Automatic segmentation of echocardiographic images using a shifted windows vision transformer architecture

Souha Nemri* **and Luc Duong**

Interventional Imaging Laboratory (LIVE), Software and IT Engineering Department, École de technologie supérieure, 1100 Notre-Dame Street West, Montreal, Quebec, Canada H3C 1K3, Canada

* Author to whom any correspondence should be addressed.

E-mail: souha.nemri.1@ens.etsmtl.ca **and** luc.duong@etsmtl.ca

Keywords: echocardiography, semantic segmentation, left ventricle, transformers, U-Net

Abstract

Echocardiography is one the most commonly used imaging modalities for the diagnosis of congenital heart disease. Echocardiographic image analysis is crucial to obtaining accurate cardiac anatomy information. Semantic segmentation models can be used to precisely delimit the borders of the left ventricle, and allow an accurate and automatic identification of the region of interest, which can be extremely useful for cardiologists. In the field of computer vision, convolutional neural network (CNN) architectures remain dominant. Existing CNN approaches have proved highly efficient for the segmentation of various medical images over the past decade. However, these solutions usually struggle to capture long-range dependencies, especially when it comes to images with objects of different scales and complex structures. In this study, we present an efficient method for semantic segmentation of echocardiographic images that overcomes these challenges by leveraging the self-attention mechanism of the Transformer architecture. The proposed solution extracts long-range dependencies and efficiently processes objects at different scales, improving performance in a variety of tasks. We introduce Shifted Windows Transformer models (Swin Transformers), which encode both the content of anatomical structures and the relationship between them. Our solution combines the Swin Transformer and U-Net architectures, producing a U-shaped variant. The validation of the proposed method is performed with the EchoNet-Dynamic dataset used to train our model. The results show an accuracy of 0.97, a Dice coefficient of 0.87, and an Intersection over union (IoU) of 0.78. Swin Transformer models are promising for semantically segmenting echocardiographic images and may help assist cardiologists in automatically analyzing and measuring complex echocardiographic images.

1. Introduction

Congenital Heart Disease (CHD) is the most common type of birth defect among humans, occurring in 0.5-0.8% of all live births, and affecting 1.5 million children worldwide [1, 2]. CHD prevalence is estimated to be 8 cases per 10,000 live births in the population. A major challenge when diagnosing a complex CHD is visualizing anatomical structures.

Echocardiography is an imaging method generally used to acquire anatomical data from the heart. It is a very simple technology used by cardiologists to visualize the heart's 4 chambers. It provides a representation of the heart's movements, producing images of the heart's valves and chambers, without the need for radiation. It allows the cardiologist to visualize the heart and assess its contraction and relaxation, as well as the valve function. The type of echocardiography the patient undergoes may vary as a function of the information needed by the clinician. The left ventricular volume and ejection fraction are two essential volumetric analyses that provide a detailed understanding of cardiac contractility, which leads to better cardiac function diagnosis.

Echocardiographic image segmentation can play a significant role in the automatic analysis and diagnosis of cardiac function. Precise segmentation of anatomical structures in medical images is an essential task for the clinical treatment of certain cardiac diseases. Some cardiac parameters such as the volumes of end systolic and end diastolic, ejection fraction and myocardium mass are good indicators of cardiac health, representing reliable diagnostic value. Clinicians can benefit from the advantages of segmentation to calculate these clinical measurements which are essential for any surgical intervention and treatment follow-up. Recent studies have shown segmentation to be essential and useful for extracting anatomical structures, facilitating the study of medical phenomena and the discovery of new treatments. In most clinical settings, the cardiologist or a trained operator still performs the segmentation step manually, which is laborious and time-consuming, as well as being subject to inter- and intra-observer variability. So automatic segmentation would help physicians in their decision making.

Some research efforts have focused on deep learning to study left ventricular segmentation and to calculate clinical measures for heart disease diagnosis. In particular, the Multi-attention Efficient Feature Fusion Network has been used for automatic segmentation in echocardiography. It incorporates a deep supervision mechanism and spatial pyramid feature fusion to improve feature extraction [3]. Similarly, the calculation of the left ventricular volume (LVEF) represents an effective measure for assessing cardiac health in children, with the deep learning model being adapted to pediatric data. In this context, physiological variations in children are taken into account, and consequently, the model provides an acceptable clinical error and supports the independent assessment of LVEF [4]. Similarly, various projects have been proposed using the U-Net model and its variants, and have achieved good results. The DPS-Net algorithm, based on the U-Net architecture, has shown effectiveness in left ventricle segmentation and ejection fraction measurement across different heart disease phenotypes [5]. However, these methods often require a large number of ground-truth labels, which is time-consuming. To address this, researchers proposed a method that combines multi-level and multi-type self-generated knowledge, using a superpixel approach and various pretext tasks [6].

Convolutional neural network (CNN) models have been the most commonly used models in many applications [7]. In the field of computer vision, CNN architectures remain dominant. They have become the cornerstone of a lot of tasks due to their ability to learn the most important features from our input data. Existing CNN approaches have even proved highly efficient for the segmentation of various medical images over the past decade. Recently, Vision Transformers (ViT) have seen their interest grow significantly in the field of computer vision [8]. The reliance on CNN is not even necessary and a pure transformer applied directly to sequences of image patches can perform very well on images. CNN usually struggle to capture long-range dependencies, especially when it comes to images with objects of different scales and complex structures.

The basic ViT model takes the input image and divides it into several fixed size patches, which are inputted into a neural network. This task is straightforward for small images, but can be computationally intensive for larger ones, such as medical images. With the latter, the basic ViT might fail to capture the spatial information between patches, which may have a negative impact on its ability to accurately perform segmentation. To tackle this limitation, the Shifted Windows (Swin) is proposed to better handle the segmentation of larger images while maintaining a high accuracy. Its hierarchical architecture, which is based on the Swin concept [9], processes high-resolution images by analyzing them in a series of stages at different levels of abstraction. The Swin Transformer's architecture is founded on the Shifted Window Attention (SWA) concept. It groups neighboring patches into a set of overlapping windows. In this context, the algorithm does not apply attention mechanism in a standard fashion; rather, the SWA allows each patch to focus more on patches that are spatially close to it. Patches that resemble each other or are spatially close to one another will have stronger relationships thanks to SWA. These stronger relationships are then used in the U-Net part to perform semantic segmentation.

The SWA of the Swin Transformer thus ensures a better detection of local relationships between patches, which is beneficial for the image segmentation task. The features extracted by the Swin Transformer can then be used by the U-Net part of the Swin U-Net to perform semantic segmentation. This method delivers excellent results in complex image segmentation tasks. The goal of this study is to evaluate a Swin U-Net architecture for the segmentation of echocardiographic images. The study is organized as follows: first, the methodology and the datasets used to validate the proposed approach are described, followed by a presentation of the evaluation of the model. Finally, a discussion and conclusion is provided.

2. Methodology

2.1. Database

Two public datasets were used to evaluate the performance of our proposed approach. The first was the Echonet Dynamic dataset [10], a database designed specifically for the interpretation and analysis of echocardiographic images. Its information allows to visualize the structure of the human heart, while evaluating its function. It contains around 10,030 apical four-chamber echocardiography videos. This dataset was collected at Stanford University Hospital between 2016 and 2018 from medical examinations performed on patients. Each video features a four-chamber apical view of the heart, as shown in figure 1.

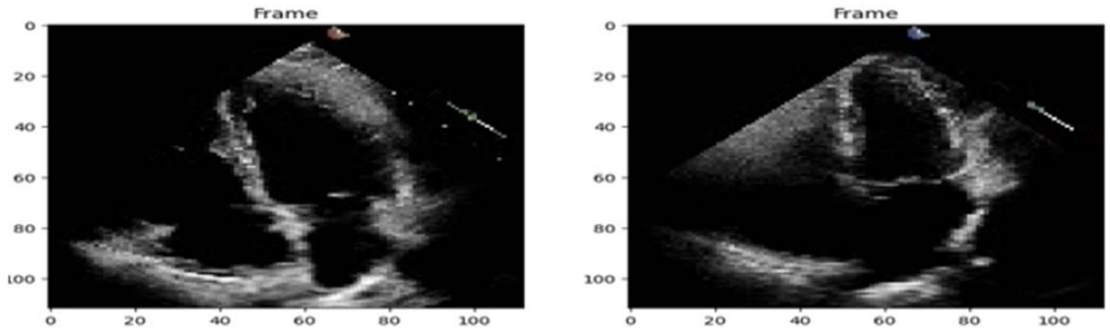

Figure 1. Sample of our database (EchoNet-Dynamic).

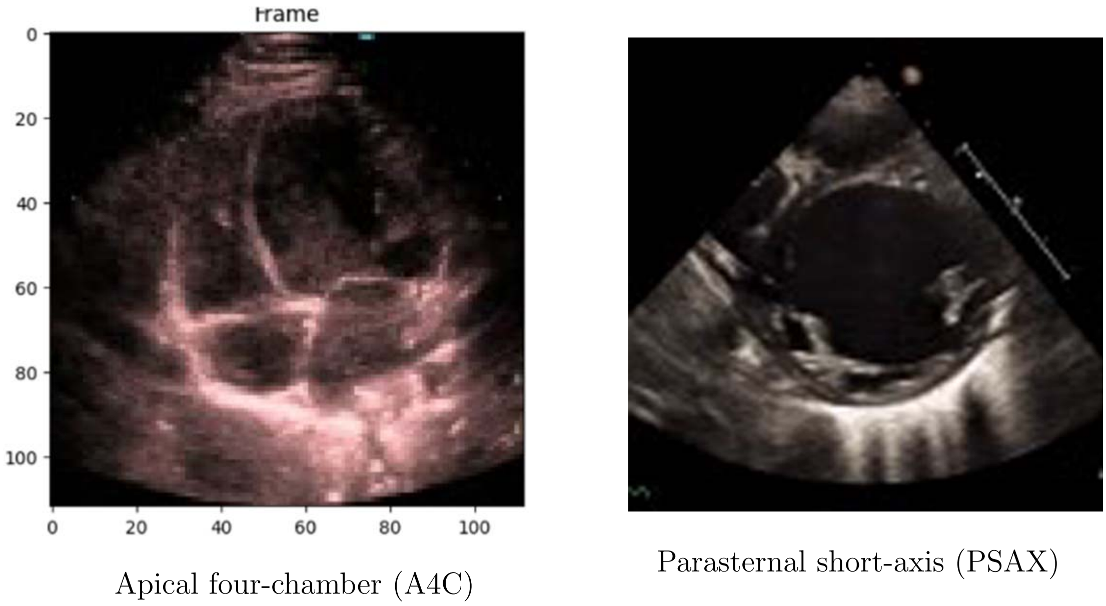

Figure 2. Sample of our database (EchoNet-Pediatric).

For each video sequence, a wide range of information is available, which is useful for diagnostics and follow-up of different cardiac diseases. For example :

i **Number of frames:** This indicates the total number of frames in each video.

ii **Split:** Assigning video to Train or Valid or Test datasets

iii **Frame:** Frame number on which left ventricular segmentation tracing was performed

Next, we used the EchoNet-Pediatric dataset [11]. It consists of a set of echocardiogram videos labeled by human experts such as to give us idea of the assessment of left ventricular function. It was obtained at Lucile Packard Children's Hospital Stanford in the context of routine clinical care, from 2014 to 2021, and from children aged between 0 and 18 years of age, and of different sizes. It contains two-dimensional grayscale clips of A4C (apical four-chamber) and PSAX (parasternal short-axis) views. For our case, however, we focus only on the four-chamber view in order to perform a comparative analysis with the results obtained with the EchoNet-Dynamic dataset. This dataset contains both anatomically normal hearts with normal ejection fraction and patients. Figure 2 provides a sample from our EchoNet-Pediatric dataset.

For both databases, the image sequence consists of a series of 112 by 112 pixel grayscale images. In order to delimit the boundary of the left ventricular cavity wall, it was essential to generate the ground truth masks for our images. To this end, we used the coordinates X1, Y1, X2 and Y2 of the EchoNet-Dynamic dataset.

These coordinates are connecting the most distant points on the ventricle surface to create our line segment, whereas for EchoNet-Pediatric, we used the X and Y columns to generate the corresponding masks. Overlaying the original image with our ground truth mask allows to see the visual results, as illustrated in figures 3 and 4. We have further displayed the

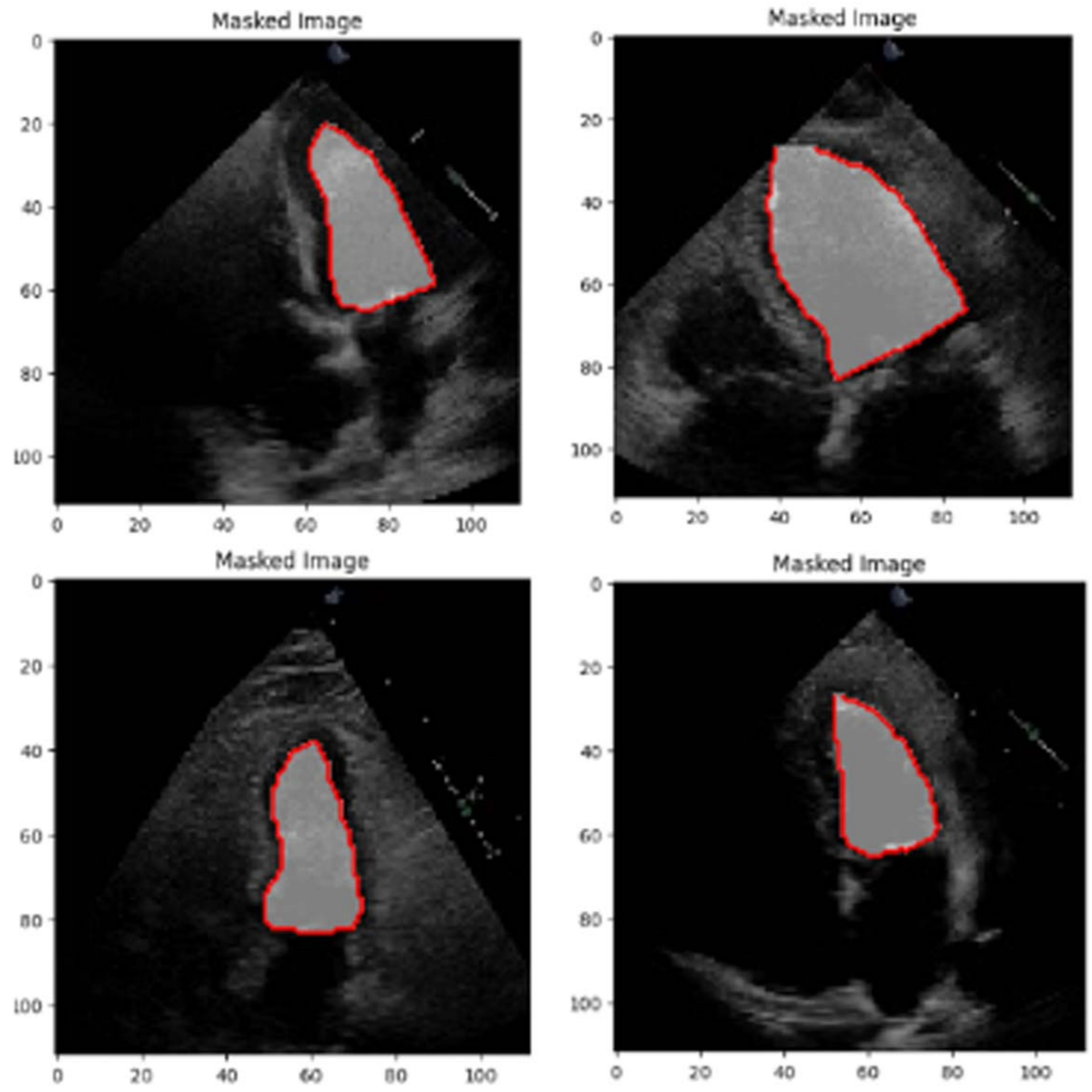

Figure 3. Masked images of EchoNet-Dynamic dataset.

contours of the segmentation mask to better visualize the results. This step is intended to validate that the mask we generated corresponds perfectly to the original image and that the predicted segmentation after training also corresponds to the original images.

2.2. Preprocessing

For each of our databases, there is a total of 15,000 images for the Echonet Dynamic dataset and 6415 for the Echonet Pediatrics dataset. For each of these, we opted to split the data as follows : 70% for training our model, 20% for validation and 10% for testing. All our data loaders were resized to match the input of our Swin Unet model. Our data's dimensions were (256,256,3), while the size of the masks was set fixed at (256,256,1). We then proceeded to normalize our entire database. This step is crucial for any deep learning problem. In this case, all our images were on a similar scale, standing between 0 and 1. Normalization was necessary to allow the neural networks to more easily ensure that the features were on a similar scale. This would guarantee the stabilization of the descent gradient and allow the optimization algorithm to converge much faster and more reliably, avoiding oscillation and slow convergence due to non-normalized or differently scaled features. We used the **Min-max normalization** (**feature scaling**) method, which performs a linear transformation on the original data. This is a common technique used to transform data into a range generally between 0 and 1. This transformation proceeds through the following formula:

$$NormalizedImage = \frac{ResizedImage - min(ResizedImage)}{max(ResizedImage - min(ResizedImage))} \quad (1)$$

Following this operation, we calculated the minimum value of our resized image, as well as its maximum value. The result was then subtracted and divided by the data range, giving values between 0 and 1.

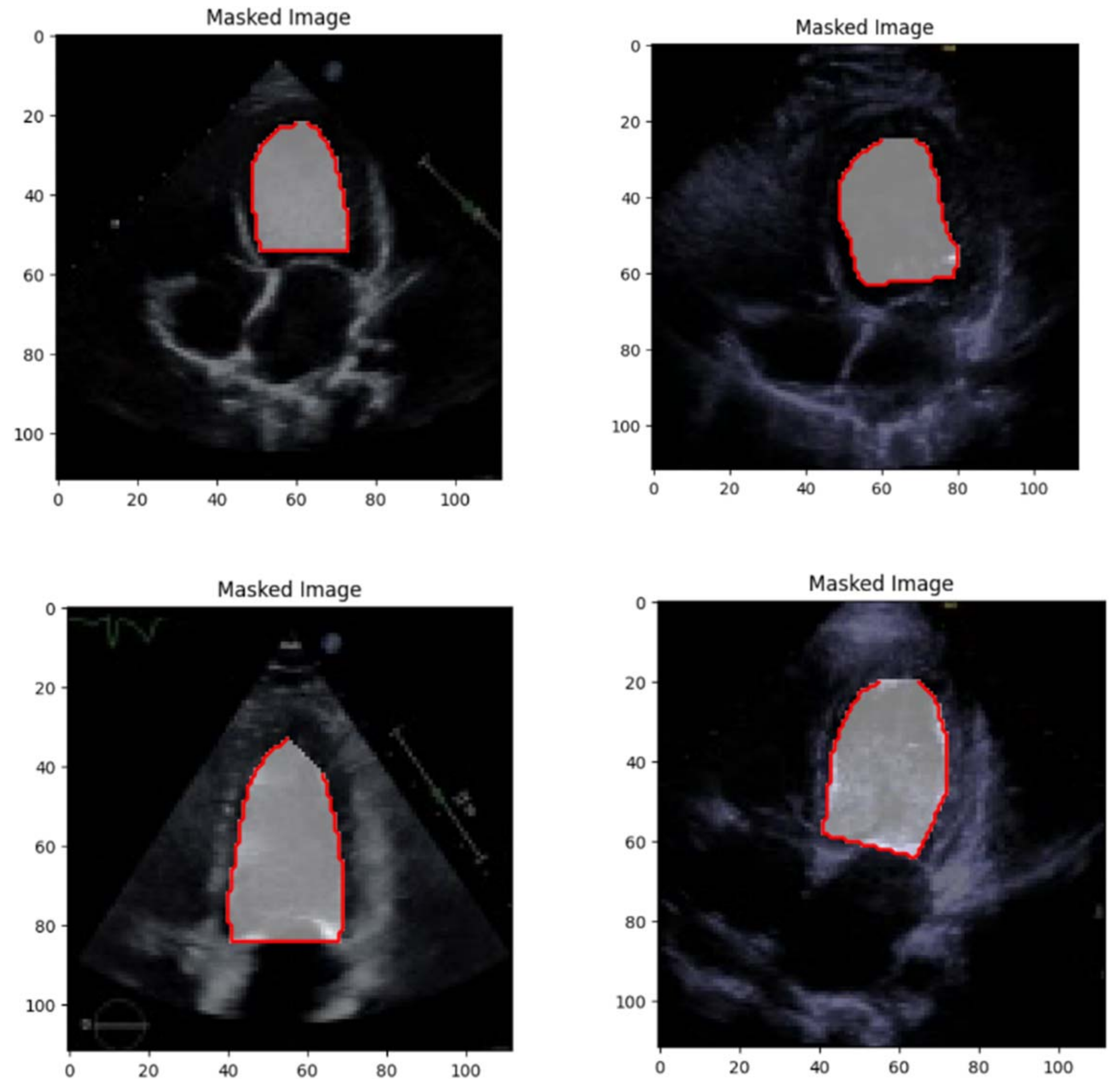

Figure 4. Masked images of EchoNet-Pediatric Dataset.

Our images and masks were then transformed to a common range, simplifying the learning process for the models to be trained. This ensured that features with high values would not dominate during the training process, thus improving the model stability.

2.3. Swin U-Net architecture

A new deep learning model concept is proposed for the segmentation of echocardiographic images, through the Swin U-Net algorithm, which combines the architecture of a CNN encoder-decoder with the structure of a transformer [12]. This U-shaped variant combines the advantages of the Swin Transformer and of U-Net architectures for semantic segmentation tasks. It allows to capture long-term dependencies using the Swin Transformer's self-attention mechanism, while preserving the high-resolution feature representation offered by U-Net.

The original image is initially divided into a set of patches, each of which is processed independently in the transform blocks. Thanks to the auto-attention mechanism, these blocks capture the local relationships within each patch. First, the Swin Transformer encoder processes the input image. Its role is to capture global dependencies and extract high-level features and transmit them to the decoder. Once the decoder has extracted the spatial details, it generates the final segmentation mask. **The encoder** processes the input image in a series of convolution and pooling operations, progressively reducing the spatial resolution. This process allows the model to capture patterns and more complex semantic information, as well as to extract abstract features. After training, we obtain a rich representation of image features. These feature maps have a much smaller spatial dimension and are transmitted to the decoder part of Swin U-Net, where spatial detail is restored.

These low-resolution feature maps, which have been generated at various stages, are now merged together. This process combines information from different scales, thus improving the model's ability to capture both global and local contexts. This fusion enables Swin U-Net's architecture to efficiently capture both global contextual information and the finest

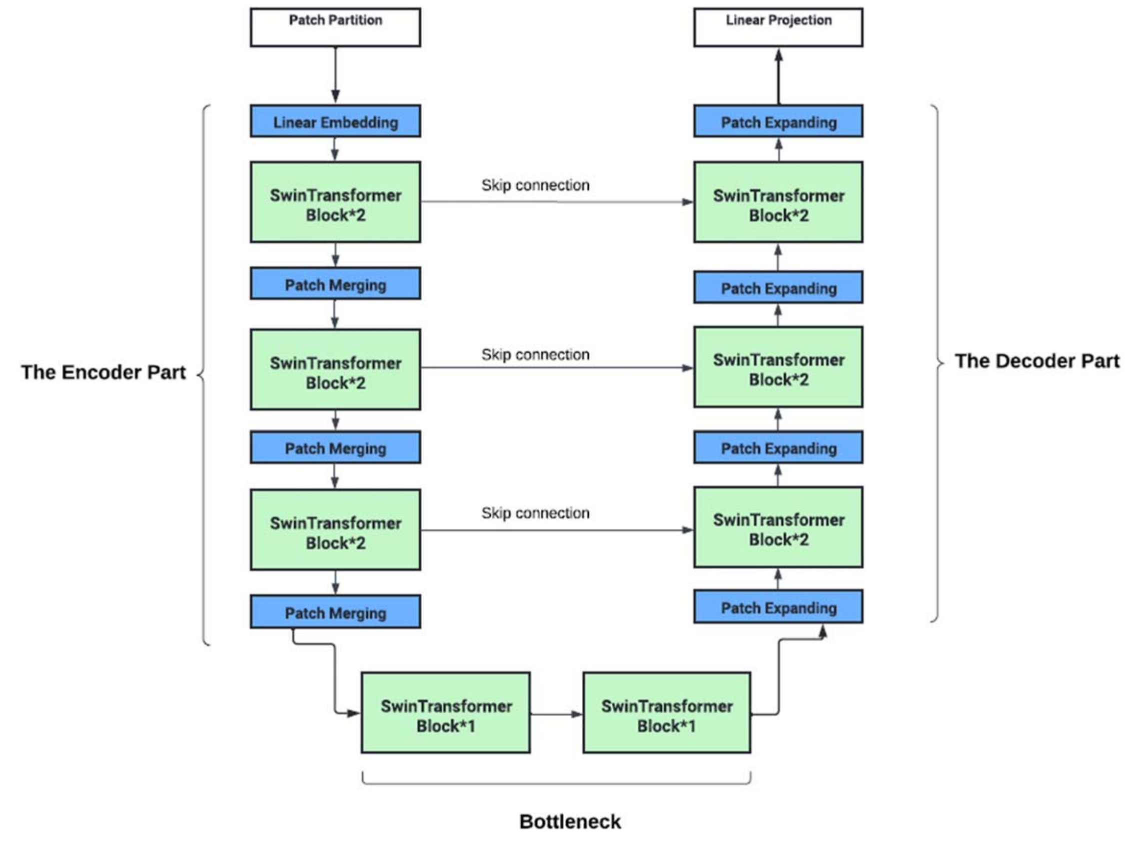

Figure 5. Swin U-net architecture.

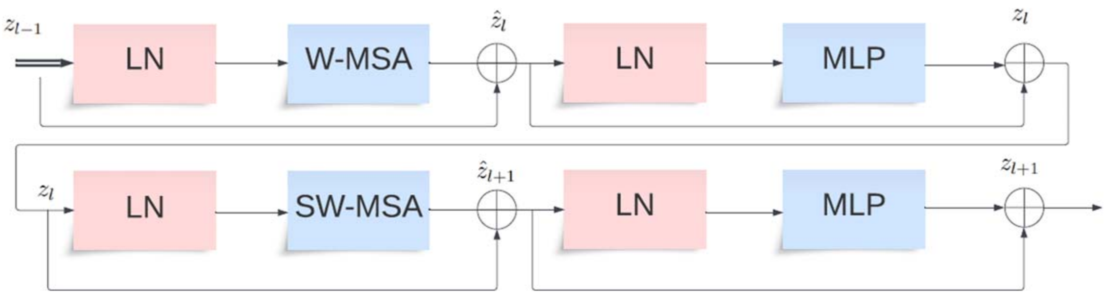

Figure 6. Swin transformer block.

details. This is known as the **Patch Merging Layer** step.

Afterwards, Swin U-Net employs the U-Net structure's **decoder**, which performs upsampling operations. The aim is to increase the spatial resolution of the feature maps received by the encoder. Also, the presence of **the skip connections** in our model is essential, as they link the encoder to the corresponding decoder layers. These connections are responsible for merging the multi-scale features from the encoder with the upsampled features. This action is necessary in order to reduce the loss of spatial information caused by the downsampling in the first part. With these connections, the decoder can simultaneously combine high-level, context-rich information from the encoder with fine, spatially-detailed information from previous layers. This allows the network to carry out more precise predictions and accurately capture complex details in the output. Ultimately, we will be able to recover the finest spatial details that have been lost.

Finally, we add the **Patch Expanding Layer**, which will restore the resolution of the feature maps to the input level. The Linear Projection Layer, for its part, will produce segmentation predictions at the pixel level. Figure 5 illustrates the architecture of the proposed model, highlighting the key components and flow of data through the network.

From figure 6, it can be seen that **the Swin Transformer block** is the basic unit of a symmetric Encoder-Decoder architecture with skip connections, named Swin U-Net. We note the presence of two consecutive Swin Transformer blocks, each composed of a LayerNorm

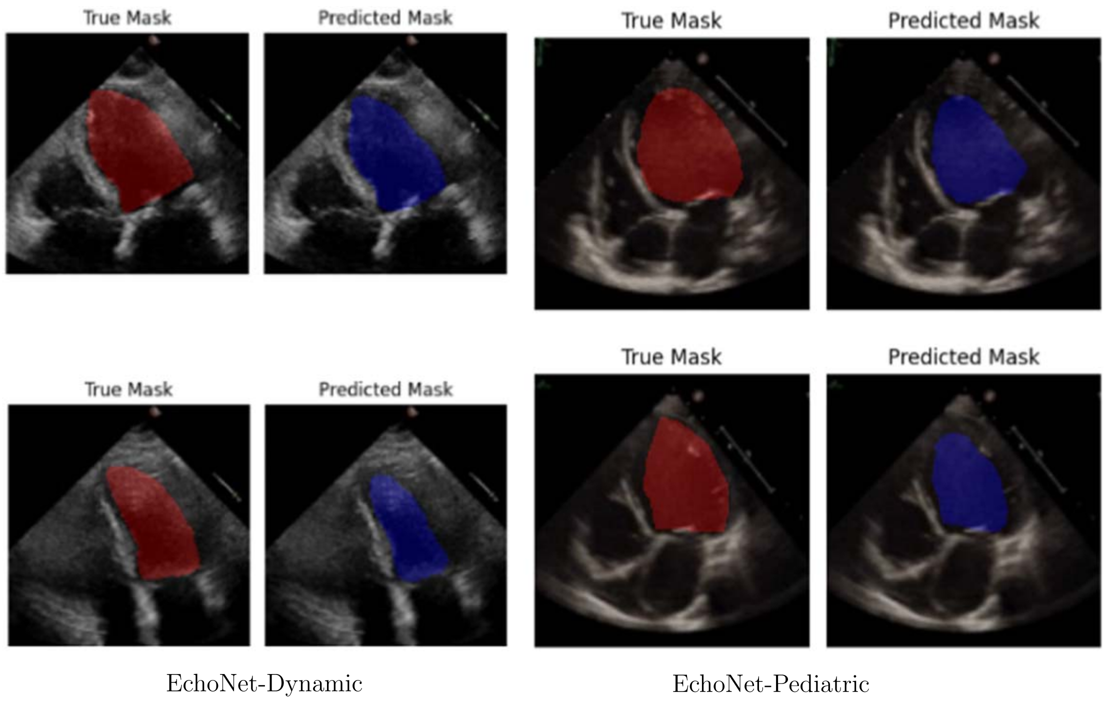

Figure 7. Representation of the segmentation of the left ventricle with Swin U-Net(blue) and its ground truth (red).

(LN) layer, a multi-headed self-attention module, a residual connection and a two-layer MLP with GELU nonlinearity. The two transformer blocks that follow implement the multihead window-based self-attention module (W-MSA) and the multihead offset window-based self-attention module (SW- MSA), respectively.

$$\hat{z}_l = \mathrm{W} - \mathrm{MSA}(\mathrm{LN}(z_{l-1})) + z_{l-1} \tag{2}$$

$$z_l = \mathrm{MLP}(\mathrm{LN}(\hat{z}_l)) + \hat{z}_l \tag{3}$$

$$\hat{z}_{l+1} = \mathrm{SW} - \mathrm{MSA}(\mathrm{LN}(z_l)) + z_l \tag{4}$$

$$z_{l+1} = \mathrm{MLP}(\mathrm{LN}(\hat{z}_{l+1})) + \hat{z}_{l+1} \tag{5}$$

Given that $\hat{z}_l$ is the output of W-MSA and SW-MSA, while z_l is the result of MLP module.

2.4. Implementation details

To define our architecture and make it suitable for the segmentation task, we opted for a value of 1 for the variable n-labels, since this is a binary segmentation and our model will predict a single mask. Our patch size used by Swin Transformers was set to (4, 4), meaning that the input image would be divided into 4x4 patches for initial processing. Prior to the start of our training, various loss functions commonly used for segmentation tasks were available, allowing to compile the model with an appropriate loss function. We chose **Binary cross entropy**, which is commonly used for binary segmentation. In our case, the goal was to delineate the left ventricle of the human heart. This option is robust to imbalance, as the background pixels far outnumber the foreground pixels. **BCE** manages this imbalance reasonably well and penalizes false positives and false negatives effectively, thus promoting balanced segmentation. To be able to evaluate the performance of our model at the end of the training, we had to define common metrics for semantic segmentation throughout the process. **Dice Coefficient** (**DSC**) or F1 is the score most widely used to measure model performance for the medical image segmentation task. Our aim was to observe the overlap between the predicted segmentation and the ground truth. In other words, we can simplify DSC to the following equation:

$$DiceCoefficient = \frac{2 * Area\ of\ Overlap}{Total\ Area} \tag{6}$$

We also used the IoU function, also known as the Jaccard index, which is another evaluation measure commonly used in segmentation tasks:

$$IoU = \frac{Area\ of\ Overlap}{Total\ of\ Union} \tag{7}$$

We have incorporated the Hausdorff distance analysis in order to provide a more comprehensive assessment of segmentation accuracy by quantifying the maximum discrepancy between two sets of points. In this case, the sets A and B are respectively the ground truth and the predicted values.

$$H(A, B) = \max\left\{\sup_{a\in A}\inf_{b\in B} d(a, b), \sup_{b\in B}\inf_{a\in A} d(a, b)\right\} \tag{8}$$

3. Results

In figure 7, it can be seen that Swin U-Net has produced a predicted mask that approximates the ground truth segmentation mask. From a visual standpoint, there is a good overlap between the results

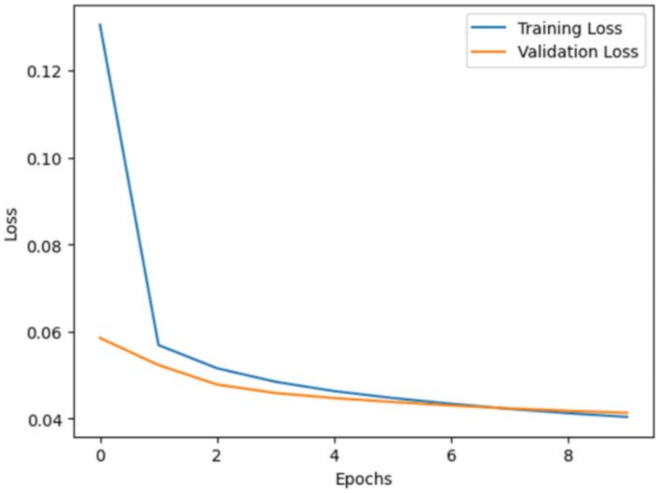

Figure 8. Swin-Unet results (EchoNet-Dynamic).

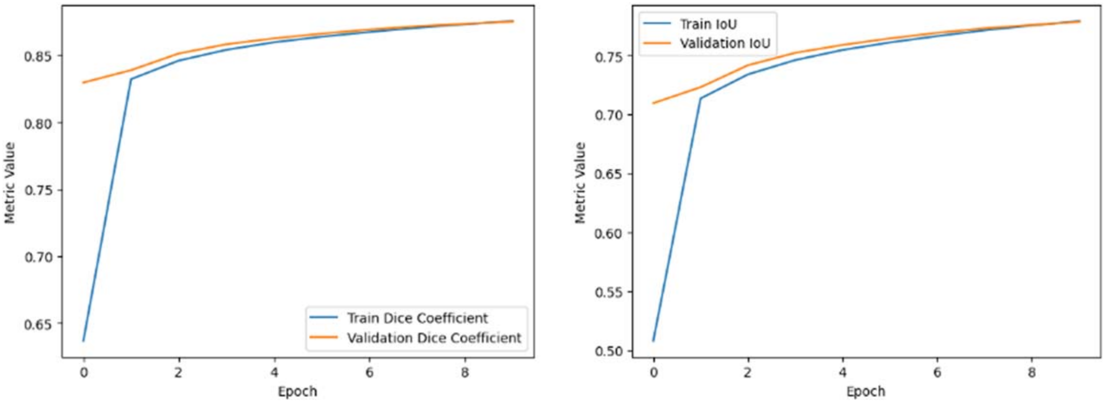

Figure 9. Swin-Unet results (EchoNet-Dynamic).

of our model and the images in the test database. This points to a good delimitation of the left ventricular border achieved by Swin U-Net.

During the training phase, IoU and the Dice coefficient were our basic metrics. Our index provides a perspective on the quality of segmentation predictions and complements the Dice coefficient as a valuable metric for monitoring the performance of our model. Figure 8 shows the results obtained by the scores at the end of the training phase. A progressive decrease can be seen in the loss function, leading to a reduction in the error between predictions and actual values. Towards the end, we were able to reach a value equal to 0.04. On the other hand, we cannot settle for the loss function reduction alone, as it is not sufficient to guarantee a high-performance model.

Figure 9 indicates that the IoU and Dice coefficient scores continue to increase progressively. This indicates a potential overlap between the predicted and the true masks. Our Swin Unet model accurately segments the left ventricle with a Dice coefficient of 0.88 and an IoU score of 0.78.

We have repeated the same process for our second dataset Echonet Pediatric in order to perform a comparative analysis with the results obtained with the EchoNet-Dynamic dataset. Figure 10 illustrates the results of segmentation of our model Swin Unet which indicates the effectiveness of our model in pediatric echocardiography. The model achieved a Dice coefficient of 80.94% which proves a good overlap between the prediction of our model and the ground truth mask. Additionally for our metric Intersection over unit (IoU), with a score approximately equal to 70%, it further confirms the ability of our model to perform a good segmentation of the left ventricle. These metrics show a high level of agreement between model predictions and ground truth annotations, demonstrating the model's ability to accurately delimit left ventricle structure. As illustrated in figure 11, the model training process was also effective, as reflected in the low loss value of 5%, showing that the model has learned to efficiently segment pediatric echocardiographic images while minimizing errors.

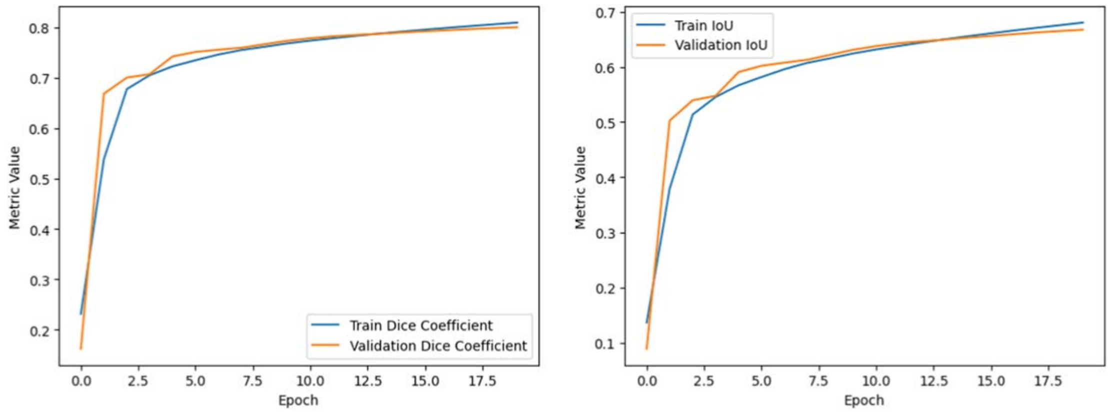

Figure 10. Swin-Unet results (EchoNet-Pediatric).

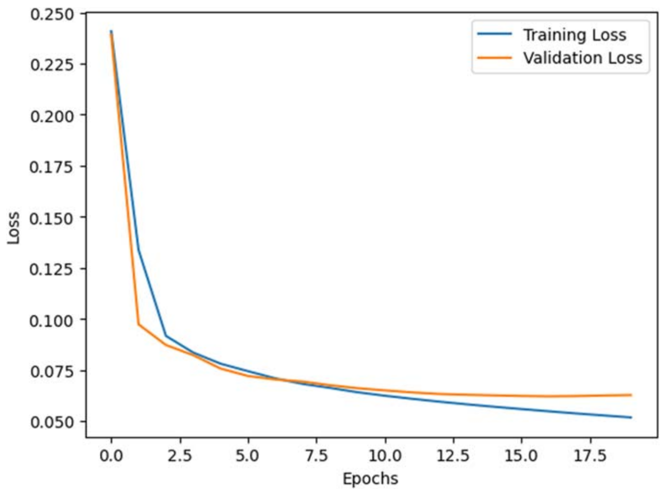

Figure 11. Swin-Unet results (EchoNet-Pediatric).

Table 1. Comparison of Swin U-Net and U-Net results.

		Performance metrics				
		Dice coefficient	IoU (%)	Binary accuracy (%)	Value error (%)	Hausdorff distance
U-Net	Echonet-Dynamic	91.66	84.63	98.45	2.09	2.15
	Echonet Pediatric	87.35	76.04	97.75	5.58	3.18
Swin U-Net	Echonet-Dynamic	88.57	78.93	97.51	4.04	2.89
	Echonet Pediatric	80.94	68.03	97.09	5.18	3.27

To best highlight the robustness of our model in both adult and pediatric databases, we set up the two summary tables below. Indeed, our aim was to emphasize that the combination of convolutional neural networks with Transformers can improve the results obtained by a CNN in a segmentation task. This is illustrated in table 1, where we compare the evaluation metrics results for the two models, U-Net and Swin U-Net, at the end of their training on the Echonet-Dynamic and EchoNet-Pediatric datasets.

4. Conclusion

This study presents a new Swin Transformer U-Net model for the segmentation of echocardiographic images. Swin UNet model has offered several benefits over traditional models for left ventricle segmentation while integrating transformers. Our model leverages the advantages of both convolutional neural networks (CNNs) and transformers, providing a robust framework for accurate segmentation tasks, particularly in challenging cases with varied heart phenotypes. The hierarchical design enables it to process images at multiple scales, ensuring that fine details and broader structural information are captured effectively. This leads to a robust model that can generalize across various types of echocardiographic images and it could be applied in diverse clinical settings, potentially improving the diagnostic accuracy and consistency of left ventricle segmentation. Similarly, our Swin UNet model shows good segmentation accuracy, particularly due to its ability to capture long-range dependencies and contextual information in echocardiographic images. For instance, unlike other models, which rely on traditional CNN architectures, Swin UNet integrates transformer-based mechanisms, enhancing its ability to model complex spatial relationships and improving segmentation performance. This provides a greater understanding and better management of the variability of echocardiographic images.

A validation was done using the EchoNet and EchoNet-Pediatric databases, and indicated a very good performance. The automatic evaluation of echocardiography is very important for diagnosis of congenital heart disease, and it may pave the way for the automatic analysis of clinical parameters such as ejection fraction measurements. The segmentation task remains important since it outlines how ejection fraction estimations are generated. This task generates the equivalent of the manual tracing for each frame between systole and diastole, and might provide validation information about the whole echocardiographic sequence, for each beat. Moreover, since echocardiography allows to visualize the heart motion in real time, it might also provide a better understanding of cardiovascular dynamics and allow the development of more comprehensive models which incorporate the complexities of congenital heart disease [13, 14]. Such information could be integrated into a virtual simulator modeling complex congenital heart diseases. Future work will involve segmentation evaluation on complex anatomies of CHD pediatric datasets. Furthermore, we will investigate our Swin-Unet segmentation model on other imaging modalities such as cardiac MRI or CT scans. This would extend the applicability of the Swin Transformer U-Net model across different imaging techniques for the study of CHD.

Acknowledgment

This work was supported by Natural Sciences and Engineering Research Council of Canada and Mitacs Globalink Graduate Fellowship.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://stanfordaimi.azurewebsites.net/datasets/834e1cd1-92f7-4268-9daa-d359198b310a; https://stanfordaimi.azurewebsites.net/datasets/a84b6be6-0d33-41f9-8996-86e5df53b005.

ORCID iDs

Souha Nemri https://orcid.org/0009-0007-8827-2762

References

[1] Hoffman J I E and Kaplan S 2002 The incidence of congenital heart disease *J. Am. College Cardiol.* **39** 1890–900

[2] Reller M D, Strickland M J, Riehle-Colarusso T, Mahle W T and Correa A 2008 Prevalence of congenital heart defects in metropolitan atlanta, 1998-2005 *J. Pediatr.* **153** 807–13

[3] Zeng Y, Tsui P-H, Pang K, Bin G, Li J, Lv K, Wu X, Wu S and Zhou Z 2023 MAEF-Net: multi-attention efficient feature fusion network for left ventricular segmentation and quantitative analysis in two-dimensional echocardiography *Ultrasonics* **127** 106855

[4] Zuercher M, Ufkes S, Erdman L, Slorach C, Mertens L and Taylor K 2022 Retraining an artificial intelligence algorithm to calculate left ventricular ejection fraction in pediatrics *J. Cardiothorac. Vasc. Anesth.* **36** 3610–6

[5] Liu X, Fan Y, Li S, Chen M, Li M, Hau W K, Zhang H, Xu L and Lee A P-W 2021 Deep learning-based automated left ventricular ejection fraction assessment using 2-d echocardiography *American Journal of Physiology-Heart and Circulatory Physiology* **321** H390–9

[6] Yu C, Li S, Ghista D, Gao Z, Zhang H, Ser J D and Xu L 2023 Multi-level multi-type self-generated knowledge fusion for cardiac ultrasound segmentation *Information Fusion* **92** 1–12

[7] Alexandre A and Luc D 2023 Automatic evaluation of the ejection fraction on echocardiography images *CMBES Proceedings* **45**

[8] Dosovitskiy A *et al* 2021 An image is worth 16x16 words: Transformers for image recognition at scale (https://arxiv.org/abs/2010.11929)

[9] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S and Guo B 2021 Swin transformer: Hierarchical vision transformer using shifted windows CoRR (https://arxiv.org/abs/2103.14030)

[10] Ouyang D *et al* 2020 Video-based AI for beat-to-beat assessment of cardiac function *Nature* **580** 252–6

[11] Reddy C D, Lopez L, Ouyang D, Zou J Y and He B 2023 Video-based deep learning for automated assessment of left ventricular ejection fraction in pediatric patients *Journal of the American Society of Echocardiography* **36** 482–9

[12] Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q and Wang M 2023 *Swin-unet: Unet-like pure transformer for medical image segmentation Computer Vision – ECCV 2022 Workshops* (Springer) 205–18

[13] Azizmohammadi F, Castellanos I N, Miró J, Segars P, Samei E and Duong L 2022 Generative learning approach for radiation dose reduction in x-ray guided cardiac interventions *Med. Phys.* **49** 4071–81

[14] Azizmohammadi F, Castellanos I N, Miró J, Segars P, Samei E and Duong L 2023 Patient-specific cardio-respiratory motion prediction in x-ray angiography using LSTM networks *Phys. Med. Biol.* **68** 025010