CARDIAC MR SEGMENTATION PIPELINE WITH FOUNDATION MODEL

KATHY H. Y. WONG[1], MIKO M. L. CHANG[1], WANTZE VONG[1], BRIAN C. S. LOH[1], CASLON CHUA[2], SWEE KING PHANG[3], PATRICK H. H. THEN[4]

[1]Faculty of Engineering, Computing and Science, Swinburne University, Sarawak Campus, Kuching, Sarawak, Malaysia
[2]School of Science, Computing and Emerging Technologies, Swinburne University, Australia
[3]School of Engineering, Taylor's University, Subang Jaya, Selangor, Malaysia
[4]Sarawak Artificial Intelligence Center, Kuching, Sarawak, Malaysia

E-MAIL: hywong@swinburne.edu.my, mchang@swinburne.edu.my

Abstract:

Cardiac segmentation is an important task to identify heart structures and assess cardiac function in medical imaging. Cardiac segmentation is primarily approached from an image domain perspective, using deep learning models to extricate and identify features not apparent to the human eye. Most efforts focus on improving performance at the model level, crafting deep learning models with specific configurations to tackle the unique task of cardiac segmentation. In this paper, a different approach is applied. Instead of model-level modifications, a pipeline consisting of two-stage detection involving cardiac localization and sub-structure detection is implemented, which then leads to cardiac sub-structure segmentation by applying the MedSAM foundation model. Overall, the pipeline achieves 91% and 88.5% accuracy for the left and right ventricle cavities, which are closely within 5% of the reported performance of SOTA model benchmarks, demonstrating the potential of foundational model for specialized tasks such as cardiac segmentation.

Keywords:

Cardiac; Segmentation; Detection; Foundation Model; MRI

1. Introduction

Cardiac segmentation is an important task to identify heart structures and assess cardiac function in medical imaging. Recent improvements in technology have rendered magnetic resonance (MR) as the gold standard for cardiac imaging due to better spatial resolution and tissue contrast.

Cardiac MR segmentation is well-established task in the field of medical image processing, due to its important contribution in applications such quantifying cardiac function for disease diagnosis, monitoring and prognosis [1].

With the development of deep learning techniques, notable breakthroughs in the field of cardiac segmentation include the U-Net [2], a deep learning architecture consisting of a U-shaped encoding-decoding pathway with skip connections, which is excellent in capturing and retaining local information. Until today, the U-Net is still widely applied as the backbone model for cardiac segmentation tasks, implemented with variations and modifications to further enhance segmentation performance.

A notable implementation is the nnU-Net by Isensee et al [3]. The nnU-Net features a self-configuring architecture of three different U-Nets, with parameters adapted to the characteristics of its input datasets. This resulted in state-of-the-art (SOTA) performance and continues to be frequently referenced as a benchmark in cardiac segmentation tasks.

Transformer variants [4] are also steadily gaining traction in cardiac segmentation tasks. Transformers are highly successful language learning models known for their ability to handle long-range dependencies, a capability the imaging community have adapted to capture global contextual information in visual tasks [5].

Qiu et al [6] employs the characteristics of the U-Net to craft an encoder-decoder network with skip connections, but replaces the convolutional layers with a novel spatially dynamic Transformer mechanism for feature extraction. The result is a model with better ability to capture features of target objects with diverse appearances, leading to improved model generalization and cardiac segmentation performance compared to other U-Net variants.

Overall, it is observed that cardiac segmentation on MR images is primarily approached from an image domain perspective, using deep learning models to extricate and identify features not apparent to the human eye. Efforts often focus on improving performance at the model level, as the models are tailored with specific configurations to tackle the unique task of cardiac segmentation.

While a well-crafted architecture is indeed important for deep learning tasks, the performance of a model is also significantly, if not more, impacted by the quality, quantity

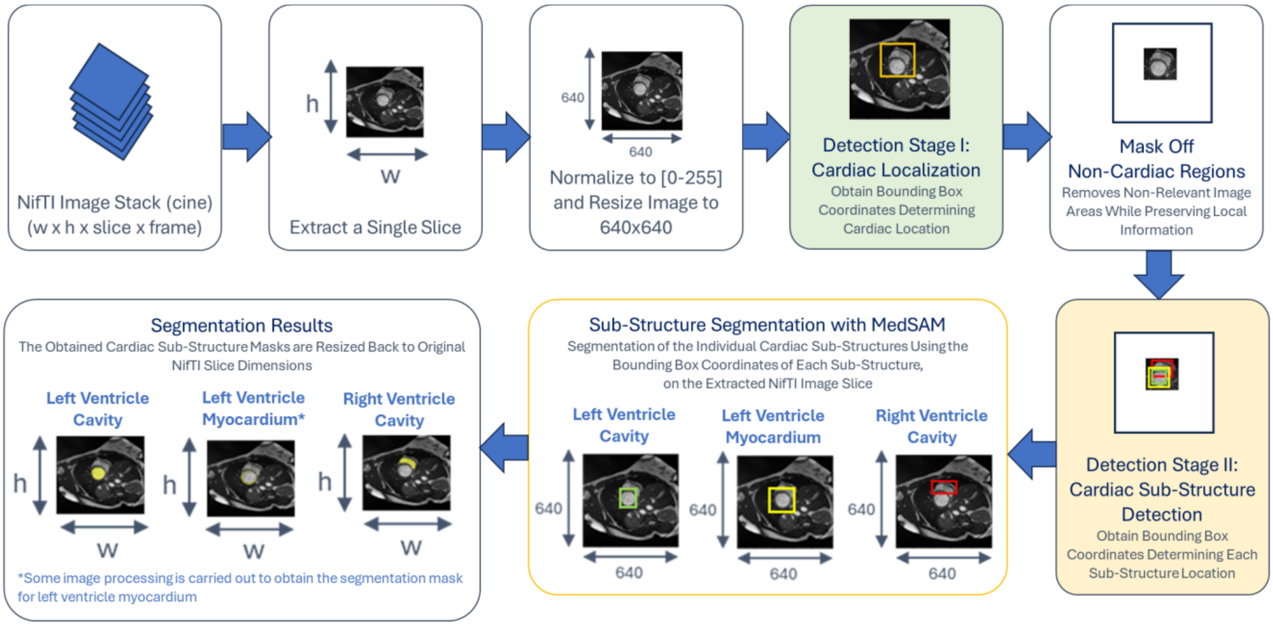
NifTI Image Stack (cine)
(w x h x slice x frame)
h
w
Extract a Single Slice
640
640
Normalize to [0-255] and Resize Image to 640x640
Detection Stage I: Cardiac Localization
Obtain Bounding Box Coordinates Determining Cardiac Location
Mask Off Non-Cardiac Regions
Removes Non-Relevant Image Areas While Preserving Local Information
Detection Stage II: Cardiac Sub-Structure Detection
Obtain Bounding Box Coordinates Determining Each Sub-Structure Location
Sub-Structure Segmentation with MedSAM
Segmentation of the Individual Cardiac Sub-Structures Using the Bounding Box Coordinates of Each Sub-Structure, on the Extracted NifTI Image Slice
Left Ventricle Cavity
Left Ventricle Myocardium
Right Ventricle Cavity
640
640
Segmentation Results
The Obtained Cardiac Sub-Structure Masks are Resized Back to Original NifTI Slice Dimensions
Left Ventricle Cavity
Left Ventricle Myocardium*
Right Ventricle Cavity
h
w
*Some image processing is carried out to obtain the segmentation mask for left ventricle myocardium

FIGURE 1. Methodology of the Implemented Two-Stage Detection and Segmentation Pipeline

and diversity of the data input for training [7].

In the era of deep learning, access to large numbers of data has led to the rise of foundation models. Compared to specialized models tailored for certain tasks depending on their training data, foundation models showcase remarkable generalizing abilities due to being trained on massive amounts of data. In short, foundation models are large-scale pre-trained models which can be adapted to a wide range of downstream tasks with minimal task-specific modifications [8]. This is especially useful for the applications of tasks involving little to no data.

A notable mention is MedSAM [8], adapted from a segmentation foundation model trained on more than a billion segmentation masks across 11 million images, then fine-tuned on more than a million medical image-mask pairs. MedSAM employs a vision transformer backbone with an encoder-decoder network which takes in both images and prompt points in the form of bounding box coordinates for object segmentation. While yet to be applied on cardiac segmentation, it has already outperformed most specialist models on different MR medical datasets.

As MedSAM is a semi-automatic model, this paper thus aims to integrate the implementation of MedSAM into a fully-automated workflow for the application of cardiac MR segmentation.

Rather than implementing model-level modifications, a fully-automated pipeline involving a two-stage cardiac detection with MedSAM is implemented. This paper aims to leverage the superior generalizing ability of foundation model to enable cardiac segmentation, which can be adapted for use in clinical workflows.

2. Methodology

2.1. Proposed Two-Stage Cardiac Segmentation Pipeline

The heart is commonly captured as cross-sectional slices in the short-axis imaging plane, parallel to the oblique axial plane of the body. These slices are commonly identified as either basal, mid-cavity or apical, according to their position relative to the heart.

The three sub-structures of the heart addressed in this paper are: the left ventricle cavity (LVC), left ventricle myocardium (LV MYO), and the right ventricle cavity (RVC), as these are the common sub-structures in cardiac MR segmentation tasks.

From a human perspective, the observation of cardiac slices for segmentation occurs with the following steps:

1. From the medical image, the heart is located.
2. The slice position is determined based on the overall appearance of the heart.
3. Each cardiac sub-structure is identified.
4. The sub-structures are segmented according to their boundaries and differences in tissue contrast.

Based on the steps mentioned above, the two-stage detection with foundation model pipeline is proposed, which consists of:

1. Detection Stages I and II:
   I. Cardiac Localization— this localizes the heart from MR imaging and determines its position.
   II. Cardiac Sub-Structure Detection— the positions of LVC, LV MYO and RVC are then identified given the localized cardiac information from Stage I.
2. Cardiac Segmentation— The sub-structures are segmented with aid from their identified positions.

For both stages of detection, the Ultralytics YOLOv11m object detection model [9] is chosen due to its lightweight and efficient nature, featuring an improved feature extraction compared to its series predecessors. Pretrained on the COCO dataset [10] consisting of more than 300 thousand images and 200 thousand labels, YOLOv11m can generalize well to a diverse range of object detection tasks. This ability can be leveraged through transfer learning on cardiac images for the cardiac localization and sub-structure detection tasks.

For cardiac segmentation, the MedSAM foundation model is applied to obtain the segmentation masks of the individual image slices.

An overview of the complete pipeline is shown in Figure 1, starting with the extraction of the MR images in NifTI format into individual image slices. The images are pre-processed before being passed into Detection Stage I, to identify the heart among the surrounding anatomical tissue.

Once that is done, the non-cardiac regions are masked off. This is used to emulate how humans would "focus" on the heart to segment the cardiac sub-structures. It is also proposed that isolating the cardiac regions ensures better accuracy for the detection of each cardiac sub-structure.

As the appearance of the heart is unique enough among the surrounding tissues, the detection model should be able to successfully identify its position among the surrounding tissues. However, each sub-structure may potentially be confused with the similar-appearing tissues surrounding it. This is due to their irregular shapes and appearances, which is expected to result in some detection inaccuracies if the heart is not initially isolated.

Thus, the implementation of the two-stage detection ensures a safe-guard against inaccurate cardiac sub-structure identification. Once the position of each sub-structure is identified, the components can be brought forward for segmentation.

2.2. Dataset

In this paper, the Automated Cardiac Diagnosis Challenge (ACDC) dataset [11] is utilized. The ACDC dataset features a total of 150 subjects with 4 pathological and 1 control group. 20 patients of each group are provided as training data, and the rest are for testing purposes. The MR images are provided in NifTI format, featuring the cross-sectional slices from the base to apex of the heart, captured across time.

As the dataset is for a cardiac segmentation challenge, the ground truth (GT) segmentation masks for the LVC, LV MYO, and RVC sub-structures of each subject are also provided, but only at end-diastolic and end-systolic stage.

2.3. Pre-processing

In pre-processing, only the relevant cardiac slices with GT data are extracted. This results in 1,076 cardiac image slices from the training set and 1,001 cardiac image slices from the testing set. For training purposes, each image is normalized to [0-255] and resized to 640x640.

To obtain the bounding box coordinates localizing the heart, a process is automated to draw a rectangle spanning the furthest points of the combined GT segmentation masks of LVC, LV MYO and RVC. The normalized center coordinates (x_{center}, y_{center}), width w and height h of the bounding boxes according to YOLO conventions are computed from the rectangle coordinates. The following procedure is computed as in Equations (1)-(9).

Given the combined GT segmentation masks of LVC, LV MYO and RVC is denoted by M for an image of height H and width W, while M' is the aggregated union of all three cardiac sub-structure segmentation masks:

$$M \in \mathbb{R}^{H \times W} \tag{1}$$

$$M' = M_{LVC} \cup M_{MYO} \cup M_{RV} \tag{2}$$

R is the aggregated region of interest, defined as a set of pixel coordinates (x, y):

$$R_{M'} = \{(x, y) \mid M\,(x, y) > 0\,\} \tag{3}$$

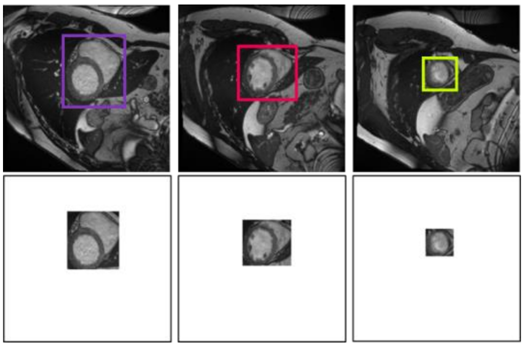

FIGURE 2. Cardiac localization bounding boxes and masking of the non-cardiac regions. From left to right: basal, mid-cavity and apical slice.

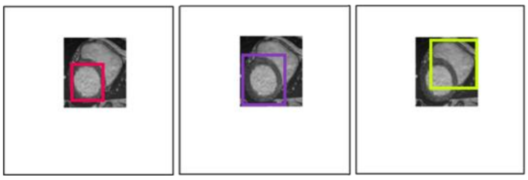

FIGURE 3. GT bounding boxes for the detection of cardiac sub-structures. From left to right: LVC, LV MYO and RVC.

Given that (x_1, y_1) and (x_2, y_2) denote the top left and bottom-right corners of the rectangle respectively, the YOLO-format bounding box conventions are as obtained:

$$x_1 = \min\{x \mid (x, y) \in R_{M'}\} \tag{4}$$

$$y_1 = \min\{y \mid (x, y) \in R_{M'}\} \tag{5}$$

$$x_2 = \max\{x \mid (x, y) \in R_{M'}\} \tag{6}$$

$$y_2 = \max\{y \mid (x, y) \in R_{M'}\} \tag{7}$$

$$x_{center} = \frac{x_1 + x_2}{2W}, y_{center} = \frac{y_1 + y_2}{2H} \tag{8}$$

$$w = \frac{x_2 - x_1}{W},\ h = \frac{y_2 - y_1}{H} \tag{9}$$

Similar to the cardiac localization, bounding box coordinates for each cardiac structure are obtained with the same automated process applied to draw rectangles spanning each cardiac sub-structure segmentation GT.

The models for both detection stages were trained with YOLOv11m object detection models, initialized with the default COCO-pretrained weights. The training employs a ratio of 70%-30% train-validation split, utilizing the default AdamW optimizer, with 120 epochs and a batch size of 14.

To improve model generalizability, the training data underwent the following image augmentations: flipping, rotation and noise addition up to 1.5% of total image pixels.

2.4. Detection Stage I: Cardiac Localization

As shown in Figure 2, cardiac slices differ in appearance, depending on their position. Thus, three main classes are assigned to identify them: basal, mid-cavity and apical. These classes are determined as according to [12]: basal and apical slices are determined by their position adjacent to the base and apex respectively, without the appearance of papillary muscles. The mid-cavity slices are those between the basal and apical, with the appearance of papillary muscles.

2.5. Detection Stage II: Sub-structure Detection

Making use of the cardiac localization coordinates, the areas outside the bounding box are masked off. This isolates the heart for the second stage of detection, while preserving the position of the heart within the image, shown in Figure 3.

The three classes of cardiac sub-structures for detection are: LVC, LV MYO and RVC, where the GT bounding boxes surround. It is worth noting at this stage the detection of LV MYO class is essentially for the detection of the entire left ventricle as the myocardium tissue encases the cavity.

2.6. Cardiac Segmentation Using Foundation Model

With the obtained bounding box coordinates for each identified cardiac sub-structure, these are input into as prompts into MedSAM. The coordinates of the bounding boxes inform the model which regions are to be segmented.

As the output, the segmentation masks for left and right ventricles can be immediately obtained, while some processing is required to obtain the LV MYO segmentation mask. MedSAM tends to only segment the LVC despite the input bounding box covering the entire left ventricle. This is likely due to the boundary line between LVC and LV MYO being quite apparent so that MedSAM just takes the LVC as the object of interest.

To mitigate this, it is found by adjusting the image contrast and inverting the image enables MedSAM to segment the left ventricle as a whole. The contrast is adjusted using an adaptive histogram method called Contrast-Limited Adaptive Histogram Equalization (CLAHE) which is generally applied to increase contrast in images.

CLAHE divides an image into equal regions known as tiles and enhances the contrast of each tile individually. This regularizes the local contrast across an image rather than causing uneven contrast, which may occur in traditional contrast enhancing methods. A tile size of 16x16 is employed

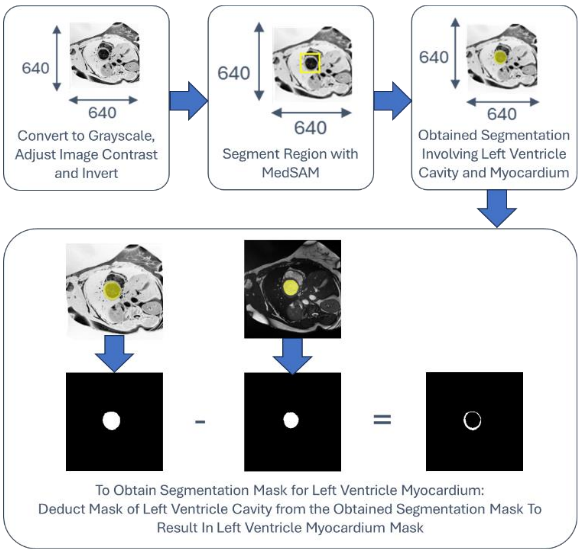
640
640
Convert to Grayscale, Adjust Image Contrast and Invert
640
640
Segment Region with MedSAM
640
640
Obtained Segmentation Involving Left Ventricle Cavity and Myocardium
-
=
To Obtain Segmentation Mask for Left Ventricle Myocardium: Deduct Mask of Left Ventricle Cavity from the Obtained Segmentation Mask To Result In Left Ventricle Myocardium Mask

FIGURE 4. Processing steps involved to obtain final MYO segmentation mask.

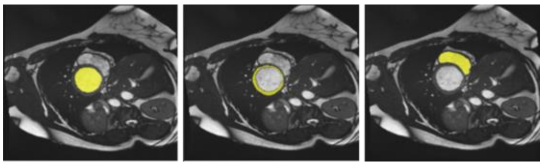

FIGURE 5. From left to right: Final segmentation output for LVC, LV MYO and RVC.

as it is observed that a larger tile size reduces the abrupt transition between image regions, such as the boundary line between LVC and LV MYO. This was found useful to enable MedSAM to segment the entire left ventricle.

As observed in Figure 4, the LVC mask is then deducted from the obtained segmentation mask of the entire left ventricle to obtain the actual LV MYO segmentation. All three cardiac sub-structure segmentation masks are then resized back to the original cardiac slice dimensions as the final output. An example of the final segmentation outputs for LVC, LV MYO and RVC are shown in Figure 5.

2.7. Evaluation Metrics

To evaluate the performance of cardiac localization and sub-structure detection in Detection Stages I and II, the Intersection-Over-Union (IOU) metric (10) is utilized. IOU is a standard performance metric often employed for object detection tasks, which quantifies how well the bounding box region of predicted object overlaps the actual GT bounding box.

$$IOU\ (GT,\ Prediction) = \frac{Area(GT \cap Prediction)}{Area(GT \cup Prediction)} \quad (10)$$

For cases of clinical applications like cardiac detection, an IOU above 85% is typically required to be considered good segmentation performance, as it indicates a strong overlap between prediction and GT.

To evaluate the cardiac sub-structure end-segmentation results, the Sørensen-Dice Coefficient (DSC) is applied to quantify the similarity between GT and predicted segmentation masks (11).

$$DSC\ (GT,\ Prediction) = \frac{2 \times Area(GT \cap Prediction)}{Area(GT + Prediction)} \quad (11)$$

Similar to IOU, DSC measures the overlap between the predicted region and GT. Unlike IOU, DSC places greater emphasis on the correctly predicted regions. DSC is often the preferred metric in clinical segmentation tasks, as it provides a more meaningful assessment in the presence of overall good alignment with minor discrepancies.

3. Results

The results of the segmentation pipeline are as presented in Table 1 and Table 2. Table 1 presents the detection performance for both cardiac detection stages of the implemented pipeline, while Table 2 shows the end performance from MedSAM model for each cardiac sub-structure segmentation with the given cardiac sub-structure bounding box as input.

For comparison purposes, all cardiac data used are from the provided testing set of the ACDC dataset. These, when extracting only cardiac slices with existing GT, sum up to a total of 1,001 cardiac image scans.

TABLE 1. Cardiac localization and sub-structures result

Detection Performance (IOU, %)						
Stage I: Cardiac Localization				Stage II: Cardiac Sub-structures Detection		
Basal	*Mid-Cavity*	*Apical*	*Irrespective of Class*	*LVC*	*RVC*	*LV MYO*
82.2	88.6	76.2	93.0	91.0	89.1	96.6

While the detection accuracies for cardiac localization stage are all less than 90%, it is worth noting that, without the consideration of the class labels denoting the cardiac slice position, 990 out of 1,001 labelled data were successfully detected with a detection accuracy of 93%, showcasing satisfactory performance and reliability in localizing the cardiac structure.

TABLE 2. Comparative studies with SOTA specialized and generalized cardiac segmentation models on the ACDC dataset

Model	**Cardiac Sub-structures Segmentation Performance (DSC, %)**		
	LVC	*RVC*	*LV MYO*
nnU-Net [3]	95.0	92.3	91.1
Fully Convolutional Transformer [4]	95.9	92.6	90.5
Transformer U-Net [6]	96.2	91.1	90.4
MedSAM	91.0	88.5	69.8

In terms of cardiac sub-structure detection, it is observed that the YOLOv11m detection model showcases satisfactory detection results of more than 90% IOU except for the RVC. In fact, the relatively poor detection performance for apical class in Stage I and RVC class in Stage II detection can be attributed to the very small appearances of the heart at the apex, and also the RVC in the basal and apical slices.

From the obtained results, it is observed that the final segmentation results are approximately within 5% of the SOTA model performances, except for the case of LV MYO class. MedSAM can perform well on both LVC and RVC, while performance on the LV MYO is relatively poor.

4. Conclusion

In this paper, a two-stage detection with foundation model pipeline is leveraged cardiac MR image segmentation. The pipeline is found to achieve satisfactory performances for LVC and RVC, while segmentation performance for the LV MYO is comparatively lower. This discrepancy is attributed to its ring-like appearance surrounding the LVC. Compared to the LVC and RVC which appear as singular masses, the LV MYO is more complex in structure.

However, as the end-result of LV MYO segmentation is approaching 70%, it shows the application of MedSAM on segmenting cardiac sub-structures is promising, as prior demonstrations of MedSAM only extend as far as single-mass objects. The foundation model itself is extensively fine-tuned on medical datasets, where items of interest mainly appear as singular masses— these include organs such as the lungs, kidneys, and abnormal body growths like polyps, lesions and tumors.

Furthermore, the instances of heart segmentation were only carried out on chest x-ray images. In such images the heart is observed as a single mass, rather than the cross-sectional view offered in common cardiac MR imaging scans. By applying MedSAM on cardiac MR imaging scans, this work showcases the potential ability of cardiac sub-structure segmentation utilizing foundation model. Overall, LVC and RVC segmentation results yielded are close within 5% of SOTA specialized and generalized model benchmarks, while LV MYO segmentation shows promising results.

Further work will focus on introducing image processing techniques such Gaussian Filtering to diffuse the boundary between LVC and LV MYO, in attempt to improve the LV MYO segmentation performance.

Acknowledgements

This work was supported by the Fundamental Research Grant Scheme (FRGS) by the Malaysian Ministry of Higher Education (MOHE) (FRGS/1/2023/ICT06/SWIN/02/1).

References

[1] H. Lyu, "A Machine Learning-Based Approach for Cardiovascular Diseases Prediction," presented at the Proceedings of the 2022 14th International Conference on Machine Learning and Computing, Guangzhou, China, 2022. [Online]. Available: https://doi.org/10.1145/3529836.3529863.

[2] 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*, Cham, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., 2015// 2015: Springer International Publishing, pp. 234-241.

[3] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods,* vol. 18, no. 2, pp. 203-211, 2021/02/01 2021, doi: 10.1038/s41592-020-01008-z.

[4] A. Tragakis, C. Kaul, R. Murray-Smith, and D. Husmeier, "The Fully Convolutional Transformer for Medical Image Segmentation," presented at the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/WACV56688.2023.00365.

[5] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ArXiv,* vol. abs/2010.11929, 2020.

[6] R. Qiu, J. Yang, S. Kumar, S. Ghosh, and A. Sotiras, *AGILEFORMER: SPATIALLY AGILE TRANSFORMER UNET FOR MEDICAL IMAGE SEGMENTATION*. 2024.

[7] A. Jain *et al.*, "Overview and Importance of Data Quality for Machine Learning Tasks," presented at the Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, CA, USA, 2020. [Online]. Available: https://doi.org/10.1145/3394486.3406477.

[8] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications,* vol. 15, no. 1, p. 654, 2024/01/22 2024, doi: 10.1038/s41467-024-44824-z.

[9] J. Q. Glenn Jocher. "Ultralytics YOLO11." https://github.com/ultralytics/ultralytics.

[10] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014*, Cham, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., 2014// 2014: Springer International Publishing, pp. 740-755.

[11] O. Bernard *et al.*, "Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?," (in eng), *IEEE Trans Med Imaging,* vol. 37, no. 11, pp. 2514-2525, Nov 2018, doi: 10.1109/tmi.2018.2837502.

[12] N. Ho and Y.-C. Kim, "Estimation of Cardiac Short Axis Slice Levels with a Cascaded Deep Convolutional and Recurrent Neural Network Model," *Tomography,* vol. 8, no. 6, pp. 2749-2760, 2022. [Online]. Available: https://www.mdpi.com/2379-139X/8/6/229.