M&Ms-2 Segmentation Report

Anastasiia Deviataieva ANASTASIIA.DEVIATAIEVA@UCDCONNECT.IE
Ehab Patrick Issa EHAB.ISSA@UCDCONNECT.IE
Dylan Forde DYLAN.FORDE@UCDCONNECT.IE
Dónal Heelan DONAL.HEELAN@UCDCONNECT.IE
Julie Sanchis JULIE.SANCHIS@UCDCONNECT.IE

Abstract

In automated cardiac MRI analysis, accurate segmentation of the right ventricle (RV) is crucial for quantifying cardiac function and detecting disease, yet this task remains challenging and has historically received less attention than left ventricular segmentation. The RV's thin wall and complex geometry have even led to it being called the "forgotten ventricle", despite its prognostic significance in conditions such as arrhythmias and cardiomyopathies. We present a compact SegFormer3D-based model, inspired by Perera et al. (2024), tailored for RV segmentation in cardiac MRI. A 4-stage encoder with overlapping $3 \times 3 \times 3$ patch embeddings (stride 2,2,1), no positional embeddings, and two MHSA→MLP blocks per stage with DropPath, yielding an anisotropic pyramid (channels 32/64/160/256). The all-MLP decoder uses $1 \times 1 \times 1$ projections to 128, upsamples to the stage-1 resolution, concatenates scales, then applies a $1 \times 1 \times 1$ head; logits are finally upsampled to input size. The model is trained and evaluated on the public M&Ms-2 challenge dataset. Models were trained with AdamW under a composite Dice+Focal loss, with standard intensity/-geometry augmentations; predictions were binarized with a sigmoid threshold of 0.5. Each study provides two complementary views – short-axis (SA) and long-axis (LA) – ensuring a heterogeneous multi-view dataset for robust evaluation. On held-out tests, we report for SA: Dice = 0.5568, HD95 = 76.8162 mm, precision = 0.4190, recall = 0.4986, F1 = 0.4327. LA results: Dice = 0.8964, HD95 = 62.4874 mm, precision = 0.5264, recall = 0.9197, F1 = 0.6547. These results demonstrate the potential effectiveness of the proposed lightweight Transformer-based approach, particularly for the LA view, in accurate RV segmentation across heterogeneous multi-center MRI data.

Keywords: Deep Learning, Segmentation, SegFormer3D, Medical Imaging, Cardiac MRI (CMR), Right Ventricle Segmentation, Transformers, M&Ms-2.

1. Introduction

Advances in computer vision methods have revolutionised medical image analysis, enabling better automated interpretation of scans and early disease detection. Cardiology in particular has benefitted massively from advancements in imaging, with cardiac magnetic resonance imaging (MRI) having quickly become the gold-standard for capturing heart structure and function (Tseng et al., 2016). A fundamental task in cardiac imaging is cardiac segmentation which involves partitioning the heart into meaningful regions, such as its 4 chambers, which allows clinicians to quantify important metrics for heart health such as ventricular volume and ejection fraction (Shams et al., 2025). Previously, cardiac images were segmented using traditional methods such as Atlas-Based Segmentation (Wachinger and Golland, 2014), however in recent years there has been a shift towards segmentation via deep learning techniques.

Convolutional Neural Networks (CNNs) are a variety of feedforward neural networks which specialise in image segmentation and classification tasks using convolutional layers and filter optimisation. CNNs along with the U-Net architecture (Ronneberger et al., 2015) have become the gold standard for cardiac MRI segmentation and consistently outperform approaches using traditional techniques (Bernard et al., 2018). However, most of this work has focused on left ventricle segmentation due to its importance in assessing conditions such as myocardial infarctions and cardiomyopathy (Hamosh and Cohn, 1971), leaving the right ventricle comparatively forgotten in segmentation research.

The right ventricle plays a critical role in circulatory task by pumping deoxygenated blood to the lungs, however due to its thinner wall and irregular geometry it historically received less attention than the left ventricle, sometimes being described as the “forgotten ventricle” (Tretter and Redington, 2018). The right ventricle can be a prognostic factor for numerous heart conditions, including congenital arrhythmogenesis and right ventricular cardiomyopathy, therefore more precise quantification of the right ventricles size and function could enable much earlier diagnosis and monitoring of heart diseases.

The Multi-Disease, Multi-View & Multi-Center Right Ventricular Segmentation in Cardiac MRI (M&Ms-2) was a segmentation challenge proposed by the University of Barcelona in 2021 (Martin-Isla et al., 2023). This challenge addressed the gap in right ventricle understanding by providing 360 annotated MRI studies and asking participants to develop machine learning models to automatically segment the right ventricle within the MRIs. The studies were obtained from 9 different scanners across 3 different health centers and contained scans of hearts with 8 different cardiac diseases present. Each study consisted of 2 different views of the heart, a long-axis (LA) view and a short-axis (SA) view and the studies were split between a training, validation and test set. Since its release, the publicly available M&Ms-2 data has inspired a variety of different segmentation approaches, with models exploring different deep learning approaches to segmenting the right ventricle. In this report, we will discuss how we built upon these prior efforts to implement our own deep learning model for the segmentation of the right ventricle.

2. Dataloader & Data Augmentation

Before the data was fed into the model, it was first prepared, split logically and augmented to optimise the model’s performance. In order to maximise the available training data for our model, we opted to use the entire dataset including all normal control and diseased hearts.

As our objective was not diagnosis but segmentation, including all conditions ensured that our model would be robust enough to segment a wider range of cardiac presentations.

First the dataset was organised based on MRI view i.e. short-axis or long-axis. For each view, the code pairs up MRI images with their corresponding ground truth label files, excluding the cine sequences. The data was then divided up into training, validation and test sets using scikit-learn's `train_test_split` function. The data was divided up into 40% training, 40% validation and 20% testing, with a fixed seed of 42 used to ensure reproducibility. This created 3 folders for training, validation and testing where each contained 2 additional subfolders for the long-axis and short-axis views of each patient.

Next a second script handled loading each study into the model. A `'get_subjects'` function was written to pair each image with its label and `'get_loaders'` was used to create a separate loader for long-axis and short-axis images so that they could be inputted and processed independently Overfitting to the training set is a common hurdle that must be overcome in all vision tasks. To counter this, the training data underwent a series of augmentations including:

1. RandomFlip – each training image had a 30% chance to be flipped along the x/y axes
2. RandomAffine – applies light transformations to the image including a $\leq 4\%$ zoom-in, a $\leq 3\%$ in-plane rotation and small-left right translations.
3. RandomNoise = each image had a small amount of noise applied to it, with a mean amount of 0 and a standard deviation of 0.01
4. RandomBiasField – mild bias field simulation was used to mimic B1 inhomogeneity
5. RandomGamma – small gamma adjustments within ±0.12 log range were added to alter the contrast of the image

After augmentation, each scan was cropped or padded to a uniform size of $256 \times 256 \times 1$ and standardised using z-score normalisation. This step was critical due to the heterogeneous dataset which uses scans from 9 different MRI scanners. The validation and test data was not augmented in order to provide an unbiased evaluation, instead only undergoing the standard cropping and normalisation preprocessing.

3. SegFormer3D Model Architecture

The model is a compact 3D SegFormer-style network for right-ventricular cardiac MRI. This design is directly inspired by SegFormer3D by Perera et al. (2024), which itself adapts the original 2D SegFormer by Xie et al. (2021) into 3D. A four-stage hierarchical Transformer encoder (overlapping 3D patch embedding + standard MHSA/MLP blocks, no positional embeddings) paired with a lightweight all-MLP decoder. The decoder projects, upsamples, concatenates the four multi-scale features and outputs the mask, then upsamples to the input size. Conceptually, this follows the pipeline shown in Perera et al. (2024), Fig. 1 (four Transformer stages → all-MLP decoder), while our implementation is a minimal 3D variant. For example, Perera et al. (2024) introduce an efficient self-attention variant to further reduce computation in 3D (compressing the token sequence before attention), whereas our implementation sticks to the standard self-attention for clarity and simplicity. But the core idea remains the same: no convoluted decoders, no UNet-style skip connections, just a straight-through Transformer with multi-scale outputs feeding into a slim decoder. Therefore, it remains efficient (≈ 3.9M params) while preserving strong multi-scale context.

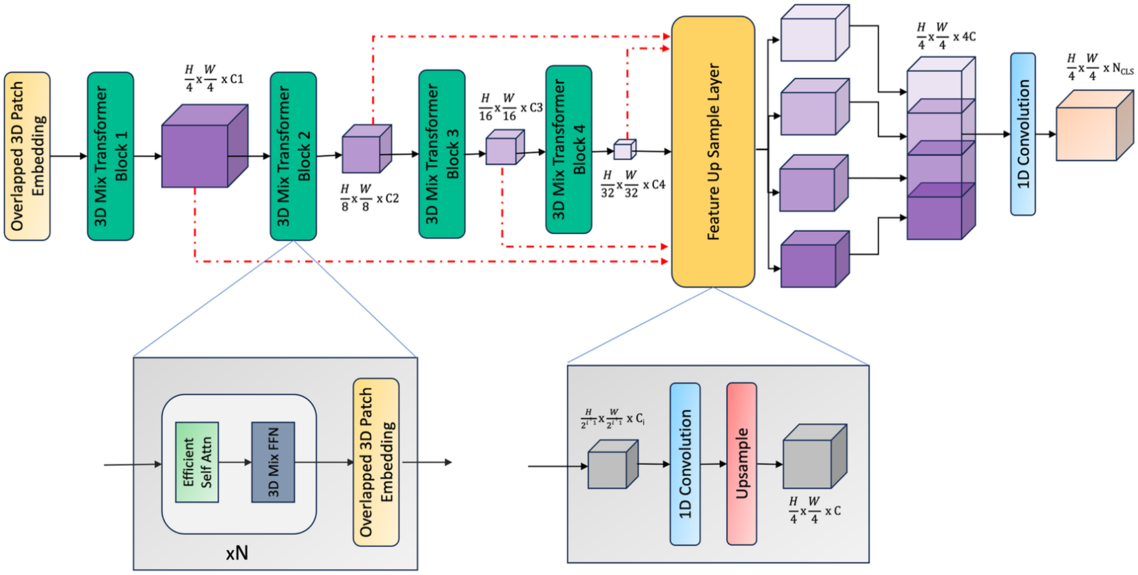
Overlapped 3D Patch Embedding
3D Mix Transformer Block 1
$\frac{H}{4}$x$\frac{W}{4}$ x C1
3D Mix Transformer Block 2
$\frac{H}{8}$x$\frac{W}{8}$ x C2
3D Mix Transformer Block 3
$\frac{H}{16}$x$\frac{W}{16}$ x C3
3D Mix Transformer Block 4
$\frac{H}{32}$x$\frac{W}{32}$ x C4
Feature Up Sample Layer
$\frac{H}{4}$x$\frac{W}{4}$ x 4C
1D Convolution
$\frac{H}{4}$x$\frac{W}{4}$ x N_{CLS}
Efficient Self Attn
3D Mix FFN
Overlapped 3D Patch Embedding
xN
$\frac{H}{2^{i+1}}$x$\frac{W}{2^{i+1}}$ x C_i
1D Convolution
Upsample
$\frac{H}{4}$x$\frac{W}{4}$ x C

Figure 1: (Perera et al., 2024) – SegFormer3D overview, used here as a schematic reference. The figure shows a four-stage hierarchical Transformer that extracts multiscale volumetric features from a 3D input and an all-MLP decoder that upsamples and fuses them into a segmentation mask. In our implementation, the input tensor is $x \in \mathbb{R}^{C\times D\times H\times W}$; each stage uses overlapping 3D patch embedding with $3\times3\times3$ kernels and stride 2,2,1, followed by standard MHSA+MLP blocks without sequence-reduction or Mix-FFN. This yields an anisotropic pyramid (depth fixed at 1 in our data; height downsampled per stage; width preserved) with channels $[32, 64, 160, 256]$. Our all-MLP decoder projects each scale to 128 channels, upsamples, concatenates, and predicts logits – i.e., it fuses multiscale features rather than explicitly aggregating "local vs. global attention" as phrased in the caption of the original figure.

3.1. Encoder

The encoder (`MixTransformer3D`) applies, at each stage, a Conv3D-based overlapping patch embedding to the input volume to produce tokens, then processes those tokens with attention-based Transformer blocks. This layer is essentially a strided 3D convolution that partitions the volume into small patches (voxels or regions) while allowing patches to overlap at the edges, thereby preserving local continuity between neighboring tokens. For each patch, the convolution produces an embedding vector (i.e. a "token"), and a LayerNorm is applied to normalize these token features. Importantly, by using overlapping patches (rather than disjoint ones), the encoder avoids losing boundary information at patch edges, which improves segmentation precision. After this patch embedding step (which also performs downsampling), the volumetric data is no longer treated as a regular image grid but as a sequence of tokens. No positional encoding is added to these tokens, consistent with SegFormer's design that positional cues can be learned implicitly by the network (for example, via the subsequent feed-forward layers). This positional-embedding-free approach means the encoder's attention layers work purely with the content of patches, enabling the model to handle varying input resolutions without needing to resize or interpolate fixed position vectors.

The encoder is organized into four stages, each corresponding to a different resolution and feature granularity. At the end of every stage, tokens are reshaped back to $[B, C, D, H, W]$ and this 3D feature map is fed to the next stage's OverlapPatchEmbed3D, which performs anisotropic downsampling (stride (2,2,1)) and increases the channel dimension. This yields a hierarchical pyramid. For the common input $[B, C, D = 1, H = 256, W = 256]$, spatial sizes evolve approximately as: Stage 1: $1 \times 128 \times 256$, Stage 2: $1 \times 64 \times 256$, Stage 3: $1 \times 32 \times 256$, Stage 4: $1 \times 16 \times 256$, with channel widths $32 \rightarrow 64 \rightarrow 160 \rightarrow 256$. Each stage then applies a sequence of Transformer blocks to the tokens. Every Transformer block consists of a multi-head self-attention (MSA) layer and a feed-forward network (FFN), each preceded by LayerNorm and followed by a residual connection. Specifically, within a block the tokens first go through LayerNorm and an MSA module that allows each token to attend to others (capturing long-range dependencies in the volume). The attention output is added back to the input (residual skip), then another LayerNorm and a two-layer MLP (the FFN) are applied to each token. The MLP is a pure Linear – GELU – Linear with expansion ratio 4 (no Mix-FFN conv). A residual addition after the FFN yields the final output of the block. Notably, DropPath (stochastic depth) is used on these residual connections during training, meaning each block can randomly drop its skip connection with some probability, which helps regularize the model. This Transformer-block architecture (LN → MSA → +skip, LN → MLP → +skip) is repeated a few times per stage (exactly 2 blocks per stage), refining the token representations at that scale.

Across the stages, the encoder produces a set of multi-scale 3D feature maps. Channel width increases with depth, providing fine-detail features from earlier stages and higherlevel semantics from deeper stages. Within each stage, computations after patch embedding happen in token space (self-attention + MLP). The model avoids heavy 3D conv stacks; instead, attention/MLP operate on tokens, and inter-stage transitions reshape tokens back to 3D. In summary, the encoder yields four feature maps of different resolutions (from coarse to fine) and different channel dimensions, encapsulating a rich hierarchy of features. This strategy allows the model to capture both local details and global context.

3.2. Decoder

The decoder, implemented as a `SegFormerDecoder3D`, is a minimalist all-MLP decoder that merges the encoder's multi-scale features into the final segmentation output. Unlike traditional UNet-like decoders, it does not employ any deep convolutional upsampling blocks or complex skip connections; instead, it follows SegFormer's principle of lightweight feature fusion. The decoder first applies a $1 \times 1 \times 1$ convolution to each of the four encoder feature maps to project them to a common decoder channel dimension (a sort of feature alignment). For example, if the decoder uses 128 channels internally, each encoder stage output (whether it originally had 32, 64, 160, or 256 channels) is linearly mapped to a 128-channel feature map via a pointwise Conv3D. This does not change the spatial size of each feature map, only their depth (channels). Next, each of the projected feature maps is upsampled to a common spatial resolution using trilinear interpolation. In this implementation, the target resolution is the stage-1 encoder output (`features[0]`). In this model, since stage 1 already preserves a relatively high resolution, all deeper stage features (which are smaller) are interpolated up to match the stage 1 resolution (e.g. upsampled to the same depth, height, and width as the first stage's feature map). Once all feature maps are at the same size and channel count, they are concatenated together along the channel dimension, yielding one merged feature volume that contains information from all scales. This concatenation essentially performs the role of "skip connections" by bringing together low-level and high-level features, but in one combined tensor rather than through multiple paths.

After concatenation, the decoder passes the fused features through a lightweight classification head to produce the segmentation mask. In practice, this head is implemented as a small stack of 3D convolutional layers (which can be seen as a kind of MLP operating on the channel dimension of each voxel). The decoder uses a simple sequence: a $1 \times 1 \times 1$ Conv3D to mix the concatenated features (reducing the channels to 128), followed by BatchNorm3d, ReLU, Dropout3d(0.1), and then a final $1 \times 1 \times 1$ Conv3D that outputs the desired number of classes (in this model, `num_classes=1`). This final conv layer acts as a linear classifier on each voxel, utilizing the rich multi-scale representation that the previous steps compiled. In `SegFormer3D.forward`, these logits are then upsampled to the original input size using trilinear interpolation. The entire decoder thus contains no heavy spatial convolutions – only pointwise ops and upsampling – which keeps it very computationally light. Despite its simplicity, this design effectively combines fine-detail and coarse-context information:

the higher-resolution features contribute fine detail (boundary and texture of the ventricle), while the lower-resolution features contribute global context (overall position and shape). By aligning and summing up these contributions in an MLP fashion, the decoder can produce accurate segmentation masks without the need for complex decoders seen in other architectures. Notably, this approach was key to SegFormer's success, as it shows that a compact decoder can suffice when the encoder has already produced strong multi-scale features. In our 3D adaptation, we maintain that philosophy, which results in a very fast and memory-efficient segmentation pipeline.

3.3. Advantages/Limitations & Potential Improvements

Compared with other architectures (e.g. UNet-like, pyramid/attention-based, and transformer -based segmenters) paired with standard encoders (e.g., ResNet/ResNeXt, EfficientNet/MobileNet, and MiT) – this model uses a lighter decoder and obtains global context via self-attention in the encoder, while remaining robust to resolution changes because no positional embeddings are interpolated. The trade-offs relative to those same baselines are clear: attention remains $\mathrm{O}(N^2)$ in the number of tokens and can become costly on large volumes; the current code path is effectively near-2D (D=1 inputs and stride (2,2,1) mean that depth is not downsampled), so true through-plane context is limited compared with fully 3D CNNs; and local inductive bias is provided only by the patch embedding and attention – here is no Mix-FFN in the feed-forward network and no heavy UNet-style decoder blocks–so very fine boundaries may require careful loss design and augmentation.

Future work follows directly from this code path. A first step is a true 3D extension: accept $D > 1$ volumes and introduce depth downsampling (e.g., stride (2,2,2) in later stages), keeping compute manageable with windowed or sequence-reduced attention. Locality can be strengthened by adding an optional Mix-FFN (a depthwise 3×3 inside the FFN) while retaining the positional-embedding-free design. Pretraining is likely to yield the largest gains: (i) inflate 2D MiT/SegFormer weights to 3D by copying filters along depth; (ii) use self-supervised masked autoencoding on large unlabeled cardiac MR stacks; and (iii) leverage cross-dataset 3D medical pretraining before fine-tuning on right-ventricle segmentation. Given the current near-2D behaviour, also explore multi-slice or multi-view fusion (stacking neighbouring slices or LA/SA fusion) to inject through-plane information without a large memory hit. Additionally, explore hybrid designs: a shallow 3D-conv stem before the first Transformer stage to strengthen local bias; conv–Transformer co-stages (e.g., a residual 3D-conv block followed by MHSA within each stage) to balance locality and global context; or a lightweight UNet-style decoder with skip connections on top of the current projection/upsampling path when boundary quality is critical. A compact cross-attention decoder that conditions high-resolution features on stage-4 tokens is another viable option. Finally, run controlled, same-protocol comparisons against other architectures to quantify speed/accuracy/memory trade-offs on M&Ms-2.

4. Training

The model training is handled by the `train_model` function. The training process involves iterating through epochs, calculating the loss, performing backprop, and updating the model weights using AdamW optimiser.

The loss function used is a combined loss, which is a sum of the Dice Loss and Focal Loss. The Dice loss component measures the overlap between the predicted segmentation and the ground truth, while the Focal Loss addresses the class imbalance by down-weighting easily correctly classified examples. The training process includes moving the model and data to the available device, using the AdamW optimiser for weight updates, employing gradient accumulation with `accumulation_steps` to simulate larger batch sizes, calculating and tracking both training and validation loss per epoch using TensorBoard, and saving the models training to a state dictionary.

The data loading and preparation are handled by the `get_loaders function`. To benefit robustness and prevent overfitting, the training data undergoes augmentation using torchio transformers. Including augmentations such as CropOrPad, ZNormalisation, RescaleIntensity, and RandomFlip. The validation and testing data only undergo the CropOrPad, ZNormalisation, and RescaleIntensity transforms, without augmentation. Separate data loaders are created for the training, validation, and testing sets for both SA and LA modalities.

5. Evaluation

After being trained, the performance of the proposed 3D SegFormer model was assessed on test sets for both Short-Axis (SA) and Long-Axis(LA) cardiac MRI images. A hold-out test set was used to perform the evaluation task, to ensure unbiased assessment of the model's generalization and prevent overfitting, providing a realistic estimate of real-world performance. The test set consisted in preprocessed MRI images and corresponding ground-truth labels, preprocessed like during training for fair comparison. The evaluation focused on quantitative metrics to assess segmentation accuracy. The metrics evaluated both overlap-based and boundary-based aspects of performance. They were computed per sample and averaged across the test set:

- Dice Coefficient (DSC): measures the spatial overlap between predicted and ground-truth segmentations, defined as $DSC = \frac{2|P \cap G|}{|P| + |G|}$, where P and G are respectively the predicted and ground-truth masks. A smoothing factor of 10^{-5} was added to avoid division by zero.

- 95th Percentile Hausdorff Distance (HD95): quantifies the maximum boundary discrepancy between predicted and ground-truth contours, focusing on the 95th percentile to mitigate outlier sensitivity. It was computed using the MONAI library.

HD95 was only calculated for samples where both the prediction and ground truth contained foreground voxels to avoid undefined distances for empty cases. This metric complements Dice Coefficient by assessing contour accuracy, which is critical in medical segmentation where precise boundaries inform clinical measurements.

- Precision, Recall and F1 Score: were derived from binary predictions and labels, using scikit-learn functions. Precision $(= \frac{TP}{TP+FP})$ assesses the accuracy of positive predictions, recall $(= \frac{TP}{TP+FN})$ assesses the completeness of detection and F1$(= \frac{2 \times Precision \times Recall}{Precision+Recall})$ provides the harmonic mean. These metrics were computed to deal with the potential foreground-background imbalance, offering some insights into false positives and false negatives.

Moreover, predictions were binarised with a sigmoid activation and a 0.5 threshold to balance sensitivity and specificity.

6. Results

The models for SA and LA were evaluated on the test loaders with a batch size of 1 for precision. We obtained those results:

Metrics	**SA model**	**LA model**
Dice score	0.5568	0.8964
Hausdorff distance	76.8162	62.4874
Precision	0.4190	0.5264
Recall	0.4986	0.9197
F1 score	0.4327	0.6547

The evaluation showed differences in performance for SA and LA models, reflecting differences in anatomical complexity and data characteristics. For the SA views, which typically capture more detail of the heart chambers, our model achieved a moderate overlap represented by a dice score of 0.5569 and showed some weaknesses with boundary precision with a high-value Hausdorff distance of 76.8162. These could be due to higher variability in slice orientations. In comparison, the LA model demonstrated better performance in overlap with a high dice score of 0.8964. The Hausdorff distance remained high highlighting some localized errors. Therefore the LA model performed well overall but there are some localized errors that could be clinically significant.

To give more context to these results, they were compared against nmU-Net ViT model's results reported in a recent study on semantic segmentation for the Mn&Ms-2

Metrics	**Our SegFormer3D model**		**nnU-Net ViT + ResBlocks + PReLU activation** (Ayoob et al., 2025)	
	SA	**LA**	**SA**	**LA**
Dice score	0.5568	0.8964	0.9096	0.8930
Hausdorff distance	76.8162	62.4874	66.4372	15.6076

Cardiac MRI dataset. The model demonstrated superior performance in SA views, particularly in Dice score, which suggests some advantages from its adaptive architecture and advanced activations in handling short-axis details. For LA-axis, our model competed well with the nmU-Net ViT model, showing comparable Dice score, but higher Hausdorff distance. These findings highlight the SegFormer3D's strengths in resource-constrained scenarios (lightweight architecture) in comparison to heavier models and identify some possibility to enhance the model, like integrating mechanisms or advanced activations.

7. Future Work & Considerations

Looking at future work to improve the model, the first direction looked into was to combine both SA and LA views together rather than training two independent models. Two practical options that might fit the pipeline used in this project are first a 2.5D approach in which each slice, stack, neighbouring slice and orthogonal view is used as an extra channel in order for the transformer to have more context while maintaining the $256 \times 256 \times 1$ dimensions (Meng et al., 2022). The other option is using the late-fusion method which keeps our SA and LA SegFormer3D encoder but learns a small fusion head that averages their logits before selecting the threshold (Liu et al., 2022). According to (Karimzadeh et al., 2025; Chen et al., 2020), Applying such multi-view schemes have shown improvements in cardiac segmentation by adding missing 3D context and implementing cross view consistency.

Upgrading the training procedure without changing the architecture is another possibility we looked at. The code used in this model contains AdamW and ReduceLROnPlateau so possibly swapping to a cosine annealing schedule with periodic resets in order to result in a more smooth convergence and better generalisation in non pretrained architectures with limited data (Loshchilov and Hutter, 2016). In parallel, adding a deep supervision in the decoder and a simple post processing step to keep the largest connected 3D component per volume which could suppress subtle false positives. These are both potential options to look into which have shown to be strong choices in biomedical segmentation frameworks such as nnU-Net according to (Isensee et al., 2021).

Although our results in the LA model achieved a high Dice score, the HD95 is still relatively large which may point to small and sharp misalignments along the ventricle boundary

(Müller et al., 2022; Jungo et al., 2018; Jungo and Reyes, 2019). Our current loss mainly optimises region overlap rather than the extent that the contour lines up. So to specifically target the contour, we looked at potentially adding a boundary loss that uses the distance transform of the ground truth mask. Another option to tackle this was to include a Hausdorff focused process that penalises the worst boundary deviations. Kervadec et al. (2019) and Foret et al. (2020) suggest that both strategies have proven to reduce edge errors.

Finally, we looked at improving reliability and interpretability. Gal and Ghahramani (2016) discuss using a Monte-Carlo dropout at inference as an uncertainty estimation in order to pick up low-confidence voxels and create uncertainty maps per case. This provided a method to keep dropout layers active and sample K times correlating with segmentation errors. We also proposed and attempted adding a GradCAM visualisation. This would be useful for 2 issues we faced with the boundary uncertainty and the 2 models having significantly different results (SA and LA). This visualisation would allow us to check per slice whether the model is focusing on the right ventricle border or drifting to other regions such as bright cavities or background and allow us to target fixes such as augmentation rather than assuming (Selvaraju et al., 2017; Kim et al., 2023). Due to the M&Ms-2 dataset spanning across multiple scanners and disease types, GradCam could also help in seeing if the model's attention shifts depending on vendor, disease or view thus improving generalisation (Martin-Isla et al., 2023; DeGrave et al., 2021).

8. Conclusions

In this paper, a lightweight SegFormer3D based model is presented for right ventricle segmentation in cardiac MRI tailored for the heterogeneous and multi-view M&Ms-2 dataset. An efficient four stage Transformer encoder with overlapping patch embeddings and an all-MLP decoder was implemented in which the model maintained a small parameter footprint while effectively capturing both fine details and global context. The evaluation on held-out short and long axis tests showed promising results with notably stronger performance for the LA view and moderate accuracy for SA view which highlights the difference in anatomical complexity and image variability. In comparison to other architectures such as nnU-Net ViT, the approach used in this paper provided some competitive overlap metrics with substantially lower computational cost, though higher Hausdorff distances indicates room for improving boundary precision. Multiple avenues for future improvement were identified including multi-view fusion, enhanced loss functions focusing on contour alignment, post-processing to reduce false positives and interpretability improvements such as uncertainty estimation and GradCAM. Overall, these results demonstrate that compact Transformer based models can be effective and efficient for right ventricle segmentation with potential to be further refined for application of diverse cardiac MRI datasets.

References

Mohamed Ayoob, Oshan Nettasinghe, Vithushan Sylvester, Helmini Bowala, and Hamdaan Mohideen. Peering into the heart: A comprehensive exploration of semantic segmentation and explainable AI on the mnms-2 cardiac MRI dataset. 30(1):12–20, 2025. doi: 10.2478/acss-2025-0002. URL https://www.sciendo.com/article/10.2478/acss-2025-0002.

Olivier Bernard, Alain Lalande, Clement Zotti, Frédéric Cervenansky, Xin Yang, Pheng-Ann Heng, Işmail Cetin, Karim Lekadir, Oscar Camara, Montserrat A. Gonzalez Ballester, Gemma Sanroma, Stefan Napel, Stefan Petersen, Georgios Tziritas, Elefttherios Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohé, . . . , and Shubham Jain. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? 37(11): 2514–2525, 2018. doi: 10.1109/TMI.2018.2837502. URL https://ieeexplore.ieee.org/document/8360453. Introduces the ACDC dataset—150 multi-equipment CMRI recordings with reference measurements and classification by two experts :contentReference[oaicite:1]index=1.

Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. Deep learning for cardiac image segmentation: A review. 7:25, 2020. ISSN 2297-055X. doi: 10.3389/fcvm.2020.00025. URL https://doi.org/10.3389/fcvm.2020.00025.

Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. Course corrections for clinical AI. 2 (12):2019–2023, 2021. ISSN 2641-1459. doi: 10.34067/KID.0004152021. URL https://doi.org/10.34067/KID.0004152021.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. abs/2010.01412, 2020. URL https://arxiv.org/abs/2010.01412.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. pages 1050–1059, 2016. URL https://proceedings.mlr.press/v48/gal16.html.

Paul Hamosh and Jay N. Cohn. Left ventricular function in acute myocardial infarction. 50(3):523–533, 1971. ISSN 0021-9738. doi: 10.1172/JCI106521. URL https://www.jci.org/articles/view/106521.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. 18:203–211, 2021. ISSN 1548-7091. doi: 10.1038/s41592-020-01008-z. URL https://doi.org/10.1038/s41592-020-01008-z.

Alain Jungo and Mauricio Reyes. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. 11769:48–56, 2019. doi: 10.1007/978-3-030-32248-9_6. URL https://doi.org/10.1007/978-3-030-32248-9_6.

Alain Jungo, Raphael Meier, Evrim Ermis, Ekin Herrmann, and Mauricio Reyes. Uncertainty-driven sanity check: Application to postoperative brain tumor cavity segmentation. abs/1806.03106, 2018. URL https://arxiv.org/abs/1806.03106.

Mahsa Karimzadeh, Hadi Seyedarabi, Ata Jodeiri, and Reza Afrouzian. Enhanced brain stroke lesion segmentation in MRI using a 2.5D transformer backbone U-Net model. 15 (8):778, 2025. ISSN 2076-3425. doi: 10.3390/brainsci15080778. URL https://www.mdpi.com/2076-3425/15/8/778.

Hoel Kervadec, Jalal Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. pages 285–296, 2019. URL https://proceedings.mlr.press/v102/kervadec19a.html.

Youngjin Kim, Dohyung Kang, Youngsun Mok, Seungho Kwon, and Jonghyun Paik. Bidirectional meta-Kronecker factored optimizer and hausdorff distance loss for few-shot medical image segmentation. 13(1):8088, 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-35164-5. URL https://doi.org/10.1038/s41598-023-35164-5.

Di Liu, Yunhe Gao, Qilong Zhangli, Ligong Han, Xiaoxiao He, Zhaoyang Xia, Song Wen, Qi Chang, Zhennan Yan, Mu Zhou, and Dimitris N Metaxas. TransFusion: Multi-view divergent fusion for medical image segmentation with transformers. 13435:485–495, 2022. ISSN 0302-9743. doi: 10.1007/978-3-031-16443-9_47. URL https://link.springer.com/chapter/10.1007/978-3-031-16443-9_47.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. abs/1608.03983, 2016. URL https://arxiv.org/abs/1608.03983.

Carlos Martin-Isla, Victor M. Campello, Cristian Izquierdo, Kaisar Kushibar, Carla Sendra-Balcells, Polyxeni Gkontra, Alireza Sojoudi, Mitchell J. Fulton, Tewodros Weldebirhan Arega, Kumaradevan Punithakumar, Lei Li, Xiaowu Sun, Yasmina Al Khalil, Di Liu, Sana Jabbar, Sandro Queiros, Francesco Galati, Moona Mazher, Zheyao Gao, ..., and Karim Lekadir. Deep learning segmentation of the right ventricle in cardiac mri: The m&ms challenge. 27(7):3302–3313, 2023. doi: 10.1109/JBHI.2023.3267857. URL https://doi.org/10.1109/JBHI.2023.3267857. Describes the M&Ms-2 (Multi-Disease, Multi-View & Multi-Center Right Ventricular Segmentation in Cardiac MRI) challenge, organized as part of STACOM 2021 at MICCAI, using a dataset of 360 CMR cases from three Spanish hospitals, nine scanners, multiple pathologies :contentReference[oaicite:1]index=1.

Qingjie Meng, Chen Qin, Wenjia Bai, Tianrui Liu, Antonio de Marvao, Declan P. O'Regan, and Daniel Rueckert. Mulvimotion: Shape-aware 3d myocardial motion tracking from multi-view cardiac mri. 41(8):1961–1974, 2022. ISSN 0278-0062. doi: 10.1109/TMI.2022.3154599. URL https://doi.org/10.1109/TMI.2022.3154599.

David Müller, Iñaki Soto-Rey, and Frank Kramer. Towards a guideline for evaluation metrics in medical image segmentation. 15(1):210, 2022. ISSN 1756-0500. doi: 10.1186/s13104-022-06057-9. URL https://doi.org/10.1186/s13104-022-06057-9.

Shehan Perera, Pouyan Navard, and Alper Yilmaz. Segformer3d: an efficient transformer for 3d medical image segmentation. pages 4981–4988, 2024. doi: 10.1109/CVPRW63382.2024.00503. URL https://doi.org/10.1109/CVPRW63382.2024.00503. Workshop version at CVPRW 2024 (DEF-AI-MIA).

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2015*. Springer, 2015. URL https://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/. Preprint (CoRR, arXiv:1505.04597).

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. pages 618–626, 2017. doi: 10.1109/ICCV.2017.74. URL https://doi.org/10.1109/ICCV.2017.74.

Pirbhat Shams, Amandeep Goyal, and Amgad N. Makaryus. Left ventricular ejection fraction, 2025. URL https://www.ncbi.nlm.nih.gov/books/NBK459131/. StatPearls [Internet]; Treasure Island (FL): StatPearls Publishing; Last Update 2025-06-14.

Justin T. Tretter and Andrew N. Redington. The forgotten ventricle?: The left ventricle in right-sided congenital heart disease. 11(3):e007410, 2018. doi: 10.1161/CIRCIMAGING.117.007410. URL https://doi.org/10.1161/CIRCIMAGING.117.007410.

Wen-Yih Isaac Tseng, Mu-Yen Michael Su, and Yung-Hsin Edward Tseng. Introduction to Cardiovascular Magnetic Resonance: Technical principles and clinical applications. 32 (2):129–144, 2016. ISSN 1011-6842. doi: 10.6515/ACS20150616A. URL https://doi.org/10.6515/ACS20150616A.

Christian Wachinger and Polina Golland. Atlas-based under-segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 315–322. Springer, 2014. doi: 10.1007/978-3-319-10404-1_40. URL https://doi.org/10.1007/978-3-319-10404-1_40.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose Manuel Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. abs/2105.15203, 2021. URL https://arxiv.org/abs/2105.15203.