Journal of Imaging Informatics in Medicine
https://doi.org/10.1007/s10278-024-01191-x

CapNet: An Automatic Attention-Based with Mixer Model for Cardiovascular Magnetic Resonance Image Segmentation

Tien Viet Pham[1] · Tu Ngoc Vu[1] · Hoang-Minh-Quang Le[1] · Van-Truong Pham[1] · Thi-Thao Tran[1]

Received: 24 December 2023 / Revised: 21 May 2024 / Accepted: 22 May 2024

Abstract

Deep neural networks have shown excellent performance in medical image segmentation, especially for cardiac images. Transformer-based models, though having advantages over convolutional neural networks due to the ability of long-range dependence learning, still have shortcomings such as having a large number of parameters and and high computational cost. Additionally, for better results, they are often pretrained on a larger data, thus requiring large memory size and increasing resource expenses. In this study, we propose a new lightweight but efficient model, namely CapNet, based on convolutions and mixing modules for cardiac segmentation from magnetic resonance images (MRI) that can be trained from scratch with a small amount of parameters. To handle varying sizes and shapes which often occur in cardiac systolic and diastolic phases, we propose attention modules for pooling, spatial, and channel information. We also propose a novel loss called the Tversky Shape Power Distance function based on the shape dissimilarity between labels and predictions that shows promising performances compared to other losses. Experiments on three public datasets including ACDC benchmark, Sunnybrook data, and MS-CMR challenge are conducted and compared with other state of the arts (SOTA). For binary segmentation, the proposed CapNet obtained the Dice similarity coefficient (DSC) of 94% and 95.93% for respectively the Endocardium and Epicardium regions with Sunnybrook dataset, 94.49% for Endocardium, and 96.82% for Epicardium with the ACDC data. Regarding the multiclass case, the average DSC by CapNet is 93.05% for the ACDC data; and the DSC scores for the MS-CMR are 94.59%, 92.22%, and 93.99% for respectively the bSSFP, T2-SPAIR, and LGE sequences of the MS-CMR. Moreover, the statistical significance analysis tests with p-value < 0.05 compared with transformer-based methods and some CNN-based approaches demonstrated that the CapNet, though having fewer training parameters, is statistically significant. The promising evaluation metrics show comparative results in both Dice and IoU indices compared to SOTA CNN-based and Transformer-based architectures.

Keywords Automatic cardiac segmentation · CapNet · Priority mixer block · Tversky shape power distance function

Introduction

The emergence of deep learning is gradually replacing traditional machine learning models as well as optimization-based algorithms like active contours and level set methods that have been widely applied in MRI image segmentation. In the case of cardiac image segmentation from MRI, deep learning models could help automatically segment the interested organs and desired areas like left ventricle, right ventricle, myocardium, and myocardial infarction areas [1–3]. The segmentation in cardiac MRI is necessary for further analysis and diagnosis of heart failures and many cardiac applications including scoring coronary artery calcium, plaque analysis, left ventricular analysis, diagnosing myocardial infarction, prognosticating coronary artery disease, arterial disease, evaluating cardiac function, and diagnosing and prognosticating heart diseases.

With its ability to automatically learn features, deep learning allows models to learn complex features from data without the need to explicitly define and extract specific features beforehand. This eliminates or reduces the dependence on human intervention in feature design and helps the model automatically discover complex patterns and rules in images, thus bringing automatic segmentation performance

✉ Thi-Thao Tran
thao.tranthi@hust.edu.vn

[1] Department of Automation Engineering, School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Hanoi, Vietnam

Published online: 09 July 2024

close to manual segmentation [4–7]. Deep learning can learn from millions or even billions of cardiac MRI images, enhancing prediction and classification capabilities through models such as convolutional neural networks (CNNs) or autoencoders. However, early-stage deep learning models require a large amount of data and high computational costs and only achieve average performance. Patch-based methods in CNNs involve dividing the input image into overlapping patches and processing each patch independently. While this method has the advantage of capturing local features and gathering spatial information, it also has a major drawback: redundancy in the inference process. The fully convolutional neural network (FCN) version addresses some issues in the pioneering work of Long et al. [8]. The FCN version improves upon CNNs by enabling the processing of arbitrary input images through an encoder-decoder structure, utilizing the sampling architecture facilitated by the transformation of convolutional kernels. Tran [9] has contributed to the segmentation of the left and right ventricles through the application of the FCN. However, the FCN has shown limitations in capturing detailed contextual information in images for accurate segmentation. To achieve more accurate segmentation, Ronneberger et al. [10] proposed the U-Net model, a famous variant of the FCN network. U-Net utilizes skip connections to avoid the loss of contextual information that FCN may suffer from. The emergence of attention blocks gradually replaced skip connections as they enhance the segmentation capability by focusing heavily on important regions of the image. For example, Attention U-Net, proposed by Ozan et al. [11], builds upon the U-Net model by incorporating an attention gate mechanism to extract coarse-scale features used for gating in skip connections, distinguishing irrelevant responses and noise. Additionally, self-attention mechanism combined with position encoding transforms the relative positional information of elements in the computation sequence, creating a transformer network architecture without the need for convolutional layers and skip connections.

Though having shown superior performance in computer vision tasks, the direct application of transformers in medical image segmentation still suffers from some shortcomings. Chen et al. [12] interpreted that transformers process input as 1D sequences and only focus on modeling global context at all stages, resulting in low-resolution features and lacking detailed localized information. This information cannot be effectively recovered by directly upsampling to full resolution, leading to coarse segmentation results. On the other hand, CNN architectures (e.g., U-Net) provide a method to extract low-level visual signals that can effectively handle such small spatial details. Therefore, Chen et al. proposed the TransUnet model that combines U-Net and transformer to leverage the benefits offered by both architectures. However, recognizing the strong dominance of CNNs in medical image segmentation, Cao et al. [13] proposed a model that utilizes a pure Swin-Transformer architecture, inspired by the U-Net-like encoder-decoder framework. In this approach, prior to entering the Swin-Transformer, the input images are divided into non-overlapping patches. After encoding, the patches are decoded through a combination of patch merging layers and the Swin-Transformer. When performing machine learning tasks, we have found that training transformer models, such as TransUnet, Swin-Unet, and MISSFormer [14], incurs significant computational costs. These models are pretrained with large memory, which can pose challenges in terms of updating and expanding them. If there are changes in the training data or new task requirements, updating and expanding pretrained models may require retraining from scratch or result in substantial time and resource expenses. When learning from small amounts of data, using transformer models with large memory becomes unnecessary and overly expensive. They may not provide substantial benefits that justify the required resources.

Motivated by the above concerns regarding the image segmentation architecture for cardiac MRI images, in the current study, we propose a new model along with a novel loss for training the neural network. In particular, the proposed model, namely CapNet, is a harmonization of attention blocks that processes local information in clusters, highlighting the global information feedback. Furthermore, the model helps minimize computational costs and model parameters while ensuring a balance in learning data processing.

Related Work

Deep learning has emerged as the primary trend in addressing healthcare automation problems in recent years. With the strong development of deep learning over the past decade, the methods utilizing deep learning in Cardiac MRI image segmentation have undergone significant diversity and transformation. There are two main approaches to implementing this problem using deep learning. The first approach involves feeding the entire 3D volume of cardiac MR images into a deep learning model [15–17], which can be challenging due to the large volume and computational time required. Therefore, the second approach, using 2D slices of the 3D volume in the deep learning model, is the approach we adopt in this study.

In the past, there have been many studies following this approach, which we categorize into two main types: CNN-based and Transformer-based. The CNN-based approach is the most common direction, with numerous studies adopting this method such as [18, 19], and [20]. Cui et al. [18] utilized Attention U-Net along with a pyramid input image to retain maximum spatial information. Chen et al. [19] employed U-Net with dropout normalization layers after concatenation to reduce noise in the U-Net Decoder. Wang et al. [20]

used U-Net with skip connections comprising multiple layers to connect low-level and high-level features. Overall, CNN-based neural networks are proficient in extracting both local and global information by employing convolutional operations with a strong inductive bias, allowing them to acquire robust representations. However, the use of multiple convolutional operations can sometimes result in inadequate handling of long-range dependencies and loss of spatial information. This issue is contradictory because segmentation tasks typically necessitate substantial spatial information. The remaining approach is Transformer-based, which was first introduced in 2020 [21], but there have been quite a few studies based on transformer-based methods used in cardiac segmentation [14, 22]. Huang et al. [14] introduced a fully transformer architecture that supports local feature context. Li et al. [22] proposed a transformer architecture, parameterized in a low-complexity form using Axial Attention [23]. In fact, transformer-based methods may yield suboptimal results when trained on insufficiently large datasets, particularly when dealing with medical datasets that pose additional challenges. In scenarios with limited data, transformer-based models often rely on pretrained weights to achieve desired outcomes effectively due to the lack of inductive bias. Furthermore, the high parameter count and complexity of transformer-based models can pose challenges during deployment. Consequently, hybrid architectures that integrate CNN-based and transformer-based approaches are garnering increasing attention from researchers. These architectures leverage the strengths of both methods to address their respective limitations, such as [12, 24, 25].

The methods discussed have demonstrated very good effectiveness in the task of cardiac MRI image segmentation, showcasing the efficacy of the approach involving dividing 3D volumes into 2D slices. However, these methods, whether CNN-based or transformer-based, entail a large number of parameters and have been applied to small datasets. It would be more effective to have a model with fewer parameters that matches the size of the data. Lightweight models have been developed to address these challenges [26, 27]. To our knowledge, there are currently few lightweight models specifically designed for cardiac MRI image segmentation tasks, and it would be very promising to have a model that strikes a balance between parameters and performance for this task.

Materials

Depthwise Separable Convolution

Depthwise separable convolutions have been proposed by Chollet in Xception model [28] for image classification tasks that reduce the computational cost of the convolution while maintaining good performance. While normal convolution uses a single filter spanning across multiple channels, depthwise separable convolution splits the computation into two steps, depthwise convolution (DW) and pointwise convolution (PW). Depthwise convolution involves applying a convolutional filter to each input channel, allowing the model to capture channel-specific information. Pointwise convolution [29] utilizes a 1×1 kernel applied individually to each pixel. The intuitive idea is that the kernel size is small, allowing it to capture fine-grained details in the image. In other words, in depthwise separable convolution, depthwise convolution step applies a separate kernel to each input channel, and pointwise convolution then combines the resulting feature maps using 1×1 convolution.

Priority Attention

Attention mechanisms in deep neural networks help the network focus on important information within domains such as channel and spatial. Recently, inspired from the greedy algorithm [30], Le et al. [31] proposed a new attention mechanism called priority attention comprising two variants: Priority Channel Attention (PCA) and Priority Spatial Attention (PSA). Both PCA and PSA employ attention mechanisms based on the variations of feature maps after convolution operations. The PCA architecture uses depthwise convolution to select features for each channel, then a channel-specific feature vector is used to compare channels that change a lot before passing through softmax to produce the attention vector. With this option, the featured output is filtered by channels and does not require additional parameters. Similar to PCA, PSA is an architecture based on the deviation in each pixel to produce an attention matrix based on the feature output of pointwise convolution. PSA carries spatial feature information and selects features to produce an effective feature set. Both PCA and PSA were used for the first time in classification problems. However, in the segmentation problem, the features obtained through attention are very important, which can increase efficiency. Therefore, applying PCA and PSA to the segmentation problem can help improve performance without increasing model parameters.

Pooling Attention Based on MLP

During the process of image dimension reduction through computation, information loss may occur as a consequence of the pooling function. However, on the other hand, applying an attention mechanism to the dimension reduction process can help retain essential information from the input. In the CPA-Unet [32], a pooling attention mechanism is utilized, which shares a structural resemblance with SE (Squeeze-and-Excitation) and ECA (Efficient Channel Attention) techniques. The module is split into two branches, incorporating a combination of average-pooling, max pooling, and MLP (multilayer perceptron) layers. This approach allows for the retention of essential information without significantly increasing the number of parameters. Therefore, incorporating PA (Pooling Attention) blocks into the encoder would be appropriate for improving the model's output.

Methodology

The Proposed Model

Our Proposed CapNet Model

Observing the surveys for cardiac segmentation [1, 2], we propose a new model for cardiac MRI image segmentation. The proposed model has the encoder-decoder symmetric architecture as the U-Net [10]. In our proposed architecture, CapNet is shown in Fig. 1. In the CapNet encoder, we construct a block consisting of Conv-block and Pooling Attention (PA). The input with the shape (B, C, H, W) of cardiac segmentation data is first passed through the Conv-block that undergoes convolution with channel (C) equal to 1 and a kernel size of 3 to extract local features, followed by batch normalization (Batch Norm), ReLU activation, then a convolution operation to further learn the selected features.

After the Conv-block, the features are fed into the Pooling Attention based on MLP. Through this attention mechanism, we aim to extract information from the feature maps while expanding the receptive field to reduce information loss when reducing the dimension through max pooling. Four Conv-blocks with Pooling Attention based on MLP are established, with the dimensions decreasing by $[\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}]$ while the filter

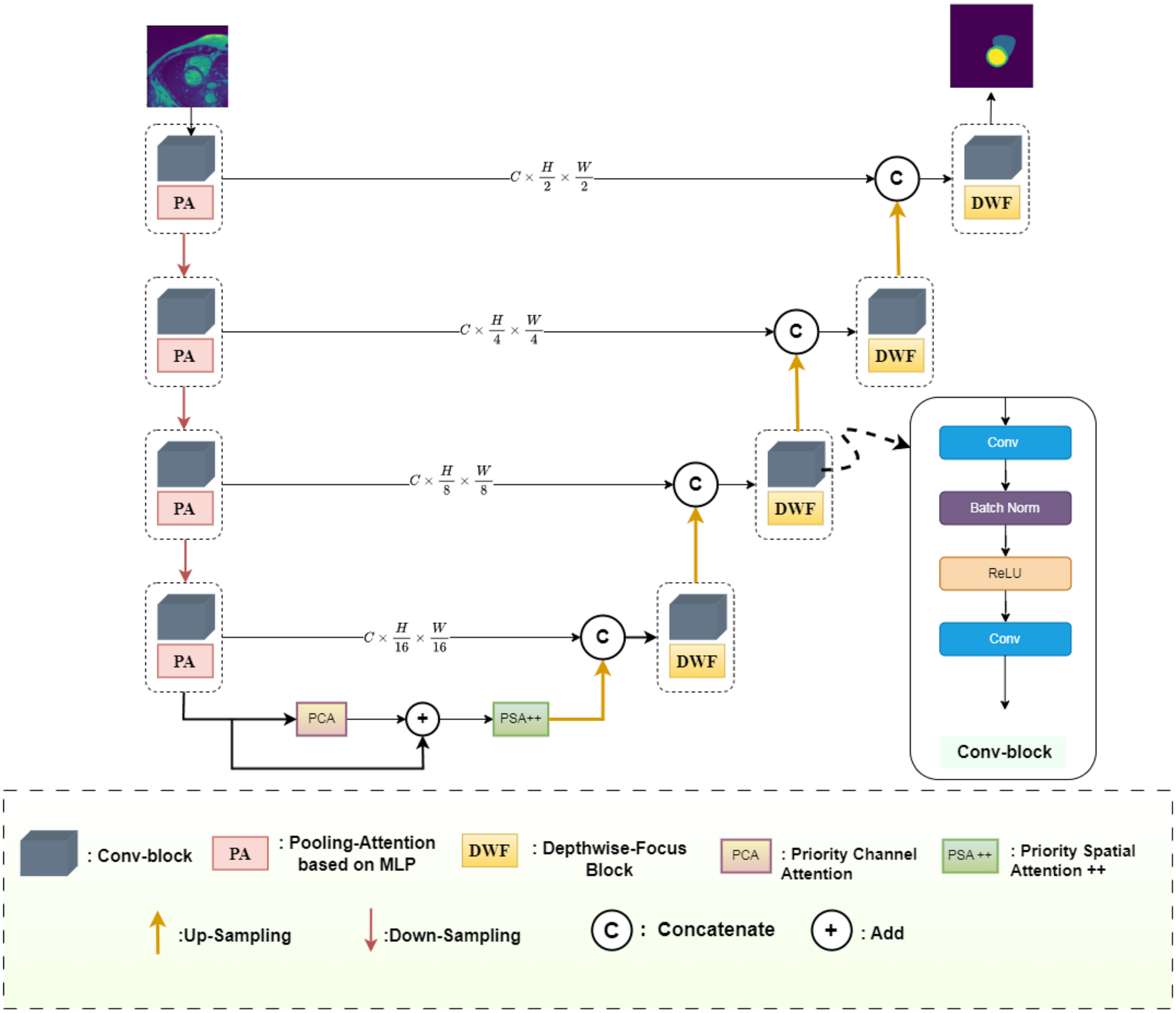

PA
DWF
$C \times \frac{H}{2} \times \frac{W}{2}$
$C \times \frac{H}{4} \times \frac{W}{4}$
$C \times \frac{H}{8} \times \frac{W}{8}$
$C \times \frac{H}{16} \times \frac{W}{16}$
C
+
PCA
PSA++
Conv
Batch Norm
ReLU
Conv
Conv-block
: Conv-block
PA : Pooling-Attention based on MLP
DWF : Depthwise-Focus Block
PCA : Priority Channel Attention
PSA ++ : Priority Spatial Attention ++
:Up-Sampling
:Down-Sampling
C : Concatenate
+ : Add

Fig. 1 Our proposed CapNet model

sizes of the encoding blocks increase to [16, 32, 64, 128] respectively. At the bottom of the architecture, also called bottleneck, the features are not directly transferred from the encoder to the decoder as in U-Net. Instead, they will traverse a bridge similar to the PASPP module in [33], or Convmixer [34] architecture for a bottleneck in [35]. In this study, we propose a new module for the bottleneck, named Priority Mixer Block that shows superior performance compared to commonly used ASPP, PASPP bottleneck, and Convmixer module. Detail on the proposed Priority Mixer Block will be described in the next subsection.

For the decoder, we gradually increase the upsampling blocks to restore the channels to their original state to generate the desired predictions. The blocks following the bridge include Conv-block and the proposed Depthwise-Focus (DWF) Block based on Wide-Focus [36]. Between these blocks, additional Upsampling layers are added to restore the dimensions, and the filters decrease inversely compared to the encoding. The skip connections from the encoder are concatenated with the corresponding blocks in the decoder.

Our Proposed Priority Mixer Block

In the current study, inspired by PCA and PSA modules in [31] that have shown effectiveness for fish classification problems, we adapt these architectures for the model of cardiac MRI segmentation. To this end, we replace depthwise separable convolution with the Priority Channel Attention block, where the Attention block has an identity branch to avoid the vanishing problem. The PCA block selectively emphasizes informative channels within feature maps, allowing for the representation of important patterns. Then, instead of using pointwise convolution in the ConvMixer, we replace it with PSA++ which is an upgrade proposed by us based on PSA that extends MLP's receptive field through height and width. The detailed description of the proposed Priority Mixer block is given in Fig. 2.

Similar to the PSA, in the proposed PSA++, we extend the attention mechanism beyond channels to encompass the spatial dimensions, H and W. This means that instead of focusing solely on channel-wise relationships, we also consider the relationships between pixels in the height and width dimensions. By incorporating spatial attention, we can better capture spatial dependencies and improve the representation power of the model. Beginning with the specific feature $x^{(B,C,H,W)}$, we diverge into three pathways, each processing pixel information from the perspectives of channel, height, and width. This allows us to capture different aspects of the input feature and extract relevant information for each dimension. By treating each dimension separately, we can better understand the relationships and patterns within the data. Reshape operations are employed for the transition from channel to height or width. This reshaping allows us to perform computations specific to each dimension. For example, in the height pathway, we reshape the input feature (B, C, H, W) to have dimensions (B, H, C, W), effectively treating each pixel along the height dimension as a separate entity. The subsequent steps follow a similar methodology as PSA. In each pathway, we compute the average across all channels for the corresponding dimension (channel, height, or width) within the reshaped blocks. This averaging operation helps to capture the overall characteristics of each dimension and summarize the information across channels. This averaging can be seen as treating the average across channels as if it were the height or width dimension. By doing so, we can effectively reduce the dimensionality of the feature and focus on the most important aspects within each pathway. Following that, all three pathways undergo a pointwise convolution (PW) operation, similar to the step performed in PCA, but with pointwise convolution (PW) instead of depthwise convolution (DW). Pointwise convolution is applied to each pathway to further process the features and capture higher-level representations. This operation helps combine information from different dimensions and channels, leading to a more comprehensive understanding of the data. Subsequently, we calculate the average across all channels for each pathway, resulting in $S'c^{(B,H,W)}$, $S'h^{(B,C,W)}$, and $S'w^{(B,H,C)}$.

These output features represent the enhanced spatial attention within each pathway. By calculating the average across channels, we obtain a summary of the attention weights for each dimension. To further refine the spatial attention, we probabilistically normalize the attention weights within each pathway. This involves subtracting the corresponding difference tensors: $(S'c - Sc)$, $(S'h - Sh)$, and $(S'w - Sw)$, which capture the changes in attention after the spatial processing. By subtracting the original attention weights, we can focus on the changes and identify the areas that have received more or less attention. Finally, we apply the softmax function to these tensors, which scales the values to range between 0 and 1 and ensures that they sum up to 1. This normalization step allows us to interpret the values as probabilities and obtain a distribution of attention weights for each dimension. These attention weights can then be used to weight the features or guide subsequent computations in the neural network model. Additionally, to maintain stability and avoid excessive fluctuations during training, the spatial attention coefficients are computed using the following formulas:

$$F_s'c = \sigma[S'c \times (1 + softmax2d(S'c - Sc)] \tag{1}$$

$$F_s'h = \sigma[S'h \times (1 + softmax2d(S'h - Sh)] \tag{2}$$

$$F_s'w = \sigma[S'w \times (1 + softmax2d(S'w - Sw)] \tag{3}$$

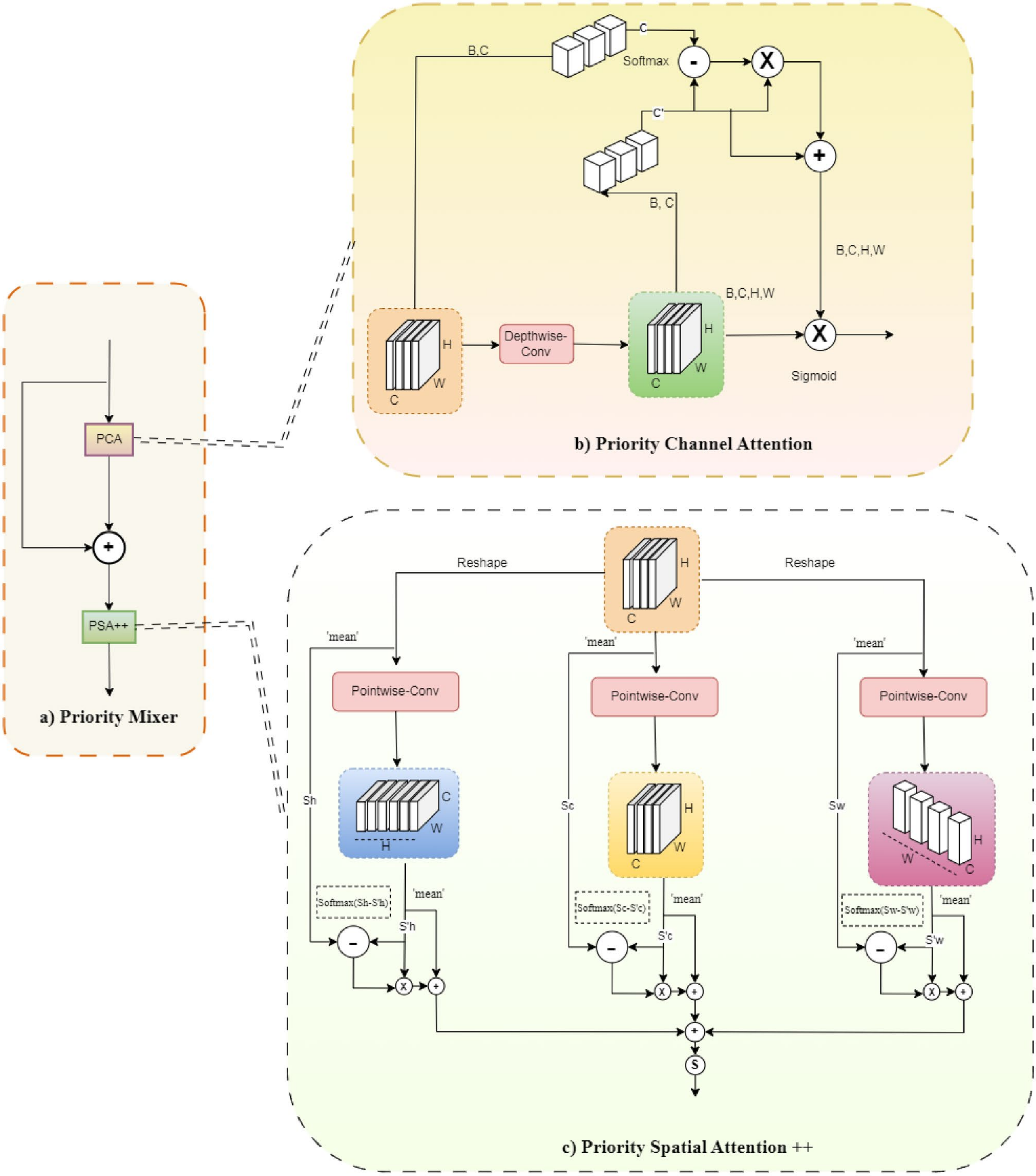


Fig. 2 Our Priority Mixer Block

$$F_s' = F_s'c + F_s'h + F_s'w \tag{4}$$

$$x = x \cdot F_s' \tag{5}$$

Our Proposed Depthwise-Focus (DWF) Block

From experiments and observations, we found out that in the encoder-decoder architecture, the decoder achieves the best performance when it simultaneously decodes local information and global context and then recovers details from the spatial source, and the previously encoded feature maps. Thus in this work, we propose the Depthwise-Focus Block, as shown in Fig. 3 that takes into consideration the above findings. In particular, we connect a Depthwise Separable Convolution layer right after the Conv-block to generate a convolutional filter for each input channel, allowing decoding of global information at the output of each channel.

To enhance the decoding process and focus on desired factors, we use additional depthwise convolutions with $kernelsize = 1 \times k$ and $kernelsize = k \times 1$. Compared to the standard convolution, it only requires $2k$ parameters instead of k^2. When $k \geq 3, 2k < k^2$, leading to a significant reduction in computational cost as k increases. In our experiments, we found that $k = 7$ yields the best performance. We also experimented with linearly increasing dilation rates, where these three convolutions are parallelly added together. We experimented with different dilation levels combined to emphasize flexibility in local detail depending on the linear dilation rate, avoiding the inefficiency of the model's accuracy after the network's learning process saturates.

Moreover, this direct emphasis helps stabilize the architecture by accurately learning focused pixels from the blocks in the encoder and the Priority Mixer bridge. Additionally, when parallelly adding the depthwise separable convolution blocks with the standard convolution as mentioned above, it improves the network's ability to replicate global maps in deeper layers. We incorporate this into the Wide-Focus module architecture introduced in [36], instead of using standard convolution with different dilations. We replace them with depthwise separable convolution, following Fig. 3, where the parallel dilation order is 1, 2, 3 for the best results. However, to achieve optimal performance and avoid information loss, we added a Residual Block parallelly with element-wise addition. This block mitigates the vanishing gradient problem from the direct connection layer from the encoder, which is connected to the decoding block using the same filter. We observed a significant improvement in results by integrating the proposed block into the architecture.

The Proposed Loss Function

The Tversky Shape Power Distance (TSPD) Loss Function

Along with the advancements in deep learning models, the loss functions have gradually evolved to shorten convergence time and capture the regions where the model performs best [37–40]. In this study, inspired by the shape distance

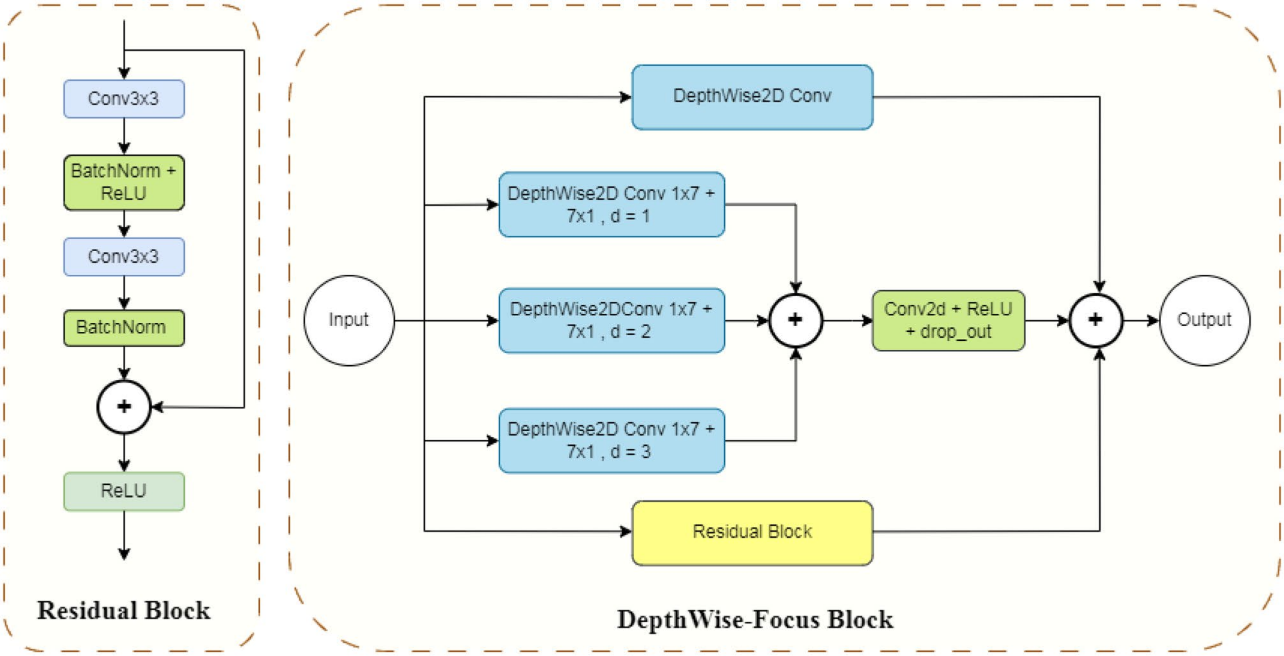


Fig. 3 Our DepthWise-Focus (DWF) Block

described by Pham et al. in [41], we propose a modified loss for training the network. Our proposed shape distance term measures the dissimilarity between the predicted mask $\hat{y}$ with $\hat{y} \in [0, 1]$ and ground truth y with $y \in \{0, 1\}$. Denote N be the number of pixels of the maps. We increase the shape distance distance by a power of m in the predicted mask. The modified shape distance is rewritten as follows:

$$L_d(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} (y_i(1 - \hat{y}_i{}^m) + \hat{y}_i{}^m(1 - y_i)) \tag{6}$$

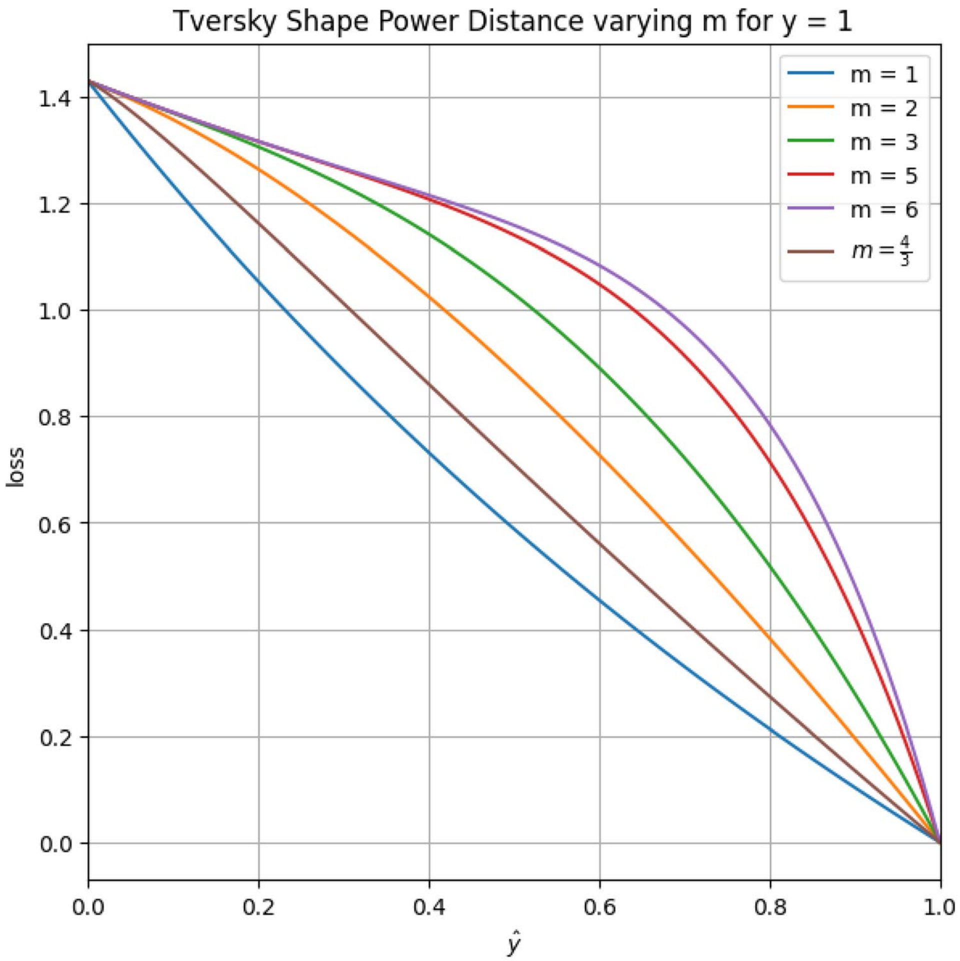
Instead of directly applying the weight of $\frac{1}{N}$ to L_d as described earlier, we reduce this weight by scaling it with the ratio of the sum of the denominators of the Tversky loss function [38]. Accordingly, we propose a loss function in the following form:

$$L_d(y, \hat{y}) = \frac{\sum_{i=1}^{N} (y_i(1 - \hat{y}_i{}^m) + \hat{y}_i{}^m(1 - y_i))}{\sum_{i=1}^{N} (y_i\hat{y}_i + \alpha(1 - y_i)\hat{y}_i + \beta(1 - \hat{y}_i)y_i)} \tag{7}$$

where α and β are hyperparameters denoted in the Tversky loss [38]. True positive (TP) is $y_i\hat{y}_i$, false postive (FP) is $(1 - y_i)\hat{y}_i$, and false negative (FN) is $(1 - \hat{y}_i)y_i$. In our simplified loss function, when $m = 1, \alpha = \beta = 1$, Eq. 7 becomes the Jaccard/IoU loss. Choosing α or β depends on the purpose of adjusting the False Positive rate or False Negative rate to be compatible with the characteristics of the datasets. Based on experiments, we observed that α and β are two parameters that follow the proportion $\alpha + \beta = 1$. We experimented with a ratio of $\alpha : \beta = 3 : 7$, which yielded good results. To find the optimal range of values for m, we assume cases where $\alpha : \beta = 3 : 7$, and the ground truth $y = 1$. To simplify, $\hat{y}$ will be gradually increased within the range [0,1]. The resulting graph is shown in Fig. 4:

In the picture shown in Fig. 4, we observed that when $m < 1$, the function focuses on accurately predicting low-density pixels that are misclassified. Testing with $m \geq 1$ slope values yields more stable results and better performance. To achieve an appropriate value for m, we described the varying value in the graph in Fig. 4. When m belongs to the range $[\frac{4}{3}, 3]$, we obtained better results. Among them, $m = 2$ is the best value that we used throughout the training process. It is worth noting

Fig. 4 Incidence of parameter *m* in Tversky Shape Power Distance loss



that the $\hat{y}$ in Eq. 6 is close to the degree of membership function in fuzzy active contour models [42]. In this formulation, the power m plays the role of a weighting coefficient on the fuzzy membership and is commonly set equal to 2 in the fuzzy logic field.

Formulation with Tversky Shape Power Distance (TSPD) Loss Function

Based on the extended description of the multiclass level set proposed by Trinh et al. in [35] and Kim and Ye in [43], we replaced the ground truth y with $\mathbf{Y}$, which is the input one-hot vector. $\mathbf{Y}$ is composed of multiple channels, where each channel contains a binary segmentation mask. These masks are used to determine the spatial domain of class k within the set 1, 2, 3..., N. Each channel in $\mathbf{Y}$ represents a specific class and distinguishes the regions assigned to that class with binary values. $\mathbf{P}$ is denoted as the output softmax of the network $P(\phi)$. Formulation with Tversky Shape Power Distance (TSPD) loss function described as follows:

$$L_d(\mathbf{Y},\mathbf{P}) = \frac{\sum_{k=1}^{N}\sum_{x\in\phi}(\mathbf{Y_{kx}}(1-\mathbf{P(\phi)_{kx}}^{m})+\mathbf{P(\phi)_{kx}}^{m}(1-\mathbf{Y_{kx}}))}{\sum_{k=1}^{N}\sum_{x\in\phi}(\mathbf{Y_{kx}}\mathbf{P(\phi)_{kx}}+\alpha(1-\mathbf{Y_{kx}})\mathbf{P(\phi)_{kx}}+\beta(1-\mathbf{P(\phi)_{kx}})\mathbf{Y_{kx}})} \tag{8}$$

The loss function we propose will gradually approach 0 as the output $P(\phi)$ of the architecture approaches the closest match to $\mathbf{Y}$. If the predicted output deviates significantly, the exponential function m that we incorporate will decrease, thereby increasing the number of correctly classified pixels in the ground truth.

Evaluation Metrics

In image segmentation, the two most commonly used evaluation metrics are the Dice similarity coefficient (DSC) and the intersection over union index (IoU) also known as the Jaccard Index. The DSC statistically measures the similarity between the segmentation map and the ground truths, and the IoU gauges the similarity and diversity of sample pixel sets.

The statistical significance analysis of a segmentation model compared with other models is demonstrated by the p-value. The assumed statistical significance level of p-value was equal to 0.05. The determination of the model's statistical significance was based on the p-value using the non-parametric Wilcoxon signed-rank test [44], which is employed for hypothesis testing. Particularly, in the current study, the segmentation scores including DSC and IoU of different models compared with the proposed model are evaluated by computing the p-value between the two models.

Experiment

Datasets

The Sunnybrook Dataset

The Sunnybrook dataset [6] provided by the Sunnybrook Health Sciences Centre, Toronto, Canada, proposed in the MICCAI 2009 LV segmentation challenge. The dataset includes cardiac cine-MRI images (1.6 GB) in the DICOM format collected from 45 patients. The patients are from a diverse range of cardiac conditions like healthy hearts, hypertrophy, heart failure with infarction, and heart failure without infarction. The data also includes manual segmentation contours by Perry Radau from the Sunnybrook Health Science Centre that includes the endocardium and epicardium from slices in various phases including the end diastolic (ED) and end systolic (ES). All the images were obtained during 10–15 s breath-holds with a temporal resolution of 20 cardiac phases over the heart cycle and scanned from the ED phase. The endocardium and epicardium images are split into 3 parts with a ratio of 70:15:15 for respectively training, validation, and testing. The data are resized to the resolution of 256 × 256 pixels.

The MRI Cardiac ACDC Dataset

The ACDC dataset [45] was generated using real clinical exams conducted at the University Hospital of Dijon. To ensure privacy, all acquired data underwent a thorough anonymization process and were handled in compliance with the regulations established by the local ethical committee of the Hospital of Dijon in France. The ACDC dataset consists of 100 patient 4D cine CMR scans. Each scan includes segmentation labels for the left ventricle (LV), the myocardium (Myo), and the right ventricle (RV) during the end-systolic and end-diastolic phases. The dataset was divided into three sets: a training set, a validation set, and a testing set, with a split ratio of 70:10:20. All images have been resized to 128 × 128 pixels in this study.

The MS-CMRSeg 2019 Dataset

The data of MS-CMRSeg 2019 (or MS-CMR) [46, 47] contained 45 multi-sequence CMRs, provided by the organizers of the Multi-sequence Cardiac MR Segmentation Challenge. The MS-CMRSeg 2019 dataset aims to capture specific aspects of cardiac imaging using different CMR sequences. Magnetic resonance imaging (MRI) is widely used to gather both anatomical and functional details of the heart. To visualize acute injuries and ischemic regions, the T2-SPAIR CMR sequence is utilized. Meanwhile, the bSSFP cine CMR sequence captures cardiac motions and establishes distinct boundaries. For visualizing myocardial infarction, the LGE CMR sequence is specifically designed. The T2-weighted, black blood spectral presaturation attenuated inversion-recovery (SPAIR) sequence generally includes a limited number of slices. For example, out of the 45 cases in the dataset, 13 cases consist of only three slices, while the remaining cases contain five (13 subjects), six (8 subjects), or seven (one subject) slices. On the other hand, the bSSFP cine CMR sequence is a balanced steady-state, free precision cine sequence that typically consists of 8 to 12 contiguous slices. These slices cover the ventricles entirely, from the apex to the basal plane of the mitral valve. Some cases may include additional slices beyond the ventricles. The images and masks of all sequences are resized to the resolution of 256×256 pixels. The train-to-valid-to-test ratio for this data is 70:10:20.

Implementation Details

We have performed the proposed network, CapNet with our proposed customized Tversky Shape Power Distance loss to segment MRI images. Our model is trained on a workstation with NVIDIA Tesla P100 16GB GPU. The minimization is performed on several epochs using AdamW optimizer [48] with an original learning rate of 1e-3. Every 5 epochs, the learning rate is divided by 2, before reaching 0.00001, and is then constantly kept through the remainder training period with 200 epochs for all three datasets. The datasets are also augmented by various techniques, such as rotation, flipping, and scaling, to further increase the diversity of the training data.

Experimental Results

Model Visualization

Deep learning models deliver unprecedented breakthrough results in computer vision tasks. Although these models exhibit outstanding performance, their complexity renders them impossible to decompose into smaller parts for interpretation. When problems arise, we can only rely on guesswork since we cannot pinpoint the exact cause. Recently, the interpretation of deep learning models has become possible using gradient-based methods from the target layer to the component neurons. This approach provides a more intuitive understanding of how deep learning models function and helps identify important neurons in deep learning networks. In this study, we interpret our model by visualizing important layers using the Grad-CAM [49] method.

In the illustration, Fig. 5 represents the output formation of all three categories RV, Myo, and LV by the proposed CapNet model. The heatmap overlaid on the diagram depicts the concentration of component layers about the target segmentation layer. It can be observed that the Encoder block diversifies feature maps, but the initial layers only extract raw information and have minimal impact on the target segmentation layer. In deeper layers, regions tend to be more pronounced and diverse, extracting higher-level information, resulting in more diverse feature maps.

Information at the end of the Encoder block is passed through the proposed Bottleneck block - Priority Mixer. When this block is introduced, information is condensed and focused using Channel Attention and Spatial Attention mechanisms. Subsequently, the feature map is passed through the Decoder, tasked with upsampling to generate the segmentation image. It is noticeable that information introduces noise upon entering the decoder, attributed to the skip connection from the encoder and additional convolutional layers within it. However, over successive layers, adjustments are made to gradually refine and synthesize information for accurate segmentation.

The layers near the end concentrate precisely on the target segmentation layer. In these layers, updates are made to the weights through gradient descent closest to the loss function. Adjusting weights or synthesizing information from preceding layers is more favorable in these layers.

To clarify the function of the Priority Mixer block, in Fig. 6, we use some samples to illustrate the internal process of this block. Before entering PCA, the feature maps are diverse and not specifically focused on any region. After passing through the PCA block, its sole task is to synthesize important features per channel, partially condensing the feature maps, but not yet clearly defined. After going through PSA++, the features are further contracted, focusing on a visible region. Additionally, we provide another example of the importance of the Priority Mixer in Fig. 7. It helps the feature maps to be more focused on the target segmentation class.

In Fig. 8, we visualized the importance of the first Depthwise-Focus Block in our proposed. The Depthwise-Focus Block consists of three branches: depthwise convolution block, wide block, and residual block. The wide block utilizes depthwise convolutions both vertically and horizontally, while employing different expansion factors to synthesize diverse essential information. Meanwhile, the depthwise convolution branch emphasizes local features, and the residual block helps retain certain characteristics. In a sample as shown, the wide block has effectively fulfilled its role. The remaining branches not only contribute to preserving information but also have the ability to diversify feature maps. Although they may introduce some noise and might not be optimal for the final segmentation purpose, their use in combination with multiple layers in the model provides better directionality and effectively exploits their function.

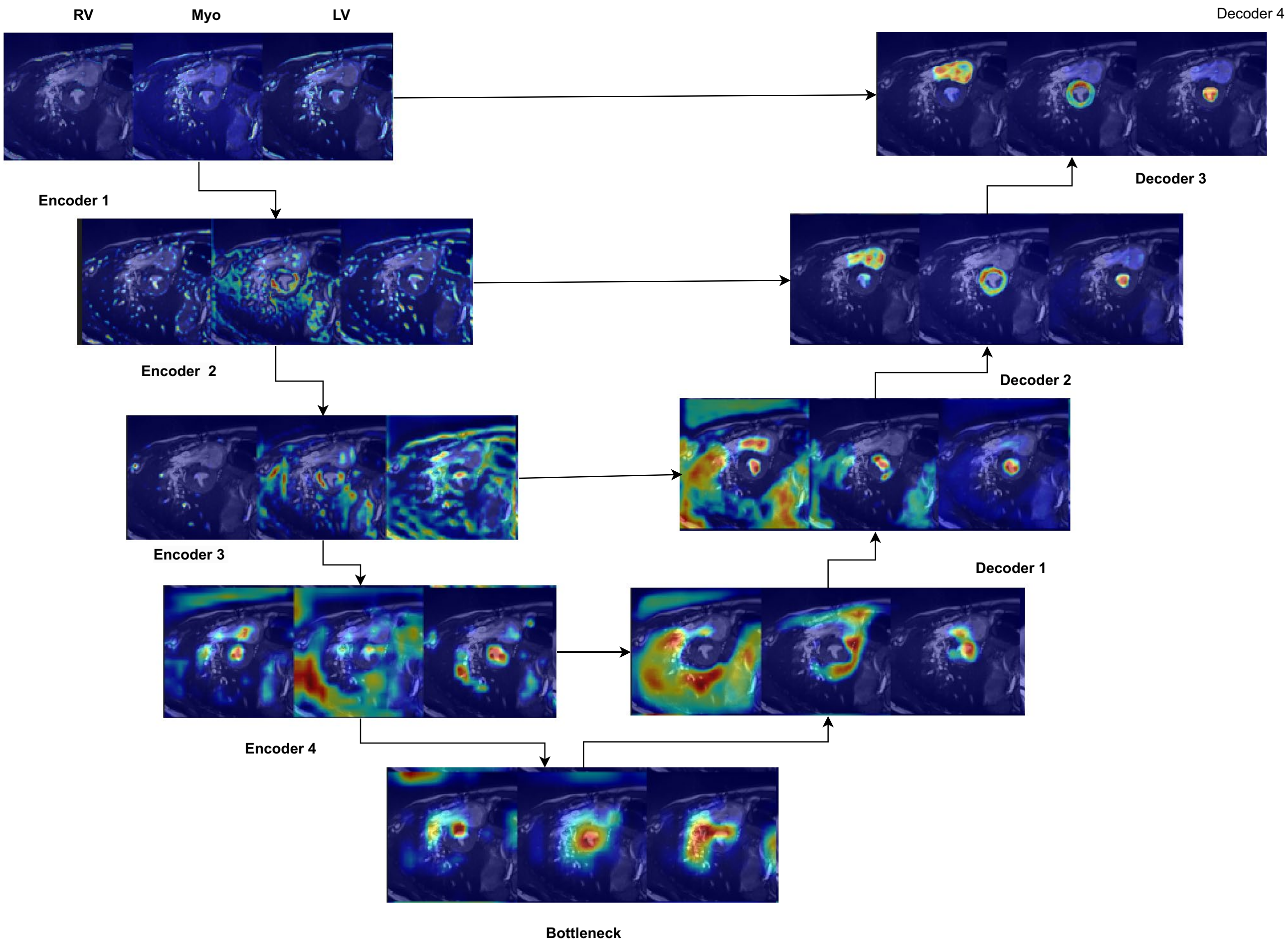


Fig. 5 Visualization of the sequential process through the proposed model to influence the segmentation output using Grad-CAM

Evaluation on the Sunnybrook Dataset

We first show the performance of the proposed model against different models on the Sunnybrook dataset. The results from Fig. 9 show that the segmentation by the proposed model is in better agreement with those by other models. For quantitative assessment, we provided the DSC and IoU scores by comparative models on the Sunnybrook data in Table 1. In addition, the statistical significance analysis by the p-value is also given in this table. The tests are made to check whether there is a difference between the segmentation quality measures by different models and our proposed model.

Table 1 shows that the proposed CapNet exhibits a statistically significant improvement in both DSC and IoU compared to TransUNet (p = 0.0336 for DSC, p = 0.0261 for IoU), Swin-Unet (p =8.727×10^{-4} for DSC, p = 3.503×10^{-4} for IoU), Res-Unet ($p = 3.710 \times 10^{-4}$ for DSC, p =1.019×10^{-4} for IoU), DS-TransUnet ($p = 1.03110^{-3}$ for DSC, p = 4.612×10^{-4} for IoU), U-Net (p = 5.776×10^{-4} for DSC, p = 2.296×10^{-4} for IoU), Attention-Unet (p = 0.0291 for DSC, p = 0.0187 for IoU), and SegNet (p = 0.0125 for DSC, p =9.536×10^{-3} for IoU) for the endocardium. For the epicardium, the statistical values are significant compared to most methods, except MSU-Net, U-Net, and U-Net++. Nevertheless, considering the number of training parameters, as shown in the second column of Table 1, the proposed CapNet has significantly fewer parameters compared to these models.

For better visualization, the boxplots showing the IoU and DSC scores by those models are also given in Fig. 10. As can be seen from this figure, with the smallest number of parameters, the proposed model has the highest values for both median and maximal values of IoU and DSC.

Evaluation on the MRI Cardiac ACDC Dataset

In the first experiment for segmentation on the ACDC data, we show the performance of left ventricle segmentation including endocardium and epicardium by the proposed model with compared models in Fig. 11. From the representative segmentation in this figure, we can observe that the predicted masks by our model are in best agreement with

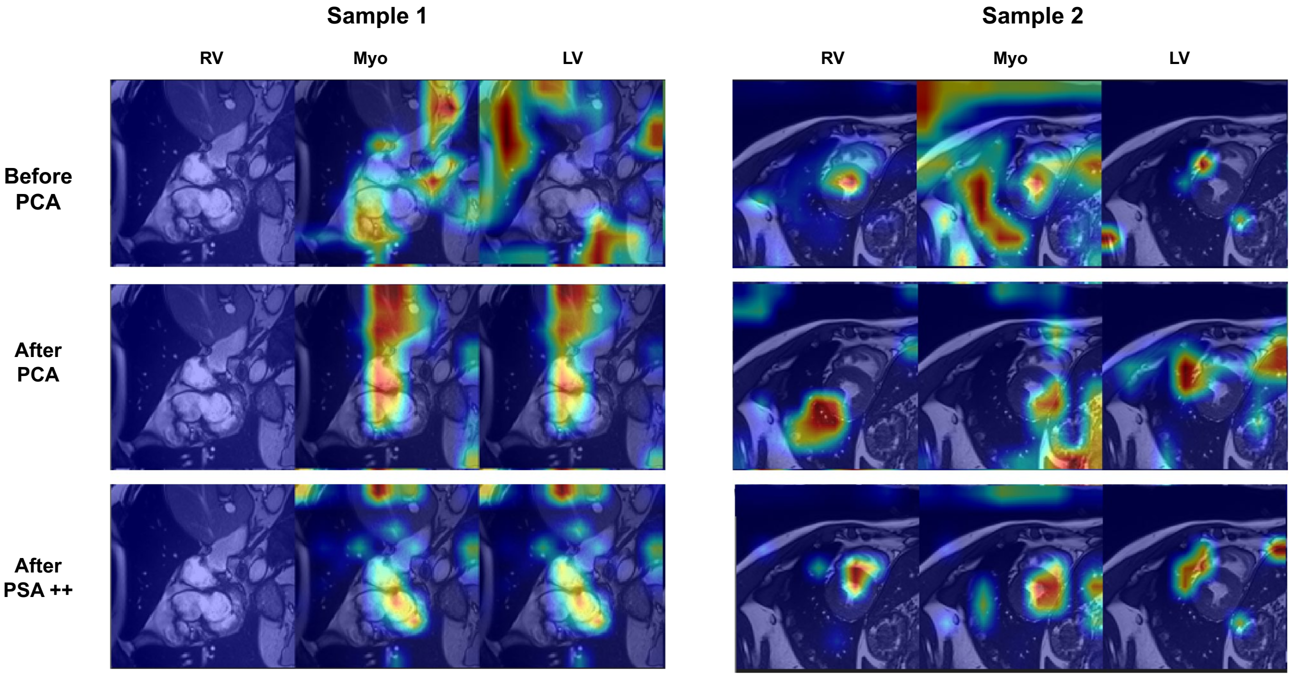


Fig. 6 Visualization of the sequential process through the proposed Priority Mixer Block to influence the segmentation output

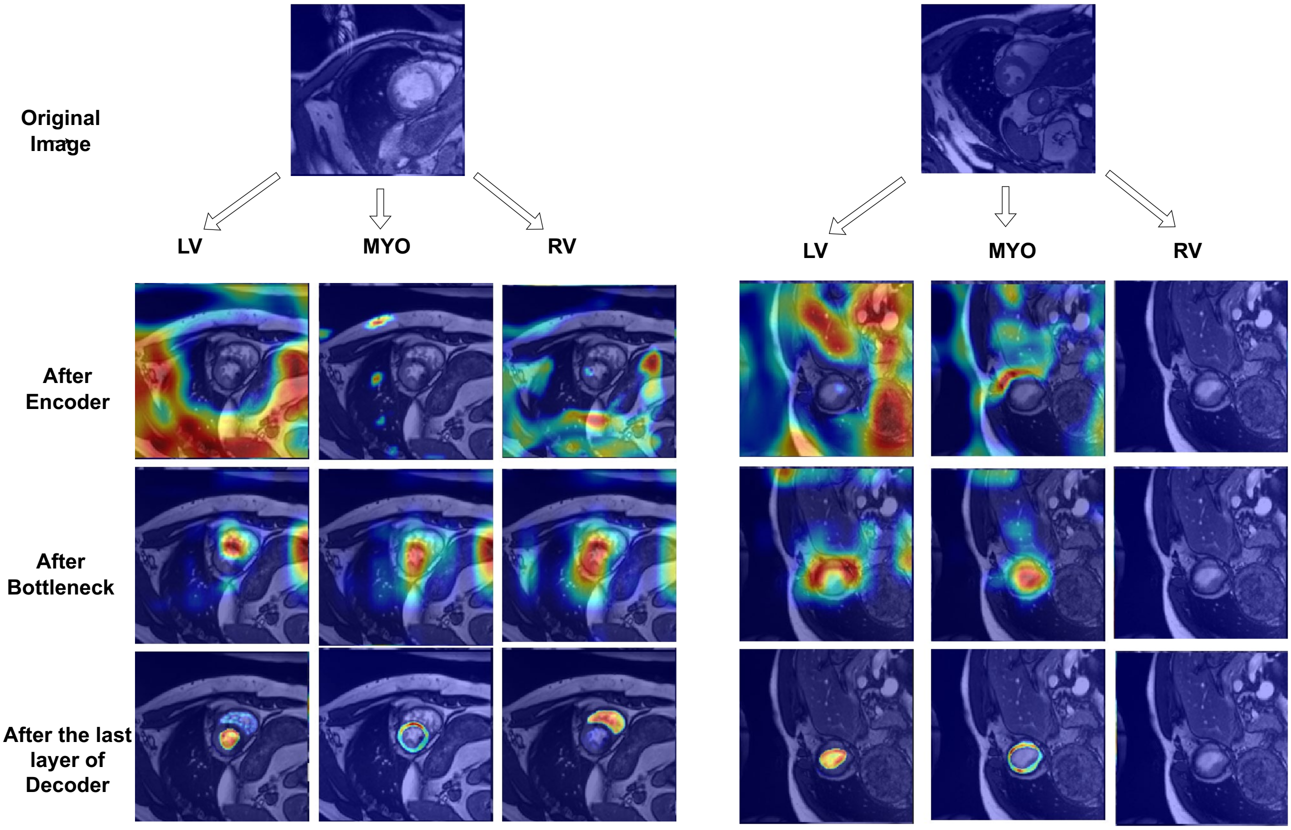


Fig. 7 The role of the Priority Mixer block in the entire model

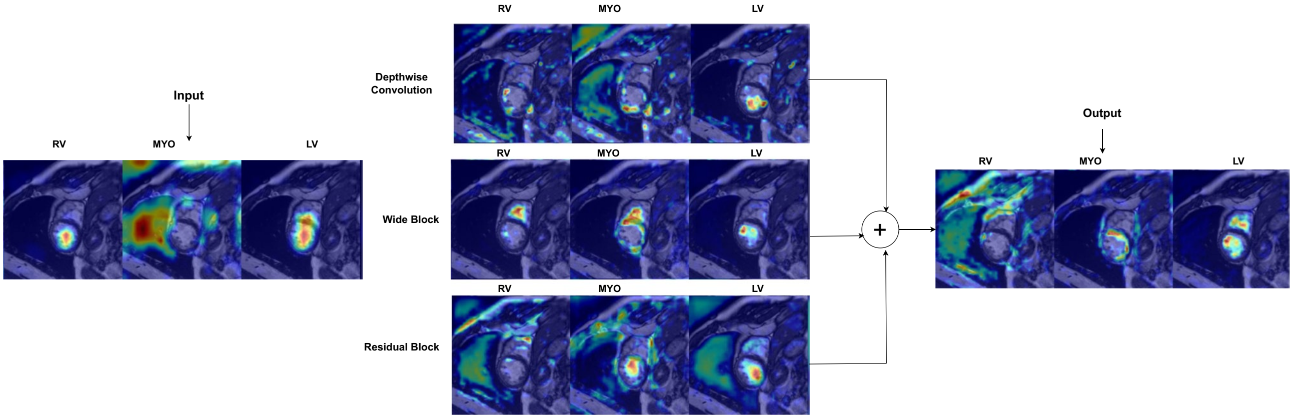
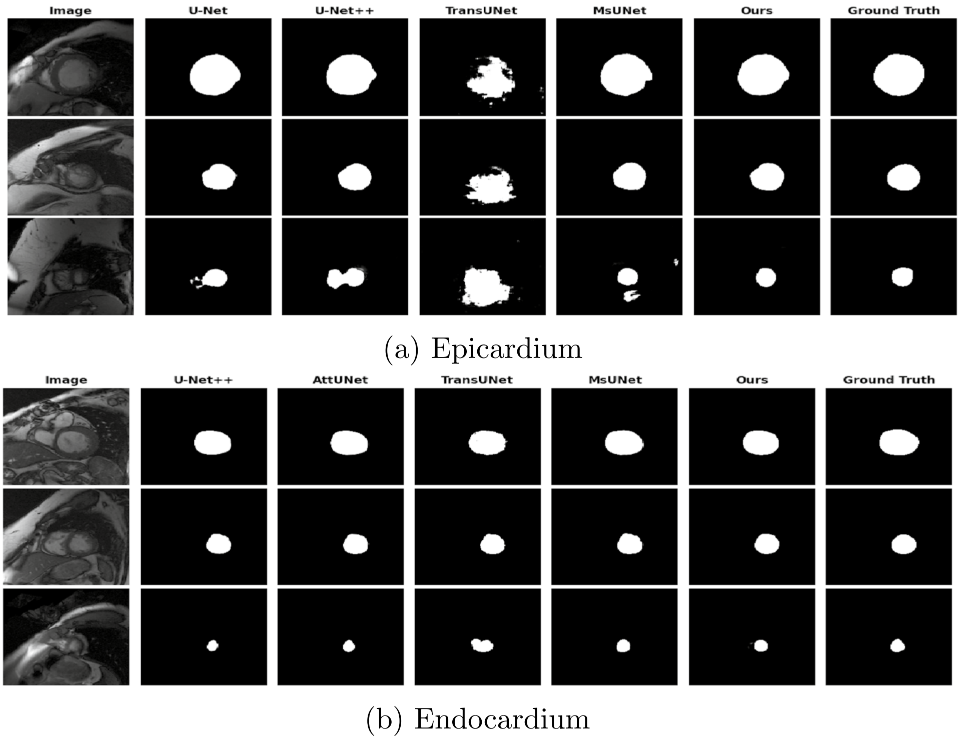


Fig. 8 Visualization of the sequential process through the proposed Depthwise-Focus Block

ground truths for both epicardial and endocardial regions in short-axis images including apex, mid, and base slices.

The quantitative results for the left ventricle are also given in Table 2. The scores in this table show that in the endocardium, the proposed CapNet gives better values for both DSC and IoU, 94.49% and 90.15% respectively, than almost all mentioned models. Specifically, U-Net++ gives a DSC of 92.65% (p=1.942 × 10^{-3}) and IoU of 86.78% (p=7.152 × 10^{-5}); SegNet achieves a DSC of 91.76% (p=7.777 × 10^{-10}) and IoU of 81.47% (p=8.933 × 10^{-13}); Res-Unet scores 90.97% DSC (p=1.910 × 10^{-5}) and 85.01% IoU (p=1.489 × 10^{-6}); DS-TransUnet records a DSC of 90.44% (p=3.008 × 10^{-8}) and IoU of 85.03% (p=3.519 × 10^{-13}); and Swin-Unet attains a DSC of 90.02% (p=9.613 × 10^{-10}) and IoU of 82.96% (p=1.132 × 10^{-14}). For the epicardium, the p-values computed for all models are smaller than 0.05, showing a significant performance of the proposed approach compared to other models. CapNet achieves superior DSC and IoU values of 96.82% and 93.93%, respectively, highlighting its robust performance. These results emphasize CapNet's remarkable improvement over other state-of-the-art models.

For better quantitative assessment, we provide the quantitative results by the boxplots of compared models in terms of DSC and IoU in Fig. 12. As can be easily observed from these figures, our proposed approach gives the highest medium and maximal scores for both DSC and IoU indices compared to comparative models.



(a) Epicardium

(b) Endocardium

Fig. 9 Segmentation results of top 5 on the Sunnybrook dataset

Table 1 The quantitative comparison between the proposed CapNet and SOTA on the Sunnybrook data. DSC and IoU scores are in mean (standard deviation)

Method	Paramter	Endocardium				Epicardium			
		DSC	*p*-value	IoU	*p*-value	DSC	*p*-value	IoU	*p*-value
U-Net [10]	31.1M	0.9055 (0.077)	5.776×10^{-4}	0.8351 (0.111)	2.296×10^{-4}	0.9492 (0.051)	0.2485	0.9071 (0.080)	0.2540
U-Net++ [50]	10.2M	0.9265 (0.059)	0.1047	0.8678 (0.087)	0.0826	0.9525 (0.044)	0.3802	0.9123 (0.071)	0.4011
Attention-Unet [11]	31.9M	0.9251 (0.074)	0.0291	0.8663 (0.102)	0.0187	0.9533 (0.036)	4.506×10^{-3}	0.9139 (0.061)	3.719×10^{-3}
SegNet [51]	29.4M	0.9176 (0.066)	0.0125	0.8539 (0.100)	9.536×10^{-3}	0.8970 (0.078)	1.332×10^{-6}	0.8211 (0.113)	8.379×10^{-8}
Res-Unet [52]	17.6M	0.9097 (0.062)	3.710×10^{-4}	0.8395 (0.091)	1.019×10^{-4}	0.9345 (0.059)	0.0129	0.8821 (0.093)	0.0113
DS-TransUnet [53]	171.4M	0.9133 (0.057)	1.031×10^{-3}	0.8467 (0.083)	4.612×10^{-4}	0.9386 (0.038)	3.810×10^{-3}	0.8870 (0.064)	3.449×10^{-3}
MSU-Net [54]	47.1M	0.9291 (0.046)	0.1302	0.8708 (0.074)	0.1029	0.9583 (0.028)	0.8270	0.9212 (0.050)	0.8296
TransUNet [12]	66.9M	0.9174 (0.083)	0.0336	0.8561 (0.114)	0.0261	0.8835 (0.069)	0.0335	0.8846 (0.104)	0.0291
Swin-Unet [13]	41.5M	0.9142 (0.053)	8.727×10^{-4}	0.8460 (0.083)	3.503×10^{-4}	0.9361 (0.035)	5.893×10^{-4}	0.8818 (0.058)	3.757×10^{-4}
CapNet (ours)	1.53M	**0.9400 (0.043)**		**0.8895 (0.069)**		**0.9593 (0.027)**		**0.9230 (0.047)**	

The signficance values have been provided and emphasized in the abstract

In addition, we show the performance of multiclass segmentation on the ACDC dataset in Fig. 13. The segmented regions include the right ventricle (RV), myocardium (MYO), and left ventricle (LV). As can be seen from this figure, the segmentation by the proposed CapNet is close to ground truths, while the under-segmentation occurs in results by the U-Net.

For multiclass segmentation, to better quantitatively assess, we show the boxplots of compared models in terms of DSC in Fig. 14. As can be seen from these figures, compared to other models, our proposed approach gives the highest medium and maximal scores in terms of DSC scores for all regions including the RV, Myo, LV areas, and AVG values.

In order to compare the evaluation scores, we provided the DSC scores by comparative models on the ACDC data in Table 3 for segmented regions including the right ventricle (DiceRV), myocardium (DiceMYO), and left ventricle (DiceLV), and the average values of the three regions (DiceAvg).

The quantitative comparison of the proposed CapNet with state-of-the-art models on the ACDC dataset is presented in Table 3. For the right ventricle (RV), CapNet achieves a DSC of 92.34%, significantly outperforming SegNet (p=1.235×10^{-3}), Res-Unet (p=4.116×10^{-3}), DS-TransUnet (p=0.0125), TransUNet (p=1.361×10^{-6}), and Swin-Unet (p=1.784×10^{-4}). In the myocardium (Myo), CapNet's DSC of 90.95% is notably better than U-Net++ (p=0.0450), SegNet (p=3.984×10^{-5}), Res-Unet (p=8.653×10^{-3}), DS-TransUnet (p=4.294×10^{-4}), TransUNet (p=1.485×10^{-9}), and Swin-Unet (p=2.353×10^{-8}). For the left ventricle (LV), CapNet achieves a DSC of 95.86%, surpassing SegNet (p=9.66×10^{-3}), Res-Unet (p=0.0276), DS-TransUnet (p=0.0195), TransUNet (p=2.975×10^{-4}), and Swin-Unet (p=1.951×10^{-5}). Overall, CapNet achieves an average DSC of 93.05%, demonstrating significant improvements over SegNet (p=1.706×10^{-5}), Res-Unet (p=2.673×10^{-4}), DS-TransUnet (p=9.892×10^{-5}), TransUNet (p=6.289×10^{-10}), and Swin-Unet (p=2.373×10^{-11}). These results clearly illustrate the remarkable performance of CapNet across all evaluated metrics. It is worth mentioning that, compared to those models, our model has the smallest number of parameters as shown in the second column of Table 3.

Evaluation on the MS-CMR Dataset

We conducted experimental studies on the MS-CMR 2019 dataset to investigate segmentation. We present the performance of segmentation specifically on three subsets (bSSFP cine, T2-SPAIR, LGE) of CMR sequence images, which were evaluated using the Dice coefficients for the right ventricle (DiceRV), myocardium (DiceMYO), and left ventricle (DiceLV). Additionally, we calculated the average Dice coefficient (DiceAvg) for the three regions. These evaluations were performed using the proposed model and compared to other models as shown in Fig. 15.

In Fig. 15, we present the top 5 models with the best mean DSC results. As shown in this figure, the segmentation results

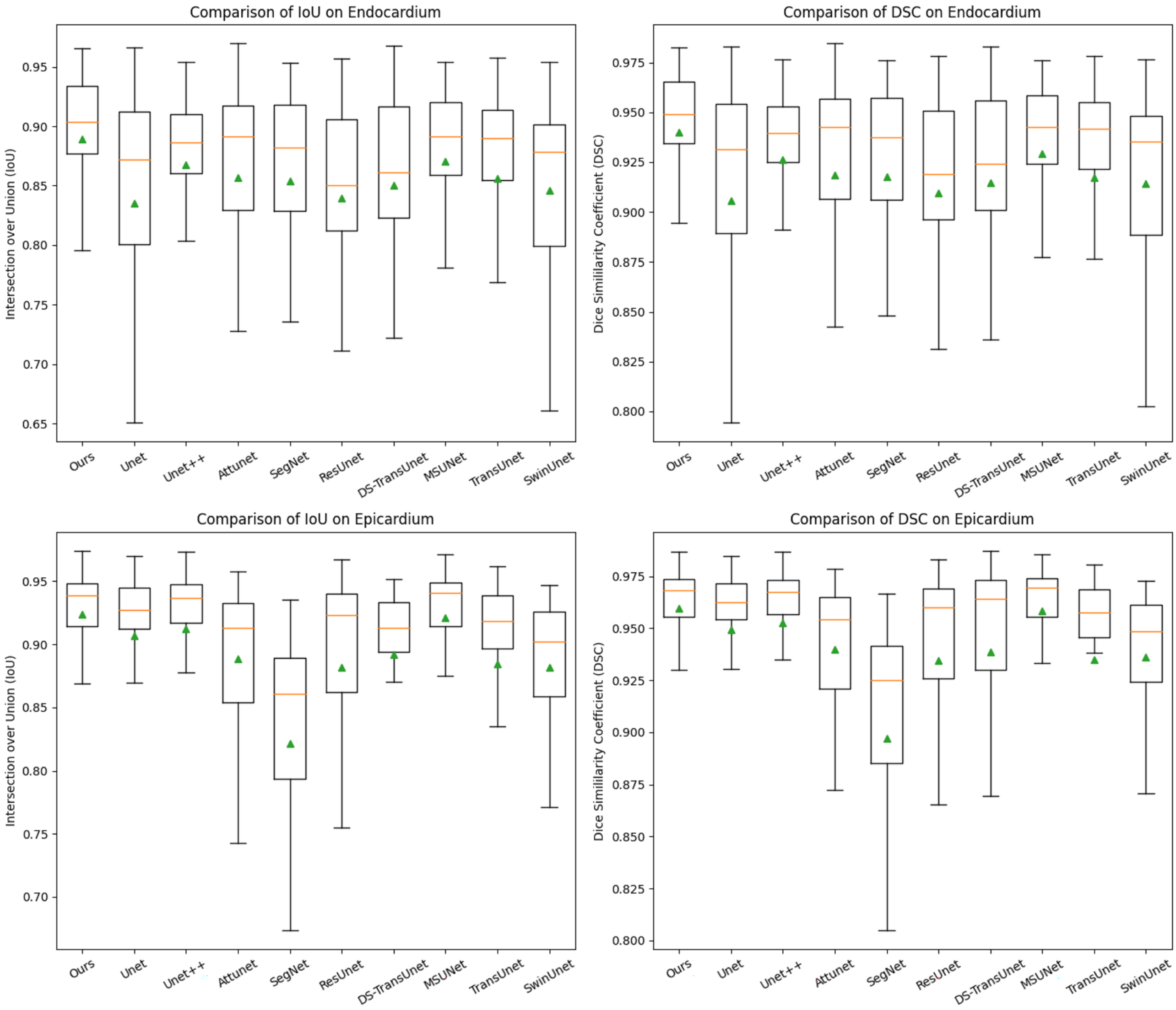


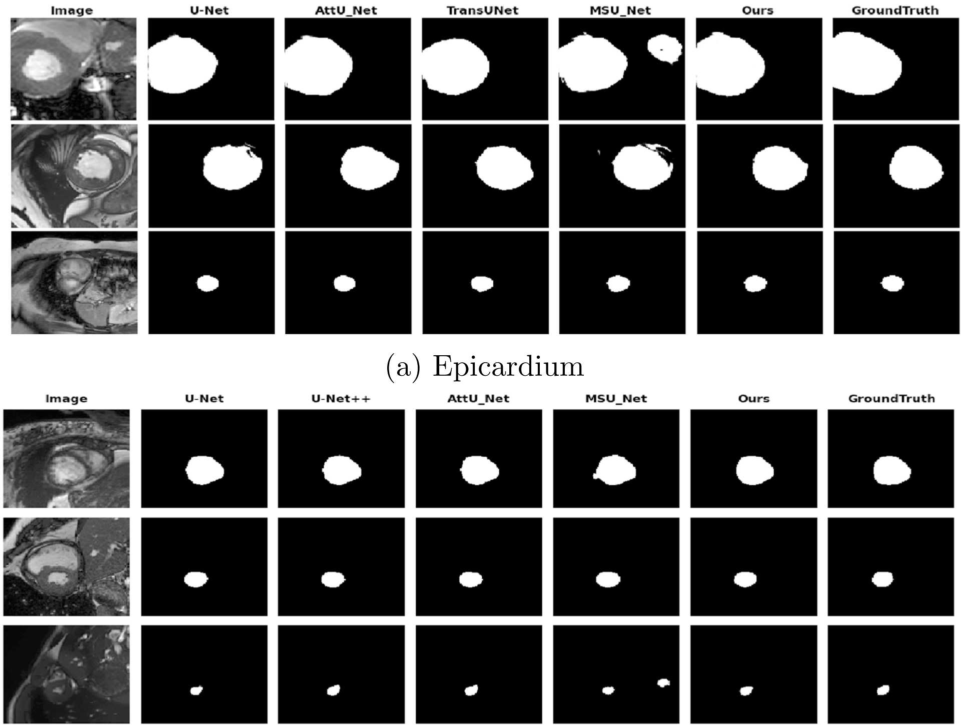
Fig. 10 Boxplots of IoU and DSC scores on Sunnybrook dataset of different models for the endocardium (top), and epicardium (bottom)

by the proposed model are best close to the ground truth on all three image sets: bSSFP, T2, and LGE. In panel a of this figure, the small slice masks (bottom) make it difficult to capture the details of Myo and RV, resulting in significant discrepancies among the compared models. Furthermore, in panel b, the slice masks (middle) exhibit over-segmentation in the compared models. Finally, in panel c, for the LGE cine sequence image masks, discrepancies with the ground truth among the models will occur in small-sized masks such as U-Net and Attention-Unet.

The quantitative results for the MS-CMR 2019 dataset are also provided in Table 4. Across the entire MS-CMR 2019 dataset, based on the computed *p*-values, we can see that the proposed model outperforms most other models, excepting the U-Net++, Attention-Unet, MSU-Net, and nnUnet (for the T2-SPAIR sequences), in terms of performance in the majority of regions (RV, Myo, LV).

Our CapNet model demonstrates outstanding performance across different CMR sequence images when compared to several state-of-the-art models, particularly those with statistically significant *p*-values (less than 0.05). In the bSSFP cine CMR sequence images, CapNet achieves DSC scores of 94.65% for RV, 92.05% for Myo, 97.06% for LV, and an average DSC of 94.59%, which significantly outperforms SegNet, Res-Unet, DS-TransUnet, TransUNet, and nnUNet in terms of DSC scores for the RV, Myo, and LV, as well as the average DSC score ($p< 0.05$).

In the T2-SPAIR CMR sequence images, CapNet achieves DSC scores of 90.47% for RV, 90.97% for Myo, 95.21% for LV, and an average DSC of 92.22%. Thus, our model shows superior performance compared to SegNet (p=0.0153 for RV, p=0.0235 for Myo, p=0.0359 for LV, and p=1.274×10^{-3} for the average DSC), Res-Unet (p=5.474×10^{-6} for RV,



(a) Epicardium

(b) Endocardium

Fig. 11 Representative segmentation results of top 5 on the left ventricle of the ACDC dataset

p=1.162 × 10^{-3} for Myo, and p=8.942 × 10^{-7} for the average DSC), DS-TransUnet (p=6.955 × 10^{-5} for RV, p=2.291 × 10^{-3} for Myo, and p=2.335 × 10^{-7} for the average DSC), and TransUNet (p=1.092 × 10^{-6} for RV, p=5.318 × 10^{-10} for Myo, p=2.997 × 10^{-5} for LV, and p=4.108 × 10^{-14} for the average DSC). Additionally, for the LGE CMR sequence images, our

Table 2 The quantitative comparison between the proposed CapNet and SOTA on the left ventricle of ACDC dataset. DSC and IoU scores are in mean (standard deviation)

Method	Params	Endocardium				Epicardium			
		DSC	p-value	IoU	p-value	DSC	p-value	IoU	p-value
U-Net [10]	31.1M	0.9355 (0.097)	0.1874	0.8894 (0.121)	0.1891	0.9606 (0.057)	0.0366	0.9202 (0.083)	4.575 × 10^{-4}
U-Net++ [50]	10.2M	0.9265 (0.097)	1.942 × 10^{-3}	0.8678 (0.131)	7.152 × 10^{-5}	0.9525 (0.029)	1.042 × 10^{-12}	0.9123 (0.049)	6.288 × 10^{-13}
Attention-Unet [11]	31.9M	0.9301 (0.108)	0.0692	0.8838 (0.128)	0.0643	0.9607 (0.038)	4.183 × 10^{-3}	0.9266 (0.059)	2.601 × 10^{-3}
SegNet [51]	29.4M	0.9176 (0.152)	7.777 × 10^{-10}	0.8147 (0.177)	8.933 × 10^{-13}	0.9243 (0.127)	1.822 × 10^{-8}	0.8756 (0.141)	5.498 × 10^{-13}
Res-Unet [52]	17.6M	0.9097 (0.122)	1.910 × 10^{-5}	0.8501 (0.149)	1.489 × 10^{-6}	0.9522 (0.076)	6.874 × 10^{-4}	0.9153 (0.094)	8.651 × 10^{-5}
DS-TransUnet [53]	171.4M	0.9044 (0.060)	3.008 × 10^{-8}	0.8503 (0.121)	3.519 × 10^{-13}	0.9400 (0.072)	4.771 × 10^{-10}	0.8902 (0.109)	3.081 × 10^{-13}
MSU-Net [54]	47.1M	0.9330 (0.091)	0.0794	0.8847 (0.119)	0.0653	0.9596 (0.050)	7.741 × 10^{-3}	0.9257 (0.071)	4.478 × 10^{-3}
TransUNet [12]	66.9M	0.9308 (0.087)	0.0334	0.8802 (0.115)	0.0180	0.9469 (0.097)	3.710 × 10^{-4}	0.9092 (0.110)	1.739 × 10^{-5}
Swin-Unet [13]	41.5M	0.9002 (0.102)	9.613 × 10^{-10}	0.8296 (0.122)	1.132 × 10^{-14}	0.9389 (0.072)	9.965 × 10^{-11}	0.8873 (0.111)	1.552 × 10^{-12}
CapNet (ours)	1.53M	**0.9449 (0.064)**		**0.9015 (0.094)**		**0.9682 (0.023)**		**0.9393 (0.039)**	

The signfificance values have been provided and emphasized in the abstract

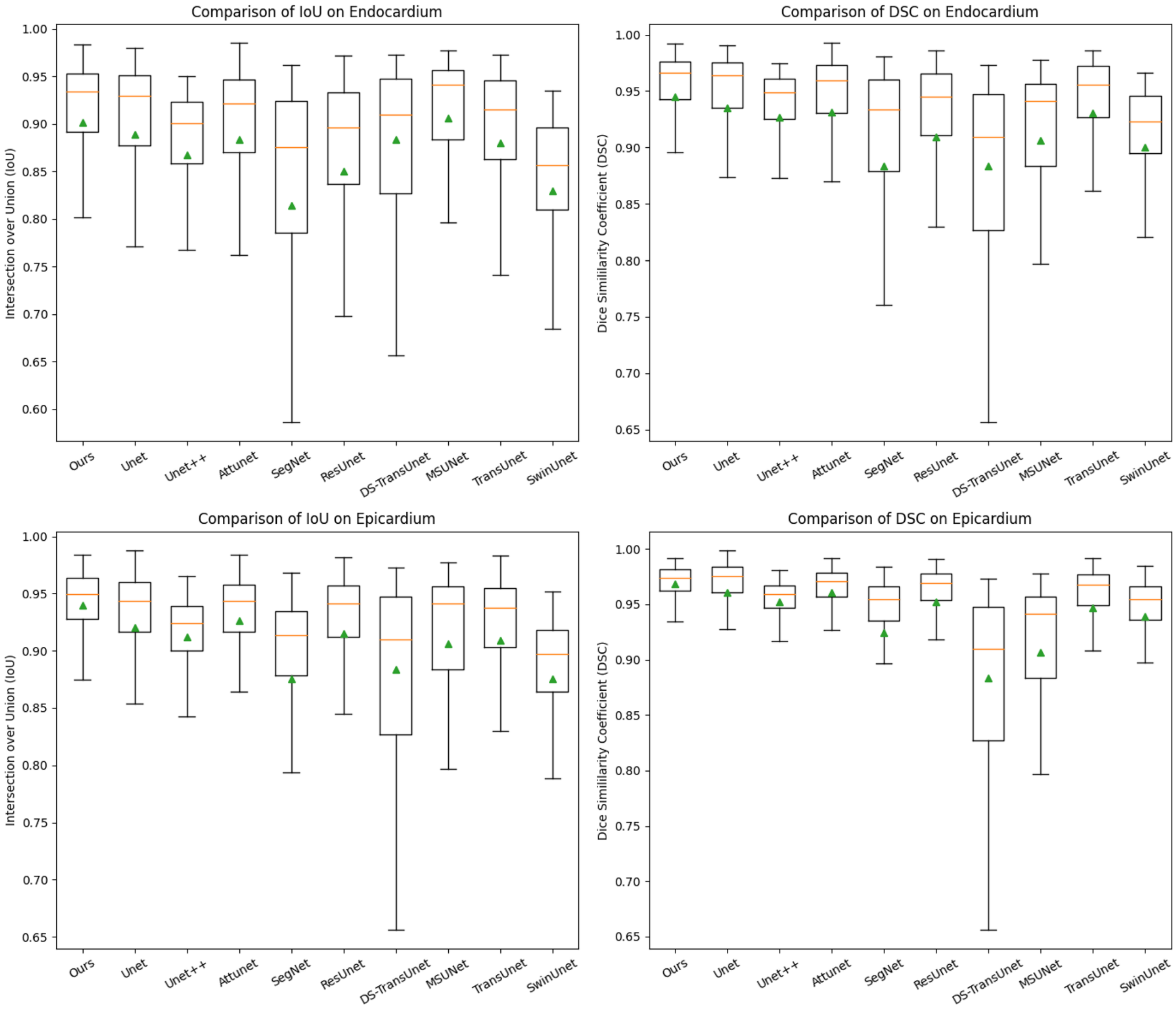


Fig. 12 Boxplots of IoU and DSC scores of endocardium (top) and epicardium (bottom) on left ventricular ACDC dataset of different models

model obtains DSC scores of 94.77% for RV, 91.24% for Myo, 95.96% for LV, and an average DSC of 93.99%. This shows that proposed model excels with significantly higher DSC scores compared to SegNet (p=2.535 × 10^{-3} for RV, p=1.241 × 10^{-5} for Myo, p=5.186 × 10^{-4} for LV, and p=1.780 × 10^{-6} for the average DSC), Res-Unet (p=3.462 × 10^{-9} for RV, p=1.218 × 10^{-9} for Myo, p=1.710 × 10^{-9} for LV, and p=6.573 × 10^{-14} for the average DSC), DS-TransUnet (p=6.159 × 10^{-3} for RV, and p=2.076 × 10^{-3} for the average DSC), TransUNet (p=4.316 × 10^{-4} for RV, p=1.500 × 10^{-5} for Myo, p=4.161 × 10^{-4} for LV, and p=1.142 × 10^{-7} for the average DSC), and Swin-Unet (p=1.607 × 10^{-3} for RV, p=0.0111 for Myo, p=4.228 × 10^{-4} for LV, and p=8.366 × 10^{-5} for the average DSC). These results clearly illustrate the exceptional performance of CapNet, making it a highly effective model for CMR image segmentation.

For better quantitative assessment, we provide the quantitative results using boxplots of compared models in terms of Dice scores in the LGE CMR sequence images of the MS-CMR dataset in Fig. 16. As can be easily observed from these figures, our proposed approach provides the highest mean and maximum scores for all the regions of interest in the dataset.

Ablation Study

Performance of the Hyperparameters α, β, and m on the Proposed Loss

To find suitable values for the hyperparameters α and β, we fixed m= 2 as the exponent parameter in the proposed loss function. Similar to the Tversky loss, we gradually vary α and β by decreasing α and increasing β, with their

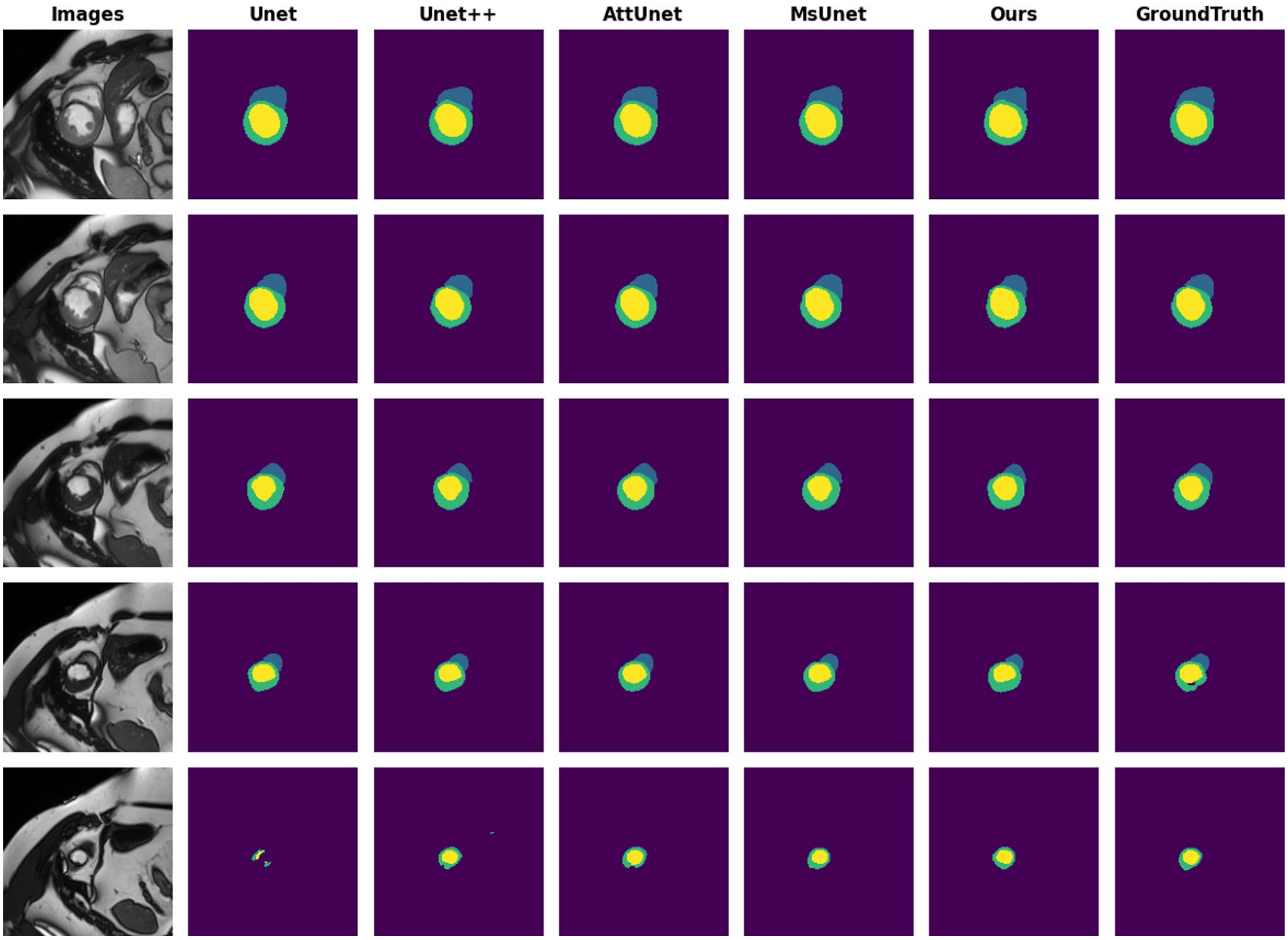


Fig. 13 Representative segmentation results of top 5 of the right ventricle, myocardium, and left ventricle on the ACDC dataset

sum equal to 1. The experimental results are evaluated as shown in Table 5. When $\alpha = \beta = 0.5$, we obtained DSC Avg with (92.37%) as the result. After slightly reducing α to 0.4 and increasing β to 0.6, there was a slight improvement in the results DSC Avg with (92.44%). Especially, when we decreased α to a ratio of 3:7 with β, we achieved the best DSC Average (AVG) value with DSC Avg (93.05%). However, when we decreased $\alpha = 0.2$, $\beta = 0.8$, and $\alpha = 0.1$, $\beta = 0.9$, the performance decreased compared to the best ratio of 3:7.

After finding suitable values for the parameters α and β, we will fix them at $\alpha = 0.3$ and $\beta = 0.7$ and gradually vary the exponent parameter m. As shown in Table 6, we obtained results for m= 1, which were not satisfactory, with DSC Average (Avg) on the ACDC data (91.64%). However, as we increased m to values greater than 1, specifically based on the data in the table, when m= $\frac{4}{3}$, the results gradually improved with a DSC Average (92.20%). Subsequently, when we increased m slightly to m= 2, we achieved good results as shown in the table, with a DSC Average (Avg) (93.05%). Overall, the best performance is achieved when the exponent parameter m ranges from $\frac{4}{3}$ to 3, with particularly good results at m= 2.

To check whether there are any statistical differences between segmentation scores when using various combinations of α, β, and m, we computed the p-values by the statistical tests. In particular, we compute the p-values on DSC scores when using other combinations with our chosen hyperparameters, $\alpha = 0.3$ and $\beta = 0.7$ (last row of Table 5), as well as the chosen $m = 2$ (last row) in Table 6. The p-values show no statistical differences when using various hyperparameter combinations. This also implies that the proposed loss is not too statistically sensitive to hyperparameters.

Performance of the Proposed Loss

Experiments to assess the performance of the proposed loss are provided in Table 7. In these experiments, the proposed model is trained with some common loss functions including the Tversky, focal Tversky, BCE-Dice, and active contour losses. In the first experiment, we conduct the binary segmentation of the endocardium and epicardium on the

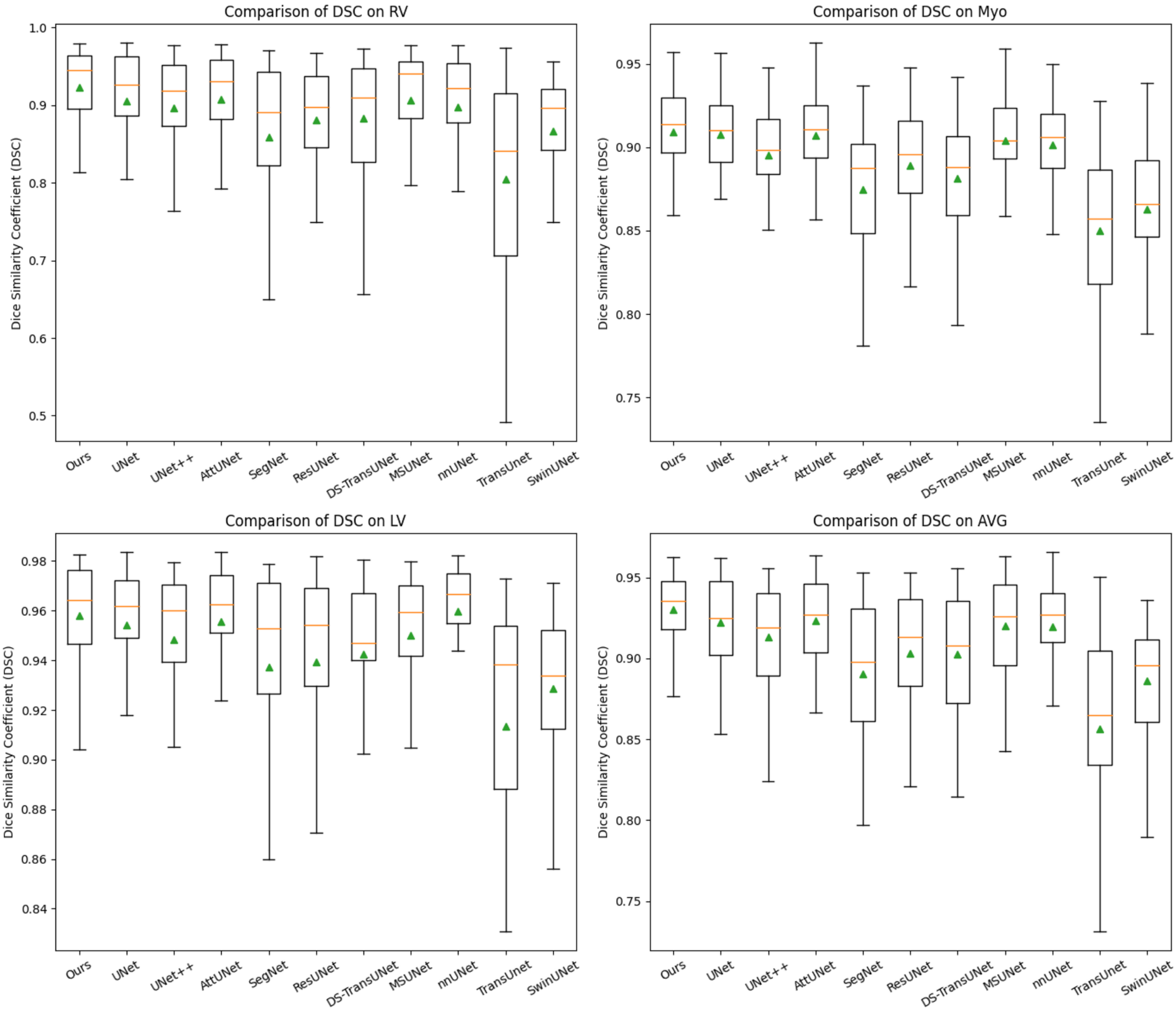


Fig. 14 Boxplots of DSC scores of different models for multiclass segmentation on ACDC dataset

Sunnybrook dataset. As shown in Table 7(a), the TSPD loss produces superior results compared to other losses in terms of the average DSC and IoU metrics for both endocardium and epicardium regions. However, considering the *p*-value, the scores by the TSPD loss are not statistically significant compared to other losses. The second experiment is evaluated on the case of multiclass segmentation performed on the ACDC dataset. The quantitative results are given in Table 7(b). As can be observed from the table, the results on the ACDC dataset clearly demonstrate the superior effectiveness of the proposed loss function. The average Dice score of 93.05% is 0.7% higher than the result ranked second. Another metric, the DiceLV, also indicates the superiority of the results in terms of the proposed loss function. The *p*-values in the last column of Table 7(b) show significant differences compared to the Dice, BCE-Dice, Tversky, and focal Tversky losses.

Performance of the PCA-PSA++ Architecture in the Bottleneck

To demonstrate the effectiveness of the PCA-PSA++ model, we have attempted to replace it with several modules to compare their performance. The compared results are provided in Table 8. Although the number of parameters may not be optimal, the trade-off is that the performance in terms of the dice score significantly surpasses that of other modules. Previous modules only demonstrated effectiveness with ASPP, but had significantly longer computation times. Compared to the previous version, PCA-PSA, the increase in the number of

Table 3 The quantitative comparison between the proposed CapNet and SOTA on multiclass of ACDC dataset. DSC score is in mean (standard deviation)

Methods	Params	RV		Myo		LV		Average	
		DSC	*p*-value	DSC	*p*-value	DSC	*p*-value	DSC	*p*-value
U-Net [10]	31.1M	0.9046 (0.073)	0.1951	0.9073 (0.028)	0.7884	0.9541 (0.031)	0.5081	0.9220 (0.030)	0.1853
U-Net++ [50]	10.2M	0.8958 (0.075)	0.0610	0.8952 (0.032)	0.0450	0.9482 (0.034)	0.1348	0.9131 (0.031)	6.881×10^{-3}
Attention-Unet [11]	31.9M	0.9069 (0.070)	0.2513	0.9071 (0.030)	0.7710	0.9555 (0.032)	0.6736	0.9232 (0.030)	0.2615
SegNet [51]	29.1M	0.8591 (0.113)	1.235×10^{-3}	0.8747 (0.042)	3.984×10^{-5}	0.9371 (0.045)	9.66×10^{-3}	0.8903 (0.053)	1.706×10^{-5}
Res-Unet [52]	17.6M	0.8808 (0.075)	4.116×10^{-3}	0.8889 (0.037)	8.653×10^{-3}	0.9393 (0.048)	0.0276	0.9030 (0.040)	2.673×10^{-4}
DS-TransUnet [53]	171.4M	0.8835 (0.084)	0.0125	0.8813 (0.039)	4.294×10^{-4}	0.9425 (0.035)	0.0195	0.9024 (0.037)	9.892×10^{-5}
MSU-Net [54]	47.1M	0.9062 (0.073)	0.2387	0.9037 (0.030)	0.4341	0.9502 (0.034)	0.2250	0.9209 (0.030)	0.1051
nnUNet [55]	37.6M	0.8974 (0.085)	0.1066	0.9016 (0.028)	0.2571	0.9597 (0.025)	0.7695	0.9196 (0.031)	0.0886
TransUNet [12]	66.9M	0.8051 (0.145)	1.361×10^{-6}	0.8501 (0.053)	1.485×10^{-9}	0.9133 (0.075)	2.975×10^{-4}	0.8561 (0.072)	6.289×10^{-10}
Swin-Unet [13]	41.5M	0.8665 (0.079)	1.784×10^{-4}	0.8627 (0.042)	2.353×10^{-8}	0.9288 (0.036)	1.951×10^{-5}	0.8860 (0.033)	2.373×10^{-11}
CapNet (ours)	1.53M	**0.9234 (0.054)**		**0.9095 (0.031)**		**0.9586 (0.022)**		**0.9305 (0.024)**	

The signficance values have been provided and emphasized in the abstract

parameters is not significant. The metrics for Myo may show relatively similar values, but the parameters for RV and LV exhibit a significant increase, particularly in the case of RV.

Performance of the Depthwise-Focus Architecture in the Decoder

With another contribution in the paper, the Depthwise-Focus, we created Tables 9 and 10 to examine whether the depthwise axial with a kernel size of 7 is truly beneficial for the model. We sequentially replaced the conventional convolution and axial convolution to compare them with the proposed depthwise axial method used in our model. In Table 9, we kept the kernel size as 7 and added increasing dilations from 1 to 4 to observe the effectiveness. All three methods yield corresponding results, with the Dice score gradually increasing from $d = 1$ to the combination of $d = 1$ and $d = 2$, reaching its peak when the combination of $d = 1, d = 2$, and $d = 3$ is used. However, when dilation 4 is added, the results relatively decrease. The depthwise axial method also demonstrates effectiveness when at the same dilation level, as the DiceLV consistently shows higher values compared to the other two methods. We also experimented with different kernel sizes such as 3, 5, and 7. The results with a kernel size of 5 were lower compared to the other two kernels, while the kernel size of 7 demonstrated dominance across all three methods. However, using a regular convolution with a kernel size of 7 would lead to a significant increase in the number of parameters. On the other hand, with depthwise convolutions, there is no significant difference in the number of parameters between kernel sizes 3, 5, and 7. Indeed, the application of depthwise axial convolutions has yielded favorable results while also reducing the number of parameters.

Discussion

Contribution of the Study

This study presents the CapNet model, which is based on the mechanism of attention clustering for local information and incorporates a smooth feature flow processing using the proposed Priority Mixer block at the bottleneck. Additionally, a decoding module namely Depthwise-Focus block is employed, leveraging creative convolutional techniques to enhance the accuracy of the predicted labels. Besides, we propose a new loss called Tversky Shape Power Distance (TSPD) loss function. We conducted experiments on various datasets to demonstrate the effectiveness of our proposed architecture and loss function compared to other methods that utilize different loss functions. Specifically, we performed experiments on well-known datasets used for cardiac segmentation: Sunnybrook, ACDC, and MS-CMR datasets. The segmentation performances are evaluated by DSC and IoU metrics, and the statistical significance analysis by a statistical test with p-value is made. Our results show that the TSPD loss function consistently outperforms other loss functions in most cases.

By experiments, we found that in the context of cardiac image segmentation based on deep learning, the CNN-based methods can outperform the transformer-based models. The transformer-based approach, though having shown performances in many computer vision areas, still suffers from drawbacks when working with limited training data. In our study, the results on the three datasets, including Sunnybrook, ACDC, and MS-CMR, show that compared to the transformer-based methods like TransUnet, DS-TransUnet, and Swin-Unet, the proposed CapNet still gives better scores in terms of DSC and IoU scores and shows statically significant differences.

On another hand, the current CNN-based or transformer-based approaches for cardiac MRI image segmentation still entail a large number of parameters. This motivated us to build a lightweight model. To the best of our knowledge, there are a few prior lightweight models specifically designed for this task. The proposed CapNet model is only with 1.53 million parameters, 20 times less than the well-known U-Net. With fewer parameters, we can reduce the memory size and computational complexity of the segmentation tasks, especially for edge devices.

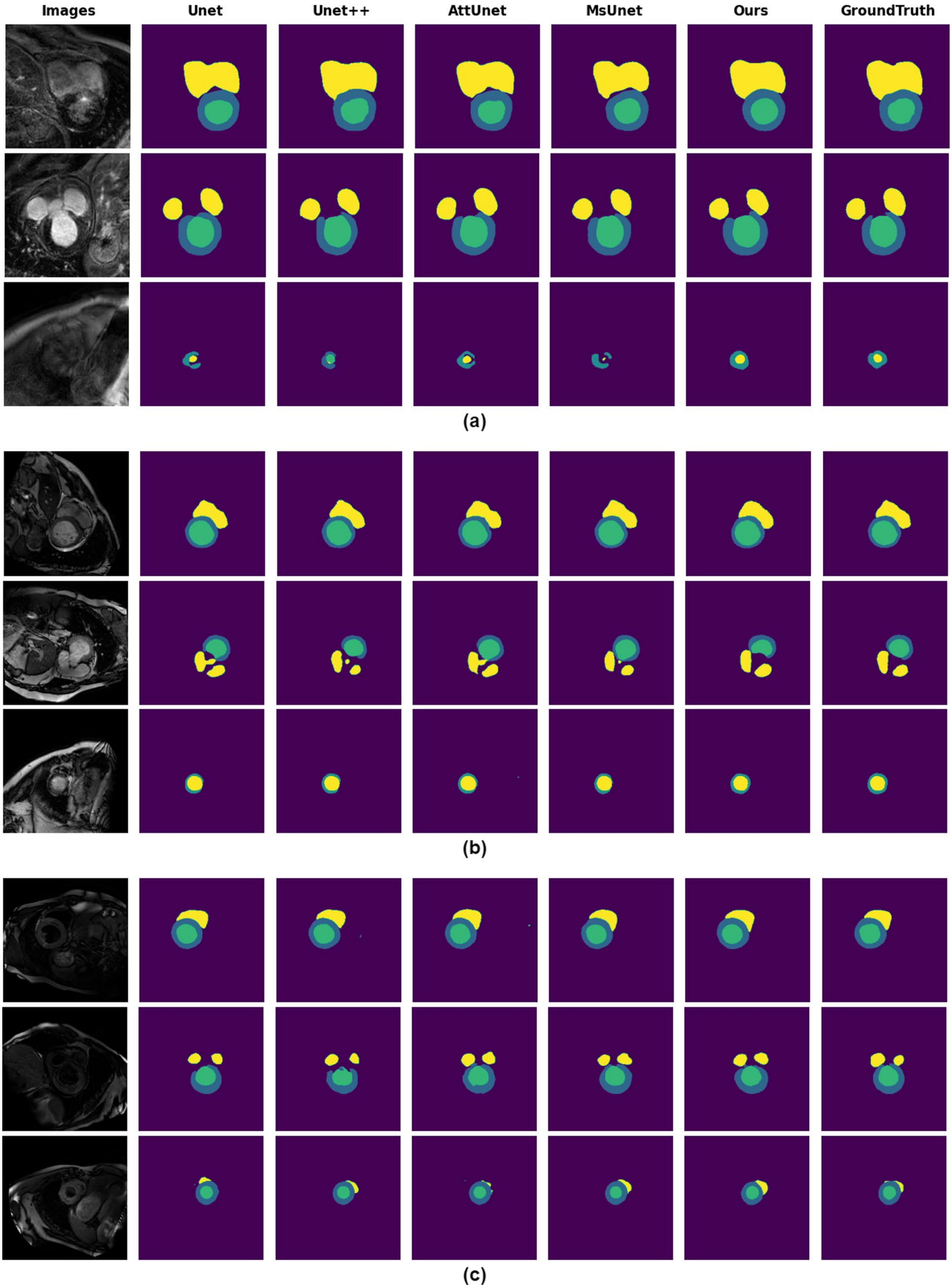
Regarding using the PCA-PSA++ architecture, the PCA, PSA, and its variant, PSA++, do not possess any learnable parameters. They function as normalization transformations or systems that reorient the outputs. Would incorporating learnable vectors or matrices into PSA/PSA++ or PCA yield significant improvements? In practice, when introducing learnable vectors or matrices, these values are random and have no predefined upper or lower bounds. The possibility of having excessively large gradients may result in the learned parameters being updated with very large or very small values, leading to instability in learning. Furthermore, when applying normalization functions to these vectors or matrices within the range of 0–1, such as sigmoid or softmax, $\lim_{x\to-\infty} \sigma(x) = \lim_{x\to-\infty} \frac{1}{1+e^{-x}} = 0$, it tends to approach 0 when x is very small. Similarly, softmax tends to have one dominant component, while the remaining components converge to zero. Consequently, there is a significant loss of information when using these functions. Another approach could be to set upper and lower bounds for the parameters, but it is not possible to exclude the possibility that all parameter values will be at the lower bound. Alternatively, using sinusoidal functions like *sin*() or *cos*() for normalization within the range of 0–1 may not guarantee satisfactory results due to their periodic nature. Therefore, we propose PSA/PSA++ and PCA without incorporating parameters to demonstrate their passive adaptive capability.

Data Sampling Size

Considering the data sampling size, it is evident that the initial datasets, including Sunnybrook, ACDC, and MS-CMR, consist of a relatively small number of 3D volume samples. For example, the ACDC dataset includes 100 samples, with 70 for training, 10 for validation, and 20 for testing. However, we have taken steps to increase the effective size of the dataset by slicing the 3D volumes along the z-axis, resulting in a larger number of 2D slices. Specifically, the 70 training samples of the ACDC were converted into 1312 slices for training, the 10 validation samples were split into 202 slices for validation, and the 20 test samples were transformed into 388 slices for testing.

In a similar way, we applied this technique for the cine-MRI dataset of Sunnybrook, with 70% for training, 15% for validation, and 15% for testing. With epicardium, 70% training was converted into 191 slices, 15% for validation to 41 slices, and 41 remaining 2D slices for testing, and then endocardium with 70% for training (369 slices), 15% for validation (79 slices), and 15% for testing (79 slices). The same approach is also applied for the MS-CMR dataset. In this data, we categorize it into 3 types: bSSFP, T2, and LGE. We have performed preprocessing and obtained 333 slices for bSSFP, 148 slices for T2, and 75 slices for LGE in the dataset. By utilizing this slicing technique, we effectively increased the size of the dataset and the number of independent samples available for training, validation, and testing. This approach not only augments the dataset but also captures the inherent variation and diversity present within the 3D volumes, enhancing the generalizability and robustness of our model.

While we acknowledge that the initial 3D volume sample size was small, the slicing technique allowed us to leverage a significantly larger number of 2D slices, mitigating potential limitations in generalizability and robustness. The increased dataset size and diversity of samples provided a more comprehensive representation of the problem domain, enabling our model to learn and generalize more effectively. Additionally, we employed various data augmentation techniques, such as rotation, flipping, and scaling, to further increase the diversity of the training data and improve the model's ability to generalize to unseen samples. We understand the importance of validating our approach on a larger and more diverse dataset, and we will continue to explore opportunities to expand our dataset further. However, we believe that the slicing technique and data augmentation strategies employed in this study have effectively addressed the potential limitations of the initial sample size.



(a)

(b)

(c)

◂**Fig. 15** Representative segmentation results of the right ventricle, myocardium, and left ventricle on the MS-CMR 2019 dataset with different CMR sequence images **a** LGE CMR, **b** bSSFP, and **c** T2 CMR

Result Discussion and Hyperparameter Settings

To evaluate the performance of the proposed approach, we have conducted experiments using the CNN and transformer-based approaches on the binary and multiclass segmentation tasks. We reimplemented all SOTA models on three datasets for assessing the quantitative results, plotting, and statistical analysis. Considering the average values of DSC and IoU scores, our CapNet model obtained better performance compared to the SOTA for both the binary and multiclass segmentation cases. The statistical significance of the proposed method is shown compared with transformer-based methods like TransUnet and Swin-Unet, with p-value < 0.01.

In particular, for the binary segmentation, with Sunnybrook data, the proposed model gets DSC of 94% and IoU of 88.95% for endocardium and DSC of 95.93% and IoU of 92.30% for epicardium. With the same data, the average scores by the U-Net are 90.55% (DSC) and 83.51% (IoU) for endocardium and 94.92% (DSC) and 90.71% (IoU) for epicardium. The scores by DS-TransUnet are DSC of 91.33% and IoU of 84.67% for endocardium and 93.86% (DSC) and 88.70% (IoU) for epicardium. For the binary segmentation on the ACDC data, the CapNet obtains the DSC of 94.49% (endocardium) and 96.82% (epicardium) while the DSC scores by the Attention-Unet are 93.01% (endocardium) and 96.07% (epicardium). The results by the proposed model outperform those by the Swin-Unet (DSC of 90.02% for endocardium and 93.89% for epicardium). The IoU by CapNet is 90.15% for endocardium and 93.93% for epicardium, whereas the IoU values by the TransUNet are 88.02% for endocardium and 90.92% for epicardium on the ACDC data.

For the multiclass segmentation case, we conducted experiments on the ACDC and three sequences of the MS-CMR data for the right ventricle (RV), myocardium (Myo), left ventricle (LV), and the average regions. Similar to the binary case, the proposed CapNet model outperforms the SOTA in terms of average values of DSC and IoU and shows statistically significant differences compared to the transformer-based approach such as DS-TransUnet, TransUNet, and Swin-Unet. For the ACDC data, the mean DSC scores of CapNet are 92.34% (RV), 90.95% (Myo), and 95.86% (LV); meanwhile, the scores for the corresponding regions by the U-Net++ are 89.58% (RV), 89.52% (Myo), and 94.82% (LV). The scores by the Res-Unet are even lower, with 88.08% (RV), 88.89% (Myo), and 93.93% (LV). With MS-CMR data, notably, the proposed CapNet gives the DSC scores of 94.65% (bSSFP sequence), 90.47% (T2 sequence), and 94.77% (LGE sequence) for the RV, while the corresponding values by the nnUNet are 91.47% (bSSFP sequence), 89.55% (T2 sequence), and 92.34% (LGE sequence), and TransUNet are 91.90% (bSSFP sequence), 82.12% (T2 sequence), and 90.00% (LGE sequence). For the statistical analysis, the p-values show the significant differences while comparing the proposed CapNet with transformer-based methods, and some CNN-based methods like SegNet and Res-Unet. However, although the proposed model gives better performance in terms of average values, the statistical tests on the IoU and DSC by the proposed CapNet show that there is no difference when comparing with some CNN-based models like U-Net, nnUNet, and Attention-Unet. Nevertheless, it is worth mentioning that the proposed model is with much less parameters, 1.53 M, while the parameters of the U-Net are 31.1 M. The number of parameters in the nnUNet is 37.6 M, and Attention-Unet are 31.9 M.

Besides building the lightweight model for less parameters, developing a suitable loss for training the neural networks is also a promising approach. Building a loss does not increase any training parameters for the model, but instead can improve the performance significantly. In this study, inspired by the traditional optimization-based segmentation framework based on active contour and level set models, we build a novel region-based loss namely the Tversky Shape Power Distance. The loss allows adjusting the false positive rate or false negative rate (by choosing the α and β) to be compatible with the data. The loss can also be extended to the multiclass segmentation as in Eq. 6. Nevertheless, one of the major concerns in building the region-based losses [38, 39] for the segmentation tasks is choosing the hyperparameters for the false positive and the false negative (α and β). In fact, we need to conduct experiments to explore various combinations to estimate a suitable range and then choose suitable hyperparameters for the cardiac data. Based on these experiments, we found that the values of α= 0.3 and β= 0.7 yielded the best performance for the binary segmentation task, i.e., the endocardium and epicardium in the Sunnybrook dataset.

Building upon these findings, we extended our experiments to the multiclass segmentation task using the loss for multiclass segmentation in Eq. 8, using the ACDC and MS-CMR datasets. Although the ACDC and MS-CMR datasets involve four classes, we found that the same values α= 0.3 and β= 0.7 also provided the best overall performance, as shown in Table 5. We believe that these hyperparameter values strike a good balance between the topological similarity term and the pixel-wise similarity term in the TPSD loss function, enabling effective segmentation for both binary and multiclass scenarios. However, we acknowledge that these hyperparameters may not be the optimal values for all datasets and segmentation tasks. In our future work, we plan to explore more advanced techniques for hyperparameter tuning, such as automated hyperparameter optimization methods to further improve the performance of our TPSD loss function.

Table 4 Comparison between the proposed CapNet model and SOTA on the MS-CMR dataset with different sequences. (a) The bSSFP cine CMR sequence images. (b) The T2-SPAIR CMR sequence images. (c) The LGE CMR sequence. DSC scores are in mean (standard deviation)

(a) The bSSFP cine CMR sequence images

Methods	Params	RV		Myo		LV		Average	
		DSC	*p*-value	DSC	*p*-value	DSC	*p*-value	DSC	*p*-value
U-Net [10]	31.1M	0.9439 (0.073)	0.6706	0.9123 (0.055)	0.7918	0.9673 (0.025)	0.2258	0.9412 (0.034)	0.1368
U-Net++ [50]	10.2M	0.9406 (0.074)	0.3564	0.9120 (0.056)	0.0705	0.9694 (0.022)	0.6497	0.9407 (0.037)	0.1147
Attention-Unet [11]	31.9M	0.9363 (0.081)	0.1298	0.9064 (0.057)	3.151×10^{-3}	0.9625 (0.031)	5.338×10^{-3}	0.9351 (0.038)	1.066×10^{-3}
SegNet [51]	29.1M	0.9245 (0.071)	5.438×10^{-4}	0.8880 (0.058)	3.217×10^{-11}	0.9347 (0.063)	5.678×10^{-4}	0.9241 (0.039)	2.139×10^{-5}
Res-Unet [52]	17.6M	0.9147 (0.094)	2.137×10^{-5}	0.8874 (0.064)	1.884×10^{-10}	0.9556 (0.040)	1.032×10^{-5}	0.9193 (0.046)	1.348×10^{-12}
DS-TransU-net[53]	171.4M	0.9277 (0.073)	3.329×10^{-3}	0.8875 (0.068)	9.532×10^{-10}	0.9605 (0.039)	2.359×10^{-3}	0.9252 (0.040)	2.006×10^{-9}
MSU-Net [54]	47.1M	0.9457 (0.061)	0.8852	0.9168 (0.050)	0.4010	0.9696 (0.022)	0.7098	0.9440 (0.033)	0.5559
nnUNet	37.6M	0.9147 (0.094)	2.174×10^{-5}	0.8874 (0.064)	1.887×10^{-10}	0.9556 (0.040)	1.023×10^{-10}	0.9193 (0.046)	1.348×10^{-12}
TransUNet [12]	66.9M	0.9190 (0.082)	7.045×10^{-5}	0.8745 (0.078)	1.332×10^{-14}	0.9524 (0.047)	1.419×10^{-6}	0.9153 (0.050)	1.310×10^{-14}
Swin-Unet [13]	41.5M	0.9381 (0.079)	0.1844	0.9104 (0.049)	0.0223	0.9659 (0.023)	0.0725	0.9381 (0.034)	0.0141
CapNet (ours)	1.53M	**0.9465 (0.068)**		**0.9205 (0.048)**		**0.9706 (0.033)**		**0.9459 (0.036)**	

(b) The T2-SPAIR CMR sequence images

Methods	Params	RV		Myo		LV		Average	
		DSC	*p*-value	DSC	*p*-value	DSC	*p*-value	DSC	*p*-value
U-Net [10]	31.1M	0.8740 (0.103)	0.0482	0.8980 (0.051)	0.1662	0.9476 (0.034)	0.4431	0.9065 (0.049)	0.0233
U-Net++ [50]	10.2M	0.8819 (0.090)	0.1144	0.9038 (0.062)	0.5212	0.9507 (0.044)	0.8295	0.9121 (0.044)	0.1399
Attention-Unet [11]	31.9M	0.8823 (0.111)	0.1570	0.9022 (0.068)	0.4338	0.9478 (0.031)	0.4343	0.9108 (0.030)	0.1274
SegNet [51]	29.1M	0.8657 (0.108)	0.0153	0.8853 (0.078)	0.0235	0.9347 (0.063)	0.0359	0.8952 (0.061)	1.274×10^{-3}
Res-Unet [52]	17.6M	0.8253 (0.124)	5.474×10^{-6}	0.8799 (0.063)	1.162×10^{-3}	0.9431 (0.031)	0.1217	0.8828 (0.057)	8.942×10^{-7}
DS-TransU-net[53]	171.4M	0.8422 (0.107)	6.955×10^{-5}	0.8795 (0.070)	2.291×10^{-3}	0.9380 (0.045)	0.0323	0.8866 (0.045)	2.335×10^{-7}
MSU-Net [54]	47.1M	0.8752 (0.111)	0.0596	0.8988 (0.078)	0.2872	0.9444 (0.054)	0.2889	0.9061 (0.062)	0.0494
nnUNet [55]	37.6M	0.8955 (0.076)	0.6718	0.8728 (0.132)	0.2744	0.9186 (0.084)	0.1174	0.8956 (0.065)	0.1175
TransUNet [12]	66.9M	0.8212 (0.116)	1.092×10^{-6}	0.8372 (0.085)	5.318×10^{-10}	0.9165 (0.063)	2.997×10^{-5}	0.8583 (0.059)	4.108×10^{-14}
Swin-Unet [13]	41.5M	0.8505 (0.110)	6.755×10^{-4}	0.8904 (0.062)	0.0367	0.9383 (0.051)	0.0526	0.8931 (0.052)	1.073×10^{-4}
CapNet (ours)	1.53M	**0.9047 (0.103)**		**0.9097 (0.062)**		**0.9521 (0.043)**		**0.9222 (0.049)**	

Table 4 (continued)

(c) The LGE CMR sequence images

Methods	Params	RV		Myo		LV		Average	
		DSC	*p*-value	DSC	*p*-value	DSC	*p*-value	DSC	*p*-value
U-Net [10]	31.1M	0.9261 (0.064)	0.0573	0.8940 (0.051)	0.1260	0.9415 (0.048)	0.0193	0.9206 (0.044)	0.0118
U-Net++ [50]	10.2M	0.9453 (0.064)	0.8315	0.9009 (0.032)	0.3654	0.9454 (0.056)	0.0985	0.9305 (0.050)	0.2645
Attention-Unet [11]	31.9M	0.9337 (0.082)	0.3021	0.8965 (0.062)	0.2219	0.9471 (0.033)	0.0357	0.9258 (0.048)	0.0831
SegNet [51]	29.1M	0.9035 (0.091)	2.535×10^{-3}	0.8412 (0.042)	1.241×10^{-5}	0.9371 (0.045)	5.186×10^{-4}	0.8910 (0.065)	1.780×10^{-6}
Res-Unet [52]	17.6M	0.8336 (0.117)	3.462×10^{-9}	0.8253 (0.069)	1.218×10^{-9}	0.8631 (0.101)	1.710×10^{-9}	0.8407 (0.080)	6.573×10^{-14}
DS-TransUnet[53]	171.4M	0.9071 (0.089)	6.159×10^{-3}	0.8880 (0.068)	0.0765	0.9409 (0.044)	0.0109	0.9120 (0.054)	2.076×10^{-3}
MSU-Net [54]	47.1M	0.9361 (0.069)	0.3342	0.8880 (0.065)	0.3421	0.9409 (0.031)	0.1293	0.9289 (0.039)	0.1256
nnUNet [55]	37.6M	0.9234 (0.063)	0.0325	0.8951 (0.051)	0.1502	0.9421 (0.048)	0.0237	0.9202 (0.044)	0.0105
TransUNet [12]	66.9M	0.9000 (0.083)	4.316×10^{-4}	0.8549 (0.065)	1.500×10^{-5}	0.9296 (0.055)	4.161×10^{-4}	0.8948 (0.051)	1.142×10^{-7}
Swin-Unet [13]	41.5M	0.9051 (0.082)	1.607×10^{-3}	0.8784 (0.066)	0.0111	0.9341 (0.058)	4.228×10^{-4}	0.9058 (0.052)	8.366×10^{-5}
CapNet (ours)	1.53M	**0.9477 (0.049)**		**0.9124 (0.069)**		**0.9596 (0.026)**		**0.9399 (0.033)**	

The signficance values have been provided and emphasized in the abstract

Limitations and Considerations

In addition to the notable strengths of this study that we have outlined above, there are still some points that we would like to further discuss regarding the limitations of our study. First, as demonstrated in Tables 1, 2, 3, and 4, the *p*-values indicating the performance superiority of the proposed CapNet model over models such as U-Net, nnUnet, and Attention-Unet still remain high. This implies that the proposed CapNet model does not convincingly exhibit higher performance than U-Net or Attention U-Net. However, when comparing the number of parameters, the proposed CapNet model is lightweight with significantly fewer parameters than Attention-Unet, nnUnet, and U-Net, by several orders of magnitude. We propose this CapNet model with the aim of balancing performance and parameter efficiency. The lack of statistical significance in the performance improvement could be due to the sample size in the data. Our future work could involve increasing the sample size, optimizing model parameters further, or exploring additional features to enhance the model's performance and achieve statistical significance.

Besides, setting the hyperparameters α, β, and m for training the initial model is quite challenging. To obtain a good set of hyperparameters, one needs to explore several parameter combinations to estimate a suitable range of values, which can be time-consuming. However, the *p*-value metric shows consistent results across different sets of hyperparameters α, β, and m within certain predefined ranges, as demonstrated in Tables 5 and 6, where α and β are both range from [0,1] and $1 \le m \le 6$. This indicates that the model remains robust and stable as the hyperparameters α, β, and m vary within predefined ranges.

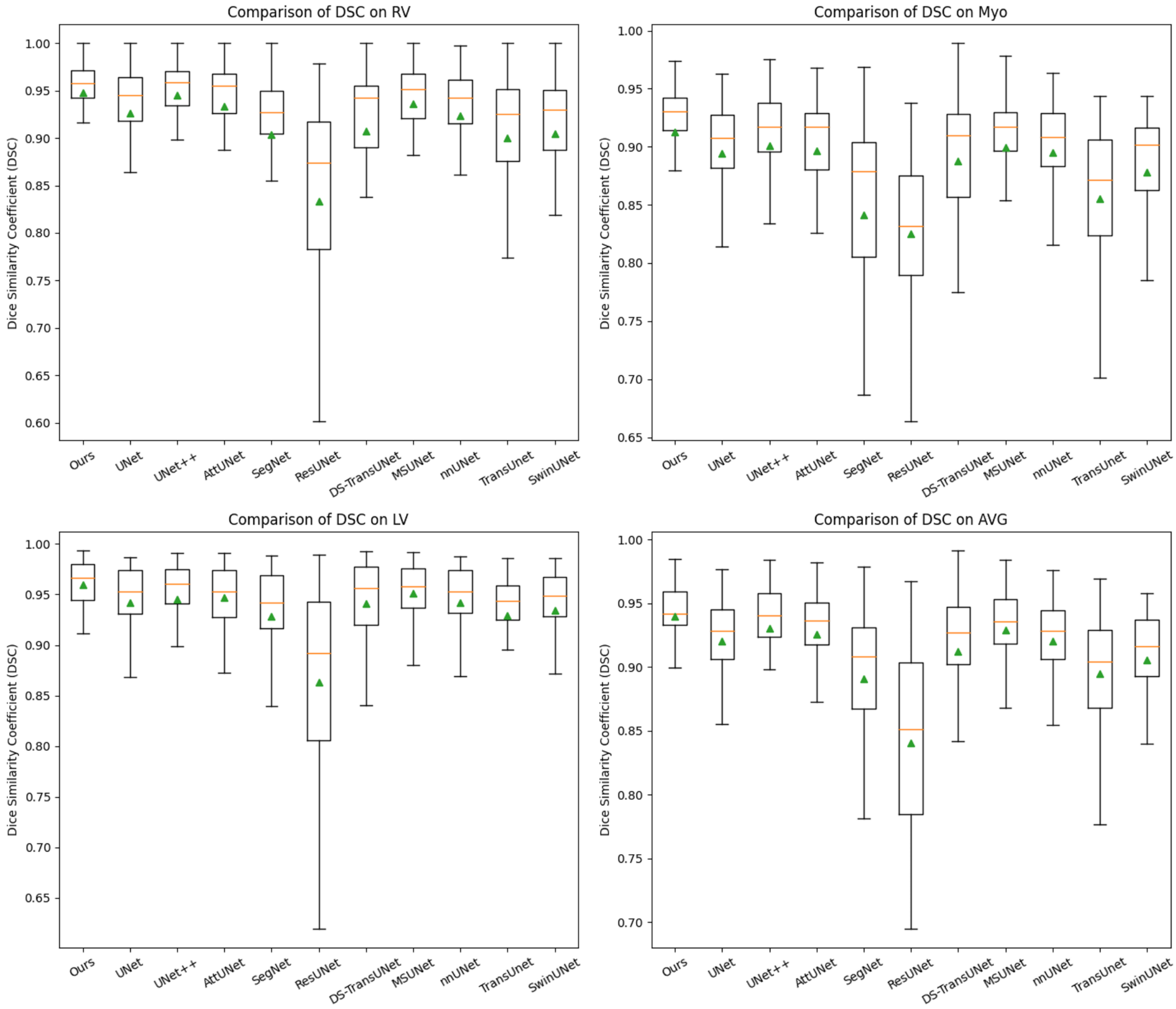


Fig. 16 Boxplots of Dice similarity coefficient on the LGE CMR sequence images in MS-CMR 2019 dataset for multiclass segmentation

Table 5 The experiment comparison between different parameters α and β on the ACDC data and statistical analysis. DSC scores are in mean (standard deviation)

Penalties	RV		Myo		LV		Average	
	DSC	*p*-value	DSC	*p*-value	DSC	*p*-value	DSC	*p*-value
$\alpha = 0.5, \beta = 0.5$	0.9108 (0.077)	0.3846	0.9052 (0.029)	0.7612	0.9551 (0.033)	0.3407	0.9237 (0.033)	0.2764
$\alpha = 0.4, \beta = 0.6$	0.9132 (0.072)	0.3925	0.9078 (0.031)	0.5112	0.9522 (0.026)	0.4082	0.9244 (0.030)	0.2461
$\alpha = 0.2, \beta = 0.8$	0.9153 (0.069)	0.4902	0.9064 (0.031)	0.6395	0.9525 (0.034)	0.3196	0.9247 (0.029)	0.2850
$\alpha = 0.1, \beta = 0.9$	0.9120 (0.063)	0.3145	0.9050 (0.032)	0.5465	0.9533 (0.023)	0.4441	0.9235 (0.030)	0.2431
$\alpha = 0.3, \beta = 0.7$	**0.9234 (0.054)**		**0.9095 (0.031)**		**0.9586 (0.022)**		**0.9305 (0.025)**	

The signficance values have been provided and emphasized in the abstract

Table 6 The experiment comparison between different parameters power m on the ACDC data and statistical analysis. DSC scores are in mean (standard deviation)

Penalties	RV		Myo		LV		Average	
	DSC	p-value	DSC	p-value	DSC	p-value	DSC	p-value
$m = 1$	0.9012 (0.077)	0.1226	0.9000 (0.030)	0.1617	0.9482 (0.037)	0.1462	0.9164 (0.034)	0.0334
$m = \frac{4}{3}$	0.9152 (0.062)	0.4825	0.9010 (0.036)	0.2696	0.9498 (0.033)	0.1936	0.9220 (0.031)	0.1735
$m = 3$	0.9115 (0.071)	0.3432	0.9063 (0.031)	0.6795	0.9578 (0.024)	0.9996	0.9252 (0.030)	0.2431
$m = 5$	0.8996 (0.088)	0.1805	0.9035 (0.036)	0.4012	0.9491 (0.031)	0.1157	0.9174 (0.040)	0.0904
$m = 6$	0.9031 (0.068)	0.1437	0.8962 (0.032)	0.0675	0.9471 (0.037)	0.1104	0.9154 (0.031)	0.0179
$m = 2$	**0.9234 (0.054)**		**0.9095 (0.031)**		**0.9586 (0.022)**		**0.9305 (0.025)**	

The signficance values have been provided and emphasized in the abstract

Table 7 Comparison between the proposed Tversky Shape Power Distance (TSPD) and other losses in the case of (a) binary segmentation and (b) multiclass segmentation

Penalties	Endocardium				Epicardium			
	DSC	p-value	IoU	p-value	DSC	p-value	IoU	p-value
Dice	0.9301 (0.085)	0.3638	0.8807 (0.111)	0.5562	0.9589 (0.033)	0.9407	0.9223 (0.056)	0.9570
BCE-Dice	0.9383 (0.028)	0.7869	0.8868 (0.050)	0.7856	0.9475 (0.035)	0.0346	0.9046 (0.054)	0.0613
Tversky [38]	0.9280 (0.111)	0.3707	0.8773 (0.115)	0.4159	0.9546 (0.039)	0.5083	0.9156 (0.065)	0.5336
Focal Tversky [39]	0.9370 (0.043)	0.8545	0.8874 (0.068)	0.8304	0.9457 (0.047)	0.1473	0.9008 (0.075)	0.1336
AC-Focal [35]	0.9398(0.040)	0.9966	0.8890 (0.066)	0.9694	0.9500 (0.048)	0.2655	0.9083 (0.078)	0.2842
Shape distance [41]	0.9341 (0.064)	0.9173	0.8873 (0.085)	0.9591	0.9502 (0.046)	0.3457	0.9091 (0.076)	0.3772
TSPD (ours)	**0.9400 (0.043)**		**0.8895 (0.069)**		**0.9593 (0.027)**		**0.9230 (0.047)**	

(a) *On the Sunnybrook data with binary segmentation*

Penalties	RV		Myo		LV		Average	
	DSC	p-value	DSC	p-value	DSC	p-value	DSC	p-value
Dice	0.9018 (0.095)	0.1526	0.9012 (0.032)	0.1912	0.9461 (0.040)	0.0685	0.9163 (0.034)	0.0334
BCE-Dice	0.9098 (0.079)	0.3342	0.9011 (0.030)	0.1617	0.9424 (0.037)	0.1462	0.9178 (0.041)	0.0389
Tversky [38]	0.9116 (0.069)	0.3780	0.8896 (0.034)	0.0057	0.9468 (0.039)	0.0828	0.9160 (0.031)	0.0159
Focal Tversky [39]	0.9025 (0.078)	0.1675	0.8999 (0.031)	0.1886	0.9437 (0.042)	0.0554	0.9154 (0.037)	0.0344
AC-Focal [35]	0.9135 (0.069)	0.4828	0.8890 (0.037)	0.0086	0.9515 (0.032)	0.2909	0.9230 (0.030)	0.2431
Shape distance [41]	0.9134 (0.065)	0.4356	0.9054 (0.032)	0.6436	0.9520 (0.031)	0.3506	0.9236 (0.031)	0.2949
TSPD (ours)	**0.9234 (0.054)**		**0.9095 (0.031)**		**0.9586 (0.022)**		**0.9305 (0.024)**	

(b) *On the ACDC data with multiclass segmentation*

The signficance values have been provided and emphasized in the abstract

Table 8 Comparison between the PCA-PSA++ architecture and other architectures on the ACDC data. Dice (DSC) scores are in mean (standard deviation)

Methods	**GFLOPS**	**Params**	**DiceRV**	**DiceMYO**	**DiceLV**	**DiceAVG**
PASPP [56]	5.28	0.84	0.9069 (0.080)	0.9055 (0.031)	0.9510 (0.033)	0.9209 (0.035)
ASPP[57]	26.14	5.11	0.9123 (0.073)	0.9071 (0.032)	0.9542 (0.021)	0.9245 (0.028)
CBAM [58]	**0.31**	**0.04**	0.9085 (0.075)	0.9038 (0.032)	0.9540 (0.031)	0.9221 (0.032)
PCA-PSA [31]	2.07	0.32	0.9150 (0.066)	0.9064 (0.030)	0.9546 (0.031)	0.9253 (0.028)
PCA-PSA++ (ours)	2.08	0.34	**0.9234 (0.054)**	**0.9095 (0.031)**	**0.9586 (0.022)**	**0.9305 (0.024)**

The signficance values have been provided and emphasized in the abstract

Table 9 Comparison of mean DSC values of RV, MYO, LV, and their average on ACDC dataset with kernel size $k = 7$ and the increasing dilation d. DSC scores are in mean (standard deviation)

Operator	Metric		$d = (1)$	$d = (1, 2)$	$d = (1, 2, 3)$	$d = (1, 2, 3, 4)$
Conv2d kernel size = (k, k)	DSC	RV	0.9069 (0.082)	0.9098 (0.079)	0.9090 (0.078)	0.9077 (0.080)
		MYO	0.9014 (0.037)	0.9022 (0.037)	0.9059 (0.041)	0.9024 (0.037)
		LV	0.9496 (0.034)	0.9489 (0.034)	0.9541 (0.022)	0.9517 (0.033)
		Average	0.9200 (0.034)	0.9206 (0.034)	0.9226 (0.031)	0.9204 (0.034)
Conv2d kernel size = $(1, k) + (k, 1)$	DSC	RV	0.9060 (0.083)	0.9075 (0.082)	0.9180 (0.066)	0.9166 (0.069)
		MYO	0.9048 (0.030)	0.9072 (0.032)	0.9064 (0.033)	0.9049 (0.032)
		LV	0.9539 (0.023)	0.9538 (0.023)	0.9525 (0.025)	0.9537 (0.023)
		Average	0.9216 (0.035)	0.9228 (0.034)	0.9256 (0.032)	0.9251 (0.032)
DW Conv kernel size = $(1, k) + (k, 1)$	DSC	RV	0.9123 (0.065)	0.9132 (0.064)	**0.9234 (0.054)**	0.9129 (0.065)
		MYO	0.9064 (0.032)	0.9067 (0.032)	**0.9095 (0.031)**	0.9041 (0.033)
		LV	0.9542 (0.026)	0.9547 (0.023)	**0.9586 (0.022)**	0.9507 (0.028)
		Average	0.9243 (0.027)	0.9249 (0.026)	**0.9305 (0.024)**	0.9225 (0.030)

The signficance values have been provided and emphasized in the abstract

Table 10 Comparison of DSC values of RV, MYO, LV, and their average value on the ACDC dataset with dilation $d = (1, 2, 3)$ and the increasing kernel size k. DSC scores are in mean (standard deviation)

Operator	Metric		$k = 3$	$k = 5$	$k = 7$
Conv2d kernel size = (k, k)	DSC	RV	0.9129 (0.064)	0.9117 (0.063)	0.9060 (0.081)
		MYO	0.9068 (0.030)	0.9051 (0.032)	0.9044 (0.036)
		LV	0.9525 (0.025)	0.9560 (0.026)	0.9520 (0.034)
		Average	0.9246 (0.028)	0.9243 (0.027)	0.9207 (0.035)
Conv2d kernel size = $(1, k) + (k, 1)$	DSC	RV	0.9120 (0.065)	0.9127 (0.066)	0.9164 (0.069)
		MYO	0.9052 (0.030)	0.9033 (0.032)	0.9049 (0.035)
		LV	0.9554 (0.031)	0.9507 (0.034)	0.9569 (0.024)
		Average	0.9241 (0.027)	0.9222 (0.030)	0.9261 (0.031)
DW Conv kernel size = $(1, k) + (k, 1)$	DSC	RV	0.9191 (0.058)	0.9178 (0.066)	**0.9234 (0.054)**
		MYO	0.9064 (0.032)	0.9041 (0.026)	**0.9095 (0.031)**
		LV	0.9544 (0.022)	0.9549 (0.029)	**0.9586 (0.022)**
		Average	0.9266 (0.026)	0.9256 (0.026)	**0.9305 (0.024)**

The signficance values have been provided and emphasized in the abstract

Conclusion

We have presented a new network model and a new loss for image segmentation of cardiovascular magnetic resonance images. The network is trained end to end with a quite small number of parameters. The proposed Priority Mixer block and Depthwise-Focus block with attention mechanism are applied for better learning information from anatomic organs which are variable in size and shape during cardiac phases. In addition, we propose a new loss called Tversky Shape Power Distance based on the dissimilarity in shape distances between the mask and the predicted label. Extensive experiments and ablation studies have been performed to prove the dominance of both the proposed architecture and the proposed loss function.

Funding This research is funded by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.05-2021.34.

Data Availability All data used are from open benchmarks associated with references.

References

1. Campello VM, Gkontra P, Izquierdo C, Martin-Isla C, Sojoudi A, Full PM, Maier-Hein K, Zhang Y, He Z, Ma J, Parreno M.. Multi-centre, multi-vendor and multi-disease cardiac segmentation: The mms challenge. IEEE Trans Med Imaging 40:3543-3554, 2021
2. Chen C, Qin C, Qiu H, Tarroni G, Duan J, Bai W, Rueckert D.Deep learning for cardiac image segmentation: A review. Front Cardiovasc Med 7:25, 2020
3. Tran TT, Pham VT, Lin C, Yang HW, Wang YH, Shyu KK, Tseng WY, Su MY, Lin LY, Lo MT, Empirical mode decomposition and monogenic signal-based approach for quantification of myocardial infarction from mr images. IEEE J Biomed Health Inform 23:731-743, 2019
4. Singh Samant S, Chauhan A, Dn J, Singh V. Glomerulus detection using segmentation neural networks. J Digit Imaging 36:1633-1642, 2023

5. Dong H, Yang G, Liu F, Mo Y, Guo Y. Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks. In Medical Image Understanding and Analysis: 21st Annual Conference, MIUA 2017, Edinburgh, UK, 506-517, 2017
6. Radau P, Lu Y, Connelly K, Paul G, Dick AJ, Wright GA. Evaluation framework for algorithms segmenting short axis cardiac MRI. The MIDAS Journal, 2009
7. Wang X, Wang F, Niu Y. Two-Stage CNN Whole Heart Segmentation Combining Image Enhanced Attention Mechanism and Metric Classification. J Digit Imaging 36:124-142, 2023
8. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, 3431-3440, 2015
9. Tran PV. A fully convolutional neural network for cardiac segmentation in short-axis MRI. arXiv preprint arXiv:1604.00494. 2016
10. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation.In Medical image computing and computer-assisted intervention-MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III1 18: 234-241, 2015
11. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B, Glocker B. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999, 2018
12. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306, 2021
13. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, Wang M. Swin-unet: Unet-like pure transformer for medical image segmentation. In European conference on computer vision, 205-218, 2022
14. Huang X, Deng Z, Li D, Yuan X, Fu Y. MISSFormer: an effective transformer for 2D medical image segmentation. IEEE Trans Med Imaging 42: 1484-1494, 2023
15. Zotti C, Luo Z, Lalande A, Jodoin PM. Convolutional neural network with shape prior applied to cardiac MRI segmentation. IEEE J Biomed Health Inform 23: 1119- 1128, 2018
16. Baumgartner CF, Koch LM, Pollefeys M, Konukoglu E. An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation.In Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers 8: 111-119, 2018
17. Patravali J, Jain S, Chilamkurthy S. 2D-3D fully convolutional neural networks for cardiac MR segmentation. In Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers 8: 130- 139, 2018
18. Cui H, Yuwen C, Jiang L, Xia Y, Zhang Y. Multiscale attention guided U-Net architecture for cardiac segmentation in short-axis MRI images. Comput Methods Programs Biomed 206, p. 106142, 2021
19. Chen C, Bai W, Davies RH, Bhuva AN, Manisty CH, Augusto JB, Moon JC, Aung N, Lee AM, Sanghvi MM, Fung K. Improving the generalizability of convolutional neural network-based segmentation on CMR images. Front Cardiovasc Med 7: 105, 2020
20. Wang Z, Peng Y, Li D, Guo Y, Zhang B. MMNet: A multi-scale deep learning network for the left ventricular segmentation of cardiac MRI images. Appl Intell 52:5225- 5240, 2022
21. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020
22. Li C, Wang L, Li Y. Transformer and group parallel axial attention co-encoder for medical image segmentation. Sci Rep 12: 16117, 2022
23. Wang H, Zhu Y, Green B, Adam H, Yuille A, Chen LC. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In European conference on computer vision, 108-126, 2020
24. Lin X, Yu L, Cheng KT, Yan Z. Batformer: Towards boundary-aware lightweight transformer for efficient medical image segmentation. IEEE J Biomed Health Inform, 2023
25. Rahman MM, Marculescu R. Medical image segmentation via cascaded attention decoding. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 6222-6231, 2023
26. Dinh BD, Nguyen TT, Tran TT, Pham VT. 1M parameters are enough? A lightweight CNN-based model for medical image segmentation. In 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 1279-1284, 2023
27. Valanarasu JM, Patel VM. Unext: Mlp-based rapid medical image segmentation network. In International conference on medical image computing and computer-assisted intervention, 23-33, 2022
28. Chollet F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, 1251-1258, 2017
29. Hua BS, Tran MK, Yeung SK. Pointwise convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 984-993, 2018
30. Jungnickel D, Jungnickel D. The greedy algorithm. Graphs, networks and algorithms. 129-53, 1999
31. Le TV, Tran TT, Pham VT. Attention ConvMixer Model and Application for Fish Species Classification. EAI Endorsed Trans Ind 10, 2023
32. Vu NT, Pham VT, Pham VT, Tran TT. CPA-Unet: A Solution for Left Ventricle Segmentation from Magnetic Resonance Images. In 2023 International Conference on System Science and Engineering (ICSSE), 33-38, 2023
33. Yan Q, Wang B, Gong D, Luo C, Zhao W, Shen J, Shi Q, Jin S, Zhang L, You Z. COVID-19 chest CT image segmentation–a deep convolutional neural network solution. arXiv preprint arXiv:2004.10987, 2020
34. Trockman A, Kolter JZ. Patches are all you need?. arXiv preprint arXiv:2201.09792, 2022.
35. Trinh MN, Nham DHN, Pham VT, Tran TT. An attention-PiDi-UNet and focal active contour loss for biomedical image segmentation. In 2022 RIVF International Conference on Computing and Communication Technologies (RIVF),635-640, 2022
36. Tragakis A, Kaul C, Murray-Smith R, Husmeier D. The fully convolutional transformer for medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision,3660-3669, 2023
37. Chen X, Williams BM, Vallabhaneni SR, Czanner G, Williams R, Zheng Y. Learning active contour models for medical image segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 11624- 11632, 2019
38. Salehi SS, Erdogmus D, Gholipour A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. In International workshop on machine learning in medical imaging 379-387, 2017
39. Abraham N, Khan NM. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019),683-687, 2019
40. Nham DHN, Trinh MN, Nguyen VD, Pham VT, Tran TT. An EffcientNet-encoder U-Net Joint Residual Refinement Module with Tversky-Kahneman Baroni-Urbani-Buser loss for biomedical image Segmentation. Biomed Signal Process Control 83:, p. 104631, 2023

41. Pham VT, Tran TT, Wang PC, Chen PY, Lo MT. EAR-UNet: A deep learning-based approach for segmentation of tympanic membranes from otoscopic images. Artif Intell Med 115:102065, 2021
42. Krinidis S, Chatzis V. Fuzzy energy-based active contours. IEEE Trans. Image Process 18:2747-55, 2009
43. Kim B, Ye JC. Mumford-Shah loss functional for image segmentation with deep learning. IEEE Trans Image Process 29:1856-1866, 2019
44. Demš¡ar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7: 1-30, 2006
45. Bernard O, Lalande A, Zotti C, Cervenansky F, Yang X, Heng PA, Cetin I, Lekadir K, Camara O, Ballester MA, Sanroma G. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?. IEEE Trans Med Imaging 37:2514-2525, 2018
46. Gao S, Zhou H, Gao Y, Zhuang X. BayeSeg: Bayesian modeling for medical image segmentation with interpretable generalizability. Med Image Anal 89: 102889, 2023
47. Zhuang X. Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE Trans Pattern Anal Mach Intell 41: 2933-2946, 2018
48. Dozat T. Incorporating nesterov momentum into adam. In Proceedings of the 4th International Conference on Learning Representations, 2016
49. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision 128:336-359, 2019
50. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. Unet++: A nested u-net architecture for medical image segmentation. InDeep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4:3-11, 2018
51. Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 39: 2481-2495, 2017
52. Diakogiannis FI, Waldner F, Caccetta P, Wu C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS J Photogramm Remote Sens 162:94-114, 2020
53. Lin A, Chen B, Xu J, Zhang Z, Lu G, Zhang D. Ds-transunet: Dual swin transformer u-net for medical image segmentation. IEEE T Instrum Meas 71:1-15, 2022
54. Su R, Zhang D, Liu J, Cheng C. Msu-net: Multi-scale u-net for 2d medical image segmentation. Front Genet 12:639930, 2021
55. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 18: 203-211, 2021
56. Yan Q, Wang B, Gong D, Luo C, Zhao W, Shen J, Ai J, Shi Q, Zhang Y, Jin S, Zhang L. COVID-19 chest CT image segmentation network by multi-scale fusion and enhancement operations. IEEE Trans Big Data 7: 13-24, 2021
57. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell 40: 834-848, 2017
58. Woo S, Park J, Lee JY, Kweon IS. Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV) , 3-19, 2018

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.