CAMS: Convolution and Attention-Free Mamba-based Cardiac Image Segmentation

Abbas Khan[1,2] Muhammad Asad[1,2] Martin Benning[3] Caroline Roney[1,2] Gregory Slabaugh[1,2]
[1] Queen Mary University of London [2] Digital Environment Research Institute
[3] University College London

Abstract

Convolutional Neural Networks (CNNs) and Transformer-based self-attention models have become the standard for medical image segmentation. This paper demonstrates that convolution and self-attention, while widely used, are not the only effective methods for segmentation. Breaking with convention, we present a ***C****onvolution and self-****A****ttention-free* ***M****amba-based semantic* ***S****egmentation* ***Net****work named CAMS-Net. Specifically, we design Mamba-based* Channel Aggregator *and* Spatial Aggregator*, which are applied independently in each encoder-decoder stage. The Channel Aggregator extracts information across different channels, and the Spatial Aggregator learns features across different spatial locations. We also propose a* ***L****inearly* ***I****nterconnected* ***F****actorized* ***M****amba (LIFM) block to reduce the computational complexity of a Mamba block and to enhance its decision function by introducing a non-linearity between two factorized Mamba blocks. Our model outperforms the existing state-of-the-art CNN, self-attention, and Mamba-based methods on CMR and M&Ms-2 Cardiac segmentation datasets, showing how this innovative, convolution, and self-attention-free method can inspire further research beyond CNN and Transformer paradigms, achieving linear complexity and reducing the number of parameters. Source code and pre-trained models are available at: https://github.com/kabbas570/CAMS-Net.*

1. Introduction

Image segmentation is an essential part of Cardiac image analysis [33]. It can help quantify the size and shape of different regions of interest, such as left ventricle (LV), right ventricle (RV), left atrium (LA), right atrium (RA), myocardium (MYO), useful for monitoring disease progression, prognosis and supporting computer-aided intervention [4]. Manual segmentation is considered a gold standard; however, with the developments in artificial intelligence, recent research has been focused on developing automatic methods for faster, cheaper, and reproducible results [17, 26].

Convolutional neural networks (CNNs) and transformer-based self-attention mechanisms have significantly evolved the landscape of medical image segmentation [51]. Although CNNs have been the commonly used choice [1], current literature suggests that self-attention-based methods produce better results than CNN architectures [23] due to their global receptive field, ability to model long-range dependencies, and the dynamic weights mechanism [8]. CNN-based methods have been criticized for their limited receptive field [40], limited ability to effectively capture long-range dependencies, and their bias toward recognizing textures rather than shapes [12]. However, attention-based methods are computationally expensive compared to CNNs due to their quadratic complexity [36] and excessive memory requirements [39]. Recent research work has focused on reducing the computational complexity of attention-based methods while maintaining accuracy, including efficient additive attention [44], efficient self-attention [11], and separable self-attention [35]. Hybrid CNN-transformer-based segmentation methods have also been a recent trend that harnesses relative strengths of CNN and self-attention [22, 46]. These methods combine CNN and self-attention to capture local and global features while reducing self-attention's computational complexity.

Recently, Mamba has gained prominence in the computer vision field and integrates Gated MLP [34] into the State Space Model (SSM) of H3 [6]. Readers are encouraged to refer to [37, 52] for a more comprehensive understanding of this topic. SSMs [14] such as Mamba [13] are considered as a potential replacement for transformers because they can capture long-range dependencies while maintaining linear computation complexity. Several architectures have been proposed to show the power of Mamba for computer vision tasks, including Vision Mamba [55], Visual Mamba (VMamba) [29], ZigMa [18], and medical image segmentation is no exception [42, 50]. The majority of existing encoder-decoder medical image segmentation

architectures inspired by Mamba [13] and vision transformers [8] differ in how the convolution layers, self-attention, and Mamba blocks are arranged.

In this paper, we go beyond the arrangements of these blocks and propose a novel convolution and attention-free CAMS-Net for medical image segmentation. We propose Mamba-based spatial and channel aggregators to extract information across different channels and spatial locations, along with a Linearly Interconnected Factorized Mamba (LIFM) block to further reduce the computational complexity and enhance its decision function. Unlike CNN-based segmentation networks, it excels at capturing global features while surpassing self-attention-based methods by modeling long-range dependencies with linear rather than quadratic complexity. Compared to other Mamba-based segmentation networks, CAMS-Net stands out for its convolutional-free design, eliminating the need for hybrid architectures. Along with these innovations, the CAMS-Net outperforms existing networks, making it a more efficient and effective solution for medical image segmentation.

2. Related Work

UNet [41] is a pioneering network architecture for medical image segmentation, and several subsequent architectures, including ResUNet [7], UNet++ [54], have extended its initial formulation. These networks use an encoder-decoder design, where the encoder extracts information from images, and the decoder reconstructs the segmentation map. Skip connections [9] mitigate the vanishing gradient problem and reuse the features from the encoder side. With the advent of the vision transformer [8], self-attention-based methods have become popular in medical image segmentation to overcome the limitations of CNN-based pipelines. These methods include Swin-UNet [3], which replaces the convolutional layers with Swin-Transformer blocks [30]. The Swin-UNet also follows a U-shaped architecture, where the encoder utilizes a hierarchical Swin Transformer with shifted windows to extract context features and a symmetric decoder with patch-expanding layers for upsampling. The UNEt TRansformers (UNETR) [16] uses a transformer-based encoder to learn sequence representations of the input images, allowing the network to capture global multi-scale information and a CNN-based decoder for localized information. Similar to Swin-UNet [3], Swin-UNETR [15] is also built on a hierarchical Swin transformer. However, it has a hybrid architecture that only uses a Swin transformer in the encoder to extract features and a CNN-based decoder to generate the segmentation map.

Mamba-UNet [50] incorporates the VMamba-based [29] encoder-decoder structure with UNet. The Cross-Scan Module from VMamba scans the input image in four ways to integrate information from all other locations for each element of the features. Mamba-UNet utilizes these VMamba [29] blocks throughout the U-shaped architecture to capture semantic contexts from intensity images. Vision Mamba UNet (VM-UNet) [42] extends Vision Mamba [55] using foundation blocks named Visual State Space. Its asymmetrical encoder-decoder structure leverages the power of SSMs to capture contextual information while maintaining linear computational complexity.

Nevertheless, these hybrid methods address the challenges posed by self-attention and CNNs and utilize both local and global features in dense prediction tasks, like segmentation. However, convolution-free methods have become a recent trend in computer vision, and some of these approaches have tried to utilize self-attention-based architectures only. For example, Kim et al. [24] proposed ReSTR for referring image segmentation, where transformer-based encoders extract features from each modality, image, and text, followed by coarse-to-fine segmentation decoder transforms to reconstruct the output from fused features. Karimi et al. [21] proposed a convolution-free 3D network for medical image segmentation. The 3D image block is divided into n^k patches (k=3, or 5) and computes a 1D embedding for each patch. Their method predicts the center patch of the block using self-attention between patch embeddings. MLP-Mixer [47] proposed an alternative architecture for image classification tasks built solely on multilayer perceptrons (MLPs). The channel- and token-mixing MLPs learn the per-location features and between different spatial locations (tokens), respectively.

Although these most recent models attempt to overcome challenges posed by CNNs, they are based on self-attention with quadratic computational complexity and high memory requirements. This paper introduces a convolution-free and self-attention-free model to mitigate the limitations of convolution-based and self-attention-based architectures while maintaining the benefits that self-attention brings, i.e., the global receptive field, dynamic weight mechanism, and long-range dependencies at the expense of linear complexity. The contributions of this work are:

1. To the best of our knowledge, we are the first to propose a convolution and self-attention-free Mamba-based segmentation network, CAMS-Net.

2. We propose a Linearly Interconnected Factorized Mamba (LIFM) block to reduce the trainable parameters of Mamba and improve its non-linearity. LIFM implements a weight-sharing strategy for different scanning directions, specifically for the two scanning direction strategies of vision Mamba [55], to reduce the computational complexity further whilst maintaining accuracy.

3. We propose the Mamba Channel Aggregator (MCA) and Mamba Spatial Aggregator (MSA) and demon-

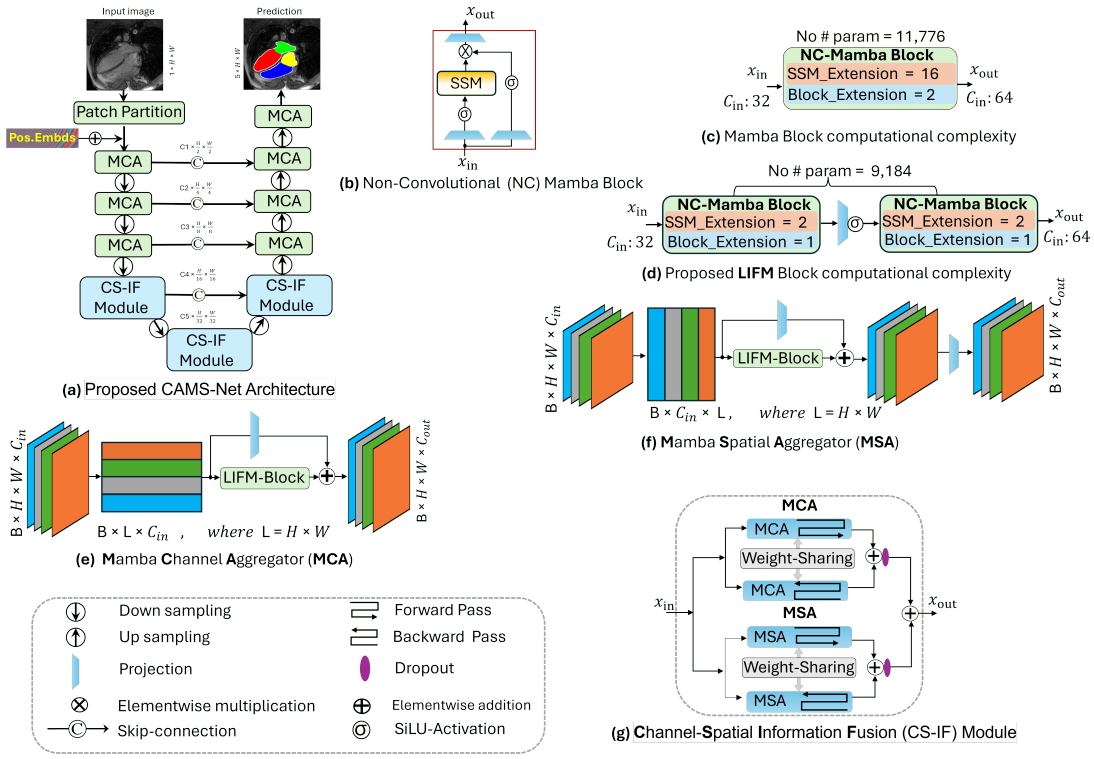

Figure 1. (a) Overall architecture of proposed CAMS-Net, (b) the Non-Convolutional (NC) Mamba Block without local convolution, (c) comparison of proposed LIFM block with original Mamba, (e) Mamba Channel Aggregator (MCA), (f) Mamba Spatial Aggregator (MSA), and (g) the Channel-Spatial Information Fusion (CS-IF) Module.

strate how they can learn information along the channel and spatial dimensions of the features, respectively.

4. Extensive experimental validation, including ablation studies, is conducted to showcase the efficacy of our proposed model. Our proposed CAMS-Net outperforms existing state-of-the-art segmentation models on the CMR and the Multi-Disease, Multi-View, and Multi-Center (M&Ms-2) segmentation datasets, including pure CNN, self-attention, and hybrid self-attention, as well as methods using the original Mamba-based architecture combined with CNNs.

3. Methodology

Figure 1 (a) shows the proposed convolution and a self-attention-free segmentation network, CAMS-Net. The input image is transformed into non-overlapping patches with a patch size of 2 × 2, reducing the in-plane spatial resolution by 2, and a linear embedding layer to project the features into dimension *C1 = 64*. It also incorporates sinusoidal positional embeddings to encode spatial context information, enabling the encoder to understand the relative positions of different regions within the image. The features are downsampled at each encoder's stage using a 2×2 average pooling layer. In the next encoder stage and bottleneck, we implement the CS-IF module, allowing the model to learn richer features along channel and spatial dimensions.

On the decoder side, the features are upsampled at each stage using a bilinear interpolation window of 2×2 to match the output dimension, followed by the CS-IF module in the first stage after the bottleneck and MCA in all other decoder stages. The skip connections [9] are also implemented at each encoder-decoder stage to reuse the features and for faster convergence. Finally, a five-class segmentation map (one for each class, LA, RA, LV, RA, and background) is generated, followed by a Softmax activation. This section will explain the components of the CAMS-Net.

3.1. Factorized Mamba with LIFM Block

Inspired by deep convolutional neural networks [45], where a stack of two 3 × 3 convolution filters have an effective receptive field of 5 × 5, we propose the idea of factorized Mamba, which makes the decision function more discriminative and also reduces the number of parameters. The 'Mamba block expansion factor' *(E)* and the 'SSM state expansion factor' *(D)* control the overall complexity of the Mamba block. More specifically, *E* expands the dimensions

of the Mamba block using linear layers with learned weights $\mathbf{W}_1$, and $\mathbf{W}_2$, while D projects the dimension within the SSM. We implemented the Mamba block with various E and D factors and analyzed their computational complexity, shown in Table 1 of the supplementary material. In the Mamba block, most of the parameters stem from E, with minimal increment from D. The majority of the Mamba-based networks use the default SSM and Mamba block extension, shown in Figure 1 (c), which is computationally expensive, and a single Mamba block brings 11,776 trainable parameters (for c_{in}=32 and c_{out}=64). Mathematically, it can be represented as,

$$\mathbf{x}_{\text{out}} = \textbf{NC-MambaBlock}(\mathbf{x}_{\text{in}}, D, E), \tag{1}$$

with,

$$\mathbf{x}_{\text{out}} = \mathbf{W}_3\Big(\sigma(\mathbf{W}_2\mathbf{x}_{in}) \odot \mathbf{SSM}(\sigma(\mathbf{W}_1\mathbf{x}_{in}))\Big), \tag{2}$$

where $\mathbf{W}_1$,$\mathbf{W}_2$,$\mathbf{W}_3$ are learnable weights for linear layers shown in Figure 1 (b) used for input x_{in} projection, $\odot$ represents element-wise multiplication and σ SiLU activation [10].

Our factorized Mamba block splits the extension parameters for SSM and Mamba, shown in Figure 1 (d). We also add a linear layer followed by SiLU activation [10] between two Mamba blocks to add more non-linearity and named it the Linearly Interconnected Factorized Mamba (LIFM) block, used throughout our proposed architecture. A single factorized Mamba block has 4,608 parameters (for c_{in}=32 and c_{out}=64), and the proposed LIFM-Block requires only 9,184 trainable parameters. For the first factorized Mamba block and linear layer, C_{in} = C_{out} = 32, and for the second factorized Mamba block, C_{out} = 64. Mathematically, we can represent the LIFM Block as

$$\begin{aligned}
\mathbf{x}_1 &= \textbf{NC-MambaBlock}(\mathbf{x}_{\text{in}}, D_1, E_1) && (3)\\
\mathbf{x}_2 &= \sigma(\mathbf{W}_{fm}\mathbf{x}_1) && (4)\\
\mathbf{x}_{\text{out}} &= \textbf{NC-MambaBlock}(\mathbf{x}_2, D_2, E_2) && (5)
\end{aligned}$$

where, D_1 = D_2 = 2 , and E_1 = E_2 = 1, $\mathbf{W}_{fm}$ represents the linear layer between factorized two Mamba blocks.

Empirically, we also found that a large Mamba block can easily overfit the data and increase the overall computational burden of the network. So, we factorized the larger Mamba blocks at each stage and used two consecutive relatively smaller ones. This factorized approach reduces the number of trainable parameters and helps the network to increase its non-linearity to learn more complex patterns and representations in the data.

3.2. Mamba Channel Aggregator

The Mamba channel aggregator (MCA) aims to learn cross-channel information, as shown in Figure 1 (e), learning the per-location features at different channels. Similar to a UNet structure, the number of channels is increased as $\{64, 128, 256, 512, 1024\}$ at each encoder stage and decreased as $\{512, 256, 128, 64\}$ at each decoder stage. For the channel aggregator, the incoming features $\mathbb{R}^{B\times C\times H\times W}$ are reshaped to $\mathbb{R}^{B\times L\times C}$, where $L = H \times W$. Then, the input is divided into two branches where, in one branch, the LIFM Block is applied, and the second branch acts as a residual connection, where a linear layer is used, followed by an element-wise addition operation with the features of the first branch. Mathematically, it can be represented as,

$$x_{\text{out}} = \overline{\mathbf{f_1}}\Big(\textbf{LIFM\_Block}\big(\mathbf{f_1}(x_{\text{in}})\big) \oplus \mathbf{W}_c(\mathbf{f_1}(x_{\text{in}}))\Big), \tag{6}$$

where, $\mathbf{f}_1 : \mathbb{R}^{B\times C\times H\times W} \rightarrow \mathbb{R}^{B\times L\times C}$ represents a reshaping function and $\overline{\mathbf{f_1}} : \mathbb{R}^{B\times L\times C} \rightarrow \mathbb{R}^{B\times C\times H\times W}$ performs the inverse operation, $\mathbf{W}_c$ is the residual linear layer of MCA, and $\oplus$ represents element-wise addition.

3.3. Mamba Spatial Aggregator

As shown in Figure 1 (f), the Mamba spatial aggregator (MSA) aims to learn information about different spatial locations and enables communication amongst them. The spatial aggregator's computational complexity depends on the features' spatial dimensions, so it is only used for the lower-dimensional features of the U-shaped network. More specifically, it is used in the bottleneck, one encoder stage before the bottleneck, and one decoder stage after the bottleneck, shown in Figure 1 (a). For the spatial aggregator, the incoming features $\mathbb{R}^{B\times C\times H\times W}$ are reshaped to $\mathbb{R}^{B\times C\times L}$. The features follow the same protocol as that of MCA, and finally, a linear layer is used either to expand (in the encoder) or to compress (in the decoder) the number of channels. In mathematical terms,

$$\mathbf{x}_{\text{out}} = \mathbf{W}_{ci}\overline{\mathbf{f_2}}\Big(\textbf{LIFM\_Block}\big(\mathbf{f_2}(\mathbf{x}_{\text{in}})\big)\oplus\mathbf{W}_s\big(\mathbf{f_2}(\mathbf{x}_{\text{in}})\big)\Big). \tag{7}$$

Here, $\mathbf{f}_2 : \mathbb{R}^{B\times C\times H\times W} \rightarrow \mathbb{R}^{B\times C\times L}$ represents a reshaping function and $\overline{\mathbf{f_2}} : \mathbb{R}^{B\times C\times L} \rightarrow \mathbb{R}^{B\times C\times H\times W}$ does the inverse, $\mathbf{W}_S$ is the residual linear layer of MSA, and $\mathbf{W}_{ci}$ is a linear layer that either increases or decreases the number of channels in MSA, to match with MCA.

3.4. Bidirectional Information Learning

Inspired by Vision Mamba [55], we implemented both MCA and MSA using a bidirectional scanning arrangement scheme, shown in Figure 1 of the supplementary material. We incorporated the bidirectional SSMs to make the network spatially aware. Unlike Vision Mamba, we found that sharing the weights for two-direction schemes results in better average performance and also lowers computational complexity, as shown in the Table 3 of the ablation study. We also experimented with a multi-directional scanning arrangement, such as a four-directional [29] and an eight-

directional scheme [18]. However, the bidirectional scanning scheme augmented with the proposed weight-sharing strategy is the best practice for the task at hand due to the smaller dataset and the method's reduced complexity.

3.5. Channel-Spatial Information Fusion Module

The Channel-Spatial Information Fusion (CS-IF) module comprises the MCA and MSA and merges the information extracted along channel and spatial dimensions, depicted in Figure 1 (g). The incoming features are passed to the MCA and MSA, where each aggregation learns the features in both the forward and backward scanning directions using the same instance of the corresponding aggregation, making it shareable in utilizing the weights. An element-wise addition operation sums up the output of both passes and to avoid overfitting, a dropout of 0.1 is applied to the output of each aggregator.

4. Experimental Validation

This section provides details of the datasets, implementation, and our experimental results showing our approach's superior performance compared to the state-of-the-art.

4.1. Datasets Description

We use the following two datasets for experimental validation of our method. The CMRxsegmentation dataset provides a balanced gender distribution (160 females, 140 males) and age diversity (mean age 26 ± 5 years), ensuring a representative analysis. It also includes multi-contrast CMR images, offering comprehensive coverage of cardiac tissue characteristics. The M&Ms-2 dataset represents clinically relevant cardiac conditions and offers diverse anatomical variations. Additionally, the dataset includes disease and healthy subjects, making the CAMS-Net robust and generalizable to various clinical scenarios.

CMR×Recon Segmentation data: The CMR×Recon MICCAI-2023 challenge [49] data has multi-contrast, multi-view, multi-slice, and multi-coil cardiac magnetic resonance imaging (MRI) data from 300 subjects. The research community uses the data for both reconstruction and segmentation tasks [38, 53]; in the proposed study, we have only used it as segmentation data. The challenge includes short-axis (SAX), two-chamber (2CH), three-chamber (3CH), four-chamber (4CH) long-axis (LAX) views, and T1 mapping and T2 mappings. We used the 4CH-LAX cine images and corresponding segmentation labels, which have been manually labeled by an expert radiologist where annotations are provided for four Cardiac chambers: LA (label=1), RA (label=2), LV (label=3), and RV (label=4). We utilize a randomly selected five-fold cross-validation split of the CMR×Segmentation dataset.

M&Ms-2 data: The M&Ms-2 is MICCAI 2021 challenge [2, 32], focused on RV segmentation and provided labels for LV, RV, and LV-myocardium (MYO). The data is collected from three clinical centers in Spain utilizing nine scanners from three vendors (Siemens, General Electric, and Philips). It contains 360 subjects, sequentially divided into 160 for training, 40 for validation, and 160 for testing. In the proposed study, we have used the LA view segmentation images, and similar to [27], all models are evaluated using a 5-fold cross-validation split.

4.2. Implementation Details

Comparative networks and the proposed framework were implemented using PyTorch, and all experiments were performed using NVidia A100 GPUs with 40GB RAM. AdamW [31] optimizer is used with β_1, β_2 = [0.5,0.55]; training was performed for 500 training epochs using Dice Loss [19] with an initial learning rate of $1e^{-4}$ which was halved after every 100 epochs. Similar to [28], we pre-trained the encoder part of the proposed CAMS-Net on ImageNet [43], followed by fine-tuning it on the segmentation data. We pre-process each input intensity image by normalizing it by its mean and standard deviation. Various intensity and geometric data augmentation are applied to improve the diversity of training, including Gaussian noise, blur, brightness contrast, random ghosting, rotation, scaling, random flipping, and random affine. For a fair comparison, all the networks are trained with the same protocol and fine-tuned where the pre-trained weights were available.

4.3. Experimental Validation with CMR Dataset

We compared and performed experiments with a number of state-of-the-art methods, including (i) CNNs-based models: UNet [41] and ResUNet [7], (ii) self-attention-based Swin-UNet [3], (iii) hybrid-architectures (CNNs+self-attention): UNETR [16] and Swin-UNETR [15], and (iv) Mamba-based models: VM-UNet [42] and Mamba-UNet [50]. Table 1 shows experimental validation results using five-fold cross-validation with the CMR-segmentation dataset. The proposed method outperforms existing methods while requiring the least model parameters. We note our method has 18.56 million trainable parameters, compared to counterparts with parameter counts ranging from 25.13 to 95.85 million. We attribute this to a combination of architectural innovations, including the LIFM block, which factorizes the Mamba block; the CS-IF module, which includes MCA and MSA capturing information along both channel and spatial dimensions; and the proposed bidirectional scanning scheme augmented with the proposed weight-sharing strategy.

Figure 2 shows the visual results of CAMS-Net and each comparative network. For rows (a) and (b), the proposed CAMS-Net column shows precise segmentation of anatomical structures, delineating clear boundaries. Specifically, row (a) shows the accurate segmentation of the LA bound-

Methods	# Params (M)↓	Dice Score (%) ↑					Hausdorff Distance- HD (mm) ↓				
		LA	RA	LV	RV	Avg	LA	RA	LV	RV	Avg
UNet [41]	31.03	85.36	79.81	90.39	85.17	85.18	4.66	10.71	6.11	5.46	6.73
ResUNet [7]	46.41	82.34	79.27	91.38	85.21	84.55	5.36	6.13	3.78	5.32	5.14
UNETR [16]	95.85	83.45	81.18	90.62	84.64	84.97	5.20	5.95	5.31	6.18	5.66
Swin-UNETR [15]	25.13	82.70	81.33	91.19	85.14	85.09	5.50	5.73	4.48	5.93	5.41
Swin-UNet [3]	41.35	83.48	81.91	90.60	84.54	85.13	4.51	6.28	4.48	6.10	5.34
TransUNet [5]	66.80	85.23	82.02	91.59	85.50	86.08	5.27	6.69	5.69	6.98	6.15
VM-UNet [42]	44.27	80.65	80.30	91.59	85.22	84.44	4.74	5.90	3.98	5.59	5.05
Mamba-UNet [50]	35.85	82.27	81.06	91.58	85.33	85.05	4.95	6.29	4.02	5.73	5.24
CAMS-Net(ours)	**18.56**	**86.06**	**84.44**	**92.53**	**87.35**	**87.59**	**4.07**	**5.43**	**3.10**	**5.40**	**4.50**

Table 1. Comparison of state-of-the-art methods using a five-fold cross-validation split of CMR-Segmentation Dataset. Best results are shown in **Bold** and model parameters (# Params) are listed in millions (M).

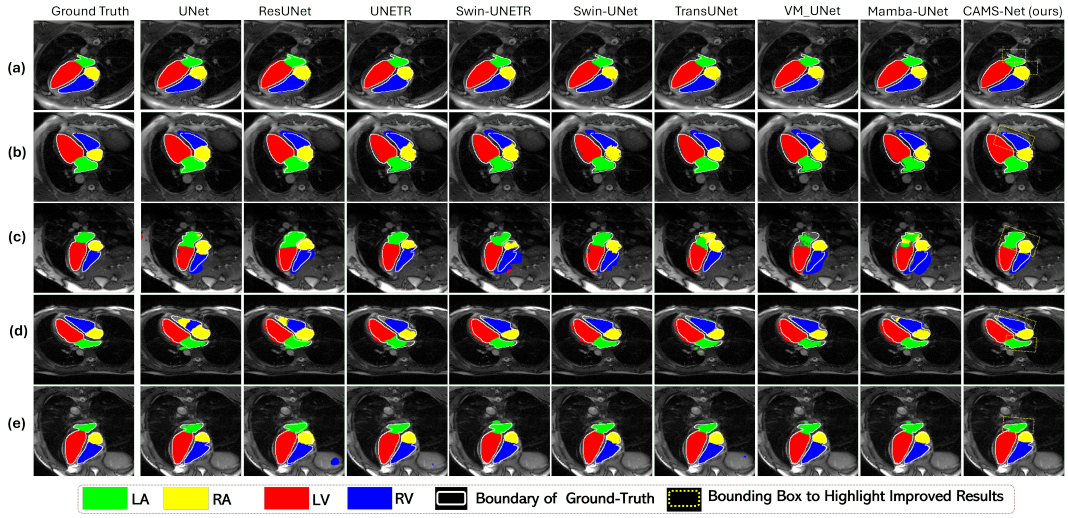

Figure 2. Qualitative comparison from CMRxSegmentation dataset, using CAMS-Net and other networks, highlighting CAMS-Net's enhanced performance in boundary separation and preserving spatial integrity across different regions. Please zoom in for details.

ary, while row (b) illustrates the distinct segmentation of the RV and RA boundaries compared to other networks. In row (c), where all other networks fail to reconstruct RA and LA correctly, the proposed CAMS-Net can segment these anatomies precisely with clear boundaries; we attribute this improvement to the incorporation of MSA, which helps maintain spatial coherence, as also demonstrated in ablation studies (also see Figure 4). For the last two rows of Figure 2, where the comparative networks either over-segment or under-segment, the CAMS-Net still performs comparably by ensuring balanced and accurate segmentation. In row (d), the CAMS-Net preserves the shape and structure of RV, compared to UNet, ResUNet, and Mamba-UNet, which segment sections of RV as RA. Finally, in row (e), where ResUNet, UNETR, and TransUNet generate a floating prediction for RV and other networks fail to delineate LA boundaries, the proposed CAMS-Net can gather spatial context from both directions because of Bidirectional scanning, which helps it to resolve ambiguities regions.

4.4. Experimental Validation with M&Ms-2 Dataset

To show the capability of CAMS-Net to work on multiple datasets, we also perform additional experimental validation using the M&Ms-2 dataset. Table 4 shows the CAMS-Net performance compared to existing methods, where it achieves the highest Dice Score across all categories and the lowest HD values, highlighting CAMS-Net's superior boundary accuracy.

The visual results shown in Figure 3 further advocate the superior performance of the proposed CAMS-Net. In rows (a) and (b), all networks struggle to differentiate between the LV and MYO's boundaries and fail to accurately capture the variable shapes of RV, compared to the proposed CAMS-Net's results, where it generates clear boundaries. In row (c), the other comparative networks, except SWIN-UNETR, are unable to segment the RV properly, resulting in higher false negatives. In row (d), except for the Mamba-based networks, all other comparative networks cannot capture the relationship between MYO and LV. Our CAMS-

Methods	Dice Score (%) ↑				HD (mm) ↓			
	LV	RV	Myo	Avg	LV	RV	Myo	Avg
UNet [41]	87.26	88.20	79.96	85.14	13.04	8.76	12.24	11.35
ResUNet [7]	87.61	88.41	80.12	85.38	12.72	8.39	11.28	10.80
InfoTrans* [25]	88.21	89.11	80.55	85.96	12.47	7.23	10.21	9.97
TransUNet [5]	87.91	88.23	79.05	85.06	12.02	8.14	11.21	10.46
MCTrans [20]	88.42	88.19	79.47	85.36	11.78	7.65	10.76	10.06
MCTrans* [20]	88.81	88.61	79.94	85.79	11.52	7.02	10.07	9.54
UTNet [11]	86.93	89.07	80.48	85.49	11.47	6.35	10.02	9.28
UTNet* [11]	87.36	90.42	81.02	86.27	11.13	5.91	9.81	8.95
SWIN-UNET [3]	90.88	86.66	79.93	85.82	10.08	10.08	6.07	8.74
UNETR [16]	91.08	87.30	81.17	86.52	8.86	9.93	5.98	8.25
SWIN-UNETR [15]	91.91	86.77	82.36	87.01	7.19	8.77	4.65	6.87
TransFusion* [27]	89.78	91.52	81.79	88.70	10.25	5.12	8.69	8.02
VM-UNet [42]	92.47	87.79	82.39	87.55	6.09	7.60	4.87	6.18
Mamba-UNet [50]	93.44	87.18	82.54	87.72	6.63	8.08	5.59	6.76
CAMS-Net (ours)	**94.45**	**91.87**	**86.21**	**90.84**	**3.64**	**4.79**	**2.61**	**4.34**

Table 2. Comparison of results obtained from different methods using a five-fold cross-validation split of M&Ms-2 dataset. Methods indicated with a $*$ use multi-view inputs. The best results are shown in **Bold**.

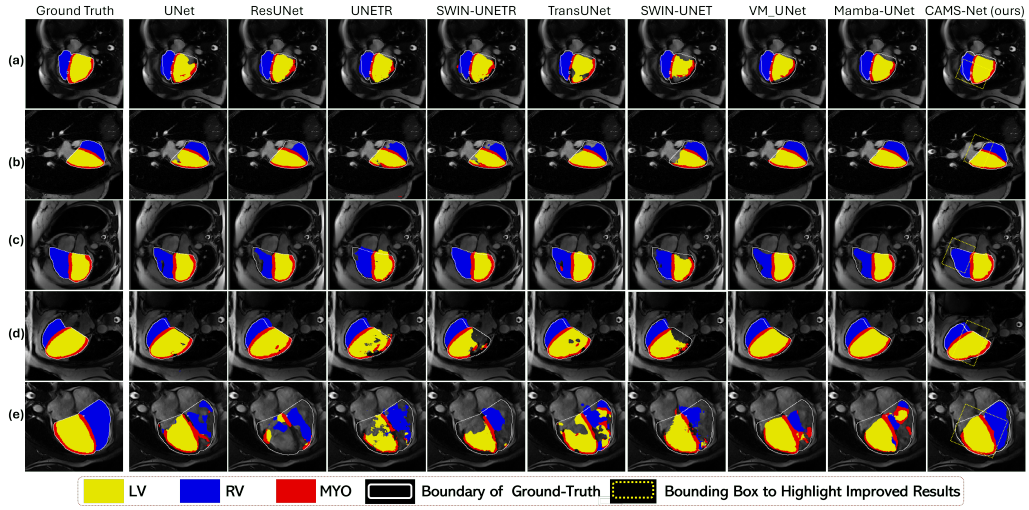

Figure 3. Qualitative comparison of visual results of CAMS-Net and other networks using M&Ms2 dataset. Please zoom in for details.

Net's predictions closely match the ground truth, confirming its capability of capturing long-range dependencies. For the last row (e), all other networks produce incomplete segmentations compared to the CAMS-Net, demonstrating spatial continuity and coherence in segmenting all three regions of interest, which is attributed to our proposed MSA module that captures information along spatial dimensions (see Figure 4 and Section 5 for further analysis).

5. Ablation Studies

We performed the following ablation studies on the CMR dataset to show how each proposed module contributes to improved accuracy. **CAMS-Net W/ and WO/ MSA:** The MSA fosters intercommunication among spatial locations, and its effectiveness is evaluated by removing it from the CS-IF module and using MCA throughout the network. The first two rows of Table 3 list the quantitative results of utilizing MSA and MCA, bringing an average improvement of 1.5% in the Dice score. Also, shown in the third row of Figure 4, the MSA enables the model to learn spatial dependencies between different regions of the image, resulting in better delineation of boundaries, better spatial coherence, fewer errors in spatial relationships, and generally improved localization of anatomical structures. Note that we have utilized the default scanning strategy from Mamba [13].

Bidirectional scanning and weight sharing strategy: The bidirectional scanning scheme incorporated at each encoder-decoder stage improves the results, shown in row 3 of the Table 3. However, this comes at the cost of ex-

MCA	MSA	Bidirectional Scanning	Weights Sharing	Positional Embeddings	Pretraining	LA	RA	LV	RV	Avg
✓	✗	✗	✗	✗	✗	80.15	78.33	89.65	83.22	82.83
✓	✓	✗	✗	✗	✗	82.17	81.09	90.30	83.79	84.33
✓	✓	✓	✗	✗	✗	83.27	81.83	90.27	84.76	85.03
✓	✓	✓	✓	✗	✗	83.20	81.10	91.46	84.68	85.11
✓	✓	✓	✓	✓	✗	84.11	82.09	91.50	84.51	85.55
✓	✓	✓	✓	✓	✓	**86.06**	**84.44**	**92.53**	**87.35**	**87.59**

Table 3. Ablation studies (Dice score %) utilizing MCA, MSA, bidirectional scanning augmented with weight sharing strategy, positional embeddings, and pretraining on ImageNet using a five-fold cross-validation split of CMR-segmentation dataset.

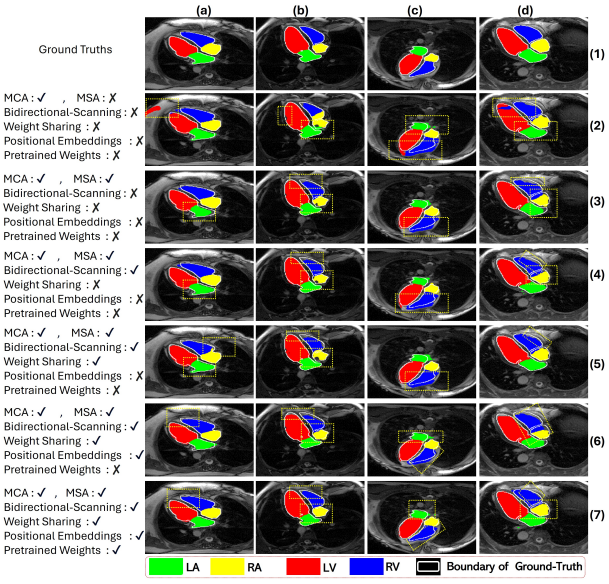

Figure 4. Visual comparison of results from different ablation studies on CMR-segmentation dataset. Please zoom in for details.

tra parameters for each forward and backward scanning scheme. The proposed bidirectional scanning, augmented with a sharing strategy, overcomes this limitation by reducing the parameters and improving the overall performance, shown in row 4 of Table 3. Rows 4 and 5 of Figure 4 depict how bidirectional scanning improves the results by capturing relationships from both directions, specifically in column (d); it helps the network better delineate the boundary between RV and RA.

Positional embeddings: CAMS-Net also utilizes the sinusoidal positional embeddings to encode spatial information about the position of each element within the input image sequence [48]. Row 5 of Table 3 lists the results of utilizing positional embeddings, which helps the network improve its average accuracy. Columns (a) and (b) of row 6 in Figure 4 exhibit how it can maintain spatial consistency across different regions, i.e., for RA here.

Effects of pre-training: We conducted experiments to examine the impact of ImageNet pre-trained parameters on the proposed CAMS-Net's performance. The last row of Table 3 lists the experimental results of this ablation and shows that utilizing the pre-trained weights improves the average Dice score by > 2%. Row 7 of Table 3 shows how pre-training leverages visual features like edges, textures, and shapes it has learned from ImageNet and improves the CAMS-Net's ability to detect boundaries more accurately.

6. Conclusion and Future Work

We are the first to propose a Mamba-based segmentation network without convolution operations and self-attention mechanisms to showcase the power of SSM-based architectures. We introduced several innovative strategies to the Mamba-based methods to increase their performance and reduce the computational complexity, including (i) a Linearly Interconnected Factorized Mamba (LIFM) block to reduce the number of trainable parameters and increase decision function, (ii) Mamba-based channel and spatial aggregators to learn the information across different channels along with spatial locations of the features, and (iii) a bidirectional weight-sharing strategy scheme. Our experiments demonstrate that the proposed CAMS-Net, an SSM-based segmentation network, outperforms the existing state-of-the-art in CNN, self-attention, and Mamba-based methods on CMR and M&Ms-2 segmentation datasets.

CAMS-Net has been implemented for 2D medical image segmentation, which has demonstrated impressive results in segmenting anatomical structures from medical scans like cardiac MRIs. However, there is significant potential to extend this work to 3D medical image segmentation, which we will explore in our future work.

Acknowledgements: This work acknowledges support from Queen Mary University of London's mini-CDT in AI-based Cardiac Image Computing, Andrena HPC, and NIHR Barts Biomedical Research Centre (NIHR203330), a partnership of Barts Health NHS Trust, Queen Mary University of London, and St George's University. Caroline Roney acknowledges the UKRI Future Leaders Fellowship (MR/W004720/1).

Methods	Dice Score (%) ↑				HD (mm) ↓			
	LV	RV	Myo	Avg	LV	RV	Myo	Avg
UNet [41]	87.26	88.20	79.96	85.14	13.04	8.76	12.24	11.35
ResUNet [7]	87.61	88.41	80.12	85.38	12.72	8.39	11.28	10.80
TransUNet [5]	87.91	88.23	79.05	85.06	12.02	8.14	11.21	10.46
MCTrans [20]	88.42	88.19	79.47	85.36	11.78	7.65	10.76	10.06
UTNet [11]	86.93	89.07	80.48	85.49	11.47	6.35	10.02	9.28
SWIN-UNET [3]	90.88	86.66	79.93	85.82	10.08	10.08	6.07	8.74
UNETR [16]	91.08	87.30	81.17	86.52	8.86	9.93	5.98	8.25
SWIN-UNETR [15]	91.91	86.77	82.36	87.01	7.19	8.77	4.65	6.87
VM-UNet [42]	92.47	87.79	82.39	87.55	6.09	7.60	4.87	6.18
Mamba-UNet [50]	93.44	87.18	82.54	87.72	6.63	8.08	5.59	6.76
CAMS-Net (ours)	**94.45**	**91.87**	**86.21**	**90.84**	**3.64**	**4.79**	**2.61**	**4.34**

Table 4. Comparison of results obtained from different methods using a five-fold cross-validation split of M&Ms-2 dataset. Methods indicated with a $*$ use multi-view inputs. The best results are shown in **Bold**.

References

[1] Abeer Aljuaid and Mohd Anwar. Survey of supervised learning for medical image processing. *SN Computer Science*, 3(4):292, 2022. 1

[2] Victor M Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martin-Isla, Alireza Sojoudi, Peter M Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, et al. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. *IEEE Transactions on Medical Imaging*, 40(12):3543–3554, 2021. 5

[3] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022. 2, 5, 6, 7, 9

[4] Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. Deep learning for cardiac image segmentation: a review. *Frontiers in cardiovascular medicine*, 7:25, 2020. 1

[5] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 6, 7, 9

[6] Tri Dao, Daniel Y Fu, Khaled K Saab, Armin W Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023. 1

[7] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020. 2, 5, 6, 7, 9

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 1, 2

[9] Michal Drozdzal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury, and Chris Pal. The importance of skip connections in biomedical image segmentation. In *International workshop on deep learning in medical image analysis, international workshop on large-scale annotation of biomedical data and expert label synthesis*, pages 179–187. Springer, 2016. 2, 3

[10] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018. 4

[11] Yunhe Gao, Mu Zhou, and Dimitris N Metaxas. Utnet: a hybrid transformer architecture for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 61–71. Springer, 2021. 1, 7, 9

[12] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2018. 1

[13] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 1, 2, 7

[14] Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2021. 1

[15] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2021. 2, 5, 6, 7, 9

[16] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022. 2, 5, 6, 7, 9

[17] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of digital imaging*, 32:582–596, 2019. 1

[18] Vincent Tao Hu, Stefan Andreas Baumann, Ming Gui, Olga Grebenkova, Pingchuan Ma, Johannes Fischer, and Bjorn Ommer. Zigma: Zigzag mamba diffusion model. *arXiv preprint arXiv:2403.13802*, 2024. 1, 5

[19] Shruti Jadon. A survey of loss functions for semantic segmentation. In *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*, pages 1–7. IEEE, 2020. 5

[20] Yuanfeng Ji, Ruimao Zhang, Huijie Wang, Zhen Li, Lingyun Wu, Shaoting Zhang, and Ping Luo. Multi-compound transformer for accurate biomedical image segmentation. In *MICCAI*, pages 326–336. Springer, 2021. 7, 9

[21] Davood Karimi, Serge Didenko Vasylechko, and Ali Gholipour. Convolution-free medical image segmentation using transformers. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 78–88. Springer, 2021. 2

[22] Abbas Khan, Muhammad Asad, Martin Benning, Caroline Roney, and Gregory Slabaugh. Crop and couple: cardiac image segmentation using interlinked specialist networks. *arXiv e-prints*, pages arXiv–2402, 2024. 1

[23] Rabeea Fatma Khan, Byoung-Dai Lee, and Mu Sook Lee. Transformers in medical image segmentation: a narrative review. *Quantitative Imaging in Medicine and Surgery*, 13(12):8747, 2023. 1

[24] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring image segmentation using transformers. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 2

[25] Lei Li, Wangbin Ding, Liqin Huang, and Xiahai Zhuang. Right ventricular segmentation from short-and long-axis mris via information transition. In *STACOM*, pages 259–267. Springer, 2021. 7

[26] Geert Litjens, Francesco Ciompi, Jelmer M Wolterink, Bob D de Vos, Tim Leiner, Jonas Teuwen, and Ivana Išgum. State-of-the-art deep learning in cardiovascular image analysis. *JACC: Cardiovascular imaging*, 12(8 Part 1):1549–1565, 2019. 1

[27] Di Liu, Yunhe Gao, Qilong Zhangli, Ligong Han, Xiaoxiao He, Zhaoyang Xia, Song Wen, Qi Chang, Zhennan Yan, Mu Zhou, et al. Transfusion: multi-view divergent fusion for medical image segmentation with transformers. In *MICCAI*, pages 485–495. Springer, 2022. 5, 7

[28] Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Yizhou Yu, Yong Liang, Guangming Shi, Shaoting Zhang, Hairong Zheng, et al. Swin-umamba: Mamba-based unet with imagenet-based pretraining. *arXiv preprint arXiv:2402.03302*, 2024. 5

[29] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024. 1, 2, 4

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 5

[32] Carlos Martín-Isla, Víctor M Campello, Cristian Izquierdo, Kaisar Kushibar, Carla Sendra-Balcells, Polyxeni Gkontra, Alireza Sojoudi, Mitchell J Fulton, Tewodros Weldebirhan Arega, Kumaradevan Punithakumar, et al. Deep learning segmentation of the right ventricle in cardiac mri: The m&ms challenge. *IEEE Journal of Biomedical and Health Informatics*, 2023. 5

[33] Carlos Martin-Isla, Victor M Campello, Cristian Izquierdo, Zahra Raisi-Estabragh, Bettina Baeßler, Steffen E Petersen, and Karim Lekadir. Image-based cardiac diagnosis with machine learning: a review. *Frontiers in cardiovascular medicine*, 7:1, 2020. 1

[34] Harsh Mehta, Ankit Gupta, Ashok Cutkosky, and Behnam Neyshabur. Long range language modeling via gated state spaces. In *The Eleventh International Conference on Learning Representations*, 2022. 1

[35] Sachin Mehta and Mohammad Rastegari. Separable self-attention for mobile vision transformers. *Transactions on Machine Learning Research*, 2022. 1

[36] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021. 1

[37] Badri Narayana Patro and Vijay Srinivas Agneeswaran. Mamba-360: Survey of state space models as transformer alternative for long sequence modelling: Methods, applications, and challenges. *arXiv preprint arXiv:2404.16112*, 2024. 1

[38] Abdul Qayyum, Hao Xu, Brian P Halliday, Cristobal Rodero, Christopher W Lanyon, Richard D Wilkinson, and Steven Alexander Niederer. Transforming heart chamber imaging: Self-supervised learning for whole heart reconstruction and segmentation. *arXiv preprint arXiv:2406.06643*, 2024. 5

[39] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *Advances in neural information processing systems*, 32, 2019. 1

[40] Mats L Richter, Julius Schöning, Anna Wiedenroth, and Ulf Krumnack. Should you go deeper? optimizing convolutional neural network architectures without training. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 964–971. IEEE, 2021. 1

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 2, 5, 6, 7, 9

[42] Jiacheng Ruan and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491*, 2024. 1, 2, 5, 6, 7, 9

[43] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 5

[44] Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17425–17436, 2023. 1

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015. 3

[46] Hui Tang, Yuanbin Chen, Tao Wang, Yuanbo Zhou, Longxuan Zhao, Qinquan Gao, Min Du, Tao Tan, Xinlin Zhang, and Tong Tong. Htc-net: A hybrid cnn-transformer framework for medical image segmentation. *Biomedical Signal Processing and Control*, 88:105605, 2024. 1

[47] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021. 2

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 8

[49] Chengyan Wang, Jun Lyu, Shuo Wang, Chen Qin, Kunyuan Guo, Xinyu Zhang, Xiaotong Yu, Yan Li, Fanwen Wang, Jianhua Jin, et al. Cmrxrecon: an open cardiac mri dataset for the competition of accelerated image reconstruction. *arXiv preprint arXiv:2309.10836*, 2023. 5

[50] Ziyang Wang, Jian-Qing Zheng, Yichi Zhang, Ge Cui, and Lei Li. Mamba-unet: Unet-like pure visual mamba for medical image segmentation. *arXiv preprint arXiv:2402.05079*, 2024. 1, 2, 5, 6, 7, 9

[51] Wenjian Yao, Jiajun Bai, Wei Liao, Yuheng Chen, Mengjuan Liu, and Yao Xie. From cnn to transformer: A review of medical image segmentation models. *Journal of Imaging Informatics in Medicine*, pages 1–19, 2024. 1

[52] Hanwei Zhang, Ying Zhu, Dan Wang, Lijun Zhang, Tianxiang Chen, and Zi Ye. A survey on visual mamba. *arXiv preprint arXiv:2404.15956*, 2024. 1

[53] Yirong Zhou, Chengyan Wang, Mengtian Lu, Kunyuan Guo, Zi Wang, Dan Ruan, Rui Guo, Peijun Zhao, Jianhua Wang, Naiming Wu, et al. Simultaneous deep learning of myocardium segmentation and t2 quantification for acute myocardial infarction mri. *arXiv preprint arXiv:2405.10570*, 2024. 5

[54] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6):1856–1867, 2019. 2

[55] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024. 1, 2, 4