This paper has been accepted for WACV 2025

Compositional Segmentation of Cardiac Images Leveraging Metadata

Abbas Khan
School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
Queen Mary's Digital Environment Research Institute (DERI), London, UK
acw676@qmul.ac.uk

Muhammad Asad
School of Biomedical Engineering and Imaging Sciences King's College London, UK
Queen Mary's Digital Environment Research Institute (DERI), London, UK
muhammad.asad@qmul.ac.uk

Martin Benning
Department of Computer Science, University College London, UK
Queen Mary's Digital Environment Research Institute (DERI), London, UK
martin.benning@ucl.ac.uk

Caroline Roney
School of Engineering and Materials Science, Queen Mary University of London, UK
Queen Mary's Digital Environment Research Institute (DERI), London, UK
c.roney@qmul.ac.uk

Gregory Slabaugh
School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
Queen Mary's Digital Environment Research Institute (DERI), London, UK
g.slabaugh@qmul.ac.uk

Abstract

Cardiac image segmentation is essential for automated cardiac function assessment and monitoring of changes in cardiac structures over time. Inspired by coarse-to-fine approaches in image analysis, we propose a novel multi-task compositional segmentation approach that can simultaneously localize the heart in a cardiac image and perform part-based segmentation of different regions of interest. We demonstrate that this compositional approach achieves better results than direct segmentation of the anatomies. Further, we propose a novel Cross-Modal Feature Integration (CMFI) module to leverage the metadata related to cardiac imaging collected during image acquisition. We perform experiments on two different modalities, MRI and ultrasound, using public datasets, Multi-Disease, Multi-View, and Multi-Centre (M&Ms-2) and Multi-structure Ultrasound Segmentation (CAMUS) data, to showcase the efficiency of the proposed compositional segmentation method and Cross-Modal Feature Integration module incorporating metadata within the proposed compositional segmentation network. The source code is available: https://github.com/kabbas570/CompSeg-MetaData.

1. Introduction

Segmenting cardiovascular anatomies in cardiac imaging involves dividing the image into semantically meaningful partitions, an essential step in numerous applications [7, 25] including diagnosis of several major cardiovascular diseases, such as dysplasia, cardiomyopathies, and pulmonary hypertension [5, 6]. Clinical data analysis can be tedious and time-consuming, with manual annotation of cardiac boundaries across different views and cycles. With the advent of deep learning, many advanced neural network-based algorithms have been proposed to automate cardiac image segmenta-

tion [1, 2, 29, 34]. However, the majority of these techniques only utilize the imaging modality as an input to the deep learning models, ignoring image-specific characteristics like acquisition parameters (scanner, vendor, field strength, number of frames, image quality), medical condition of the patients (disease, blood volume in ventricles, ejection fraction) and demographic specifications (sex, age).

As shown in Figure 1, acquisition parameters such as vendor and scanner and image-related characteristics, like disease, can affect image quality, appearance, and intensity patterns. For instance, images from Philips scanners in the dataset have higher intensity values than those from General Electric (GE) or Siemens. Also, images from patients without a disease (NOR) have a distinct intensity pattern when compared to images with underlying heart conditions, such as those affected by Hypertrophic Cardiomyopathy (HCM), Arrhythmogenic Cardiomyopathy (ARR), or Tetralogy of Fallot (FALL), for example, HCM may result in thicker ventricular walls, and ARR images show irregular heart shapes. In addition, the physiological aspects, such as age and sex, also contribute to the image characteristics; for example, older age patients may have poor image quality due to factors such as changes in tissue density and sex affects the heart size and position. Incorporating these correlations between the images and metadata within training can aid a segmentation model in accurately identifying patterns within the imaging data, resulting in improved robustness and accuracy.

Existing deep learning methods for medical image segmentation can be divided into two categories: (1) single-stage methods and (2) two-stage methods. For single-stage methods, the entire image is directly fed to the network [8, 30]. For two-stage approaches, the search area is limited by localizing the organ(s) and further segmenting each class [10, 33]. The single-stage methods are efficient regarding the end-to-end training and inference time. However, they may struggle to precisely segment the cardiac regions due to their complex anatomy, motion, high variability in shape between individuals, and the challenges posed by different image quality and modalities. While two-stage methods, which involve a localization network (coarse segmentation or a regression network) followed by a detailed segmentation network, can achieve higher accuracy by focusing on the relevant areas and refining the segmentation, they increase computational complexity, requiring more resources and time for inference and training.

This paper addresses these challenges to enable accurate anatomical segmentation. To this end, we propose a super-to-sub segmentation compositional approach to localize and segment the heart simultaneously and a Cross-Modal Feature Integration (CMFI) module to incorporate the metadata associated with each cardiac image to handle the variability in image characteristics that result from different equipment, protocols, patient conditions and demographics depending upon the metadata availability. We validated the proposed method on two different imaging modalities, MRI and Ultrasound, using Multi-Disease, Multi-View, and Multi-Centre (M&Ms 2) [3, 27] and Multi-structure Ultrasound Segmentation (CAMUS) dataset [22], respectively.

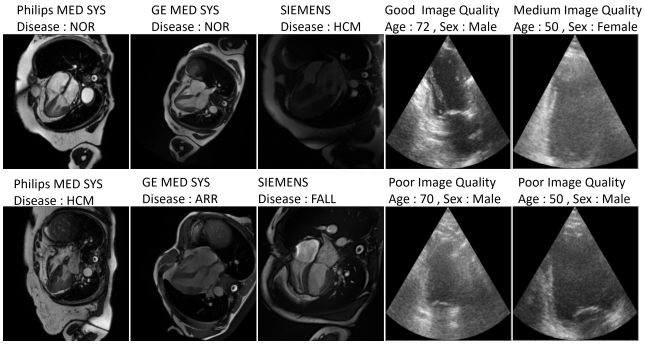

Figure 1. Influence of metadata on image quality, appearance, and intensity patterns: Analysis of MRI (M&Ms-2 dataset) and Ultrasound (CAMUS dataset) scans across different acquisition parameters and patient-specific information, i.e, disease, sex, and age.

2. Related work

UNet [30] pioneered an encoder-decoder model architecture for medical image segmentation. A number of methods have been proposed to further improve UNet [9, 18, 24]. The 'No New-Net UNet' (nnUNet) [18] is an extension to UNet with automatic hyperparameter configuration to target a range of medical image segmentation tasks. InfoTrans [24] utilized nnUNet for cardiac segmentation, where information transition was proposed to utilize the long-axis (LA) view to assist with the segmentation of a short-axis (SA) view. The predicted LA views were utilized to locate and crop the SA views. Tempera [12] proposed a Spatial Transformer Feature Pyramid-based Network that uses hybrid 2D/3D convolutions to segment the right ventricle (RV). A multi-view SA-LA model is proposed by Jabbar et al. [19] to segment the RV on the SA and LA cardiac MR images. Their method is trained and validated on 2D slices from MRI volumes, where the bottleneck layers of both views are coupled as input to the decoder.

Recent improvements shown by vision transformers [11] have inspired several medical image segmentation architectures [8, 13, 20, 26]. TransUNet [8] improved the UNet architecture with self-attention within the encoder only. UTNet [13] proposed an efficient self-attention mechanism with reduced computational complexity, incorporating self-attention in both the encoder and decoder. MCTrans [20] utilized multi-view inputs and performs intra- and inter-scale self-attention of different convolutional features. TransFusion [26] merged multi-view imaging information using a Divergent Fusion Attention (DiFA) to capture long-range correlations between unaligned data and a Multi-Scale Attention

block for learning the global correspondence of multi-scale feature representations.

Our proposed compositional approach and metadata utilization strategy are inspired by several methods, including cascaded architectures, [10, 32], FiLMed-UNet [23], and SwiftFormer [31]. Similar to two-stage methods [10, 32], firstly, the heart is localized using *super-segmentation* decoder, and then *sub-segmentation* decoder simultaneously segments the heart into LV, LA, RV, and MYO. Lemay et al. proposed FiLMed-UNet [23], where the authors used Feature-wise Linear Modulation (FiLM) layers to integrate metadata at different encoder-decoder stages of a UNet, leading to improved segmentation accuracy. However, our proposed approach learns metadata as an auxiliary task, using a classifier based on Multilayer Perceptron (MLP) [16]. We also propose a novel way of integrating the metadata features into the segmentation network using the CMFI module.

Our work is also inspired by [31], which proposes an efficient additive attention (E-2A) mechanism to reduce the quadratic computational complexity of self-attention to a linear element-wise multiplication. In our proposed CMFI module, we have utilized the E-2A to intermingle the image and metadata features. More specifically, we perform cross-attention between the segmentation network features and metadata features to provide additional context about the image's content and improve the segmentation accuracy. Our contributions are as follows:

1. We propose a novel compositional segmentation approach that simultaneously localizes the heart (super-segmentation) and segments the heart structures (sub-segmentation).

2. We propose a Cross-Modal Feature Integration (CMFI) module to utilize the image metadata, including acquisition parameters, medical condition, and demographic of the patient, to conditionally modulate the segmentation network.

3. Extensive quantitative and qualitative experimental comparisons demonstrate that our proposed method outperforms the existing state-of-the-art. We evaluate the proposed approaches on two different modalities, MRI and ultrasound, and show that our approach excels in these diverse domains. The consistent performance improvements observed in both modalities indicate that our method could yield similar accuracy enhancements in other domains and modalities.

3. Methods

Figure 2 shows the proposed multi-task compositional cardiac image segmentation network. The inputs to the model are both cardiac images and related metadata information. The image encoder extracts features from the image, and the metadata is learned via an MLP. Our proposed CMFI module conditions the segmentation network based on the metadata for a given image. Here, we emphasize that a medical image, such as a cardiac MRI, is influenced by characteristics of the imaging equipment employed for the acquisition of the image, such as manufacturer (vendor), scanner type, field strength, image quality, underlying anatomical characteristics (pathological conditions, blood volume in ventricles) depicted in the image [28], and patient's demographics. Hence, the image segmentation network is guided based on the hallmarks associated with the intensity images to enhance the segmentation performance.

In the proposed hierarchical decoder strategy, the super-segmentation decoder gets the modulated features and localizes the heart as a single region (super-segmentation). As a simultaneous step, the sub-segmentation decoder copies the decoder features from the super-segmentation and refines part-based segmentation further into different Regions of Interest (ROIs), such as LV, LA, RV, and MYO (sub-segmentation). Our compositional segmentation network is trained end-to-end simultaneously for both super and sub-segmentation and classification/regression tasks.

3.1. Encoders Strategy

Image Encoder: Convolutional layers are used to extract features from the image. Each block consists of two consecutive 3×3 convolutions followed by batch normalization and LeakyReLU activation [17]. The number of feature maps for the image encoder is increased to 32, 64, 128, 256, and 320 while down-sampling spatially by a factor 2× using a 3×3 convolution with stride 2.

Metadata MLP: For metadata, an MLP is implemented, shown in the bottom left of Figure 2. The metadata is passed through a series of linear layers where the number of linear layers equals the number of encoder or decoder stages, and the number of neurons at each layer is equal to the number of feature maps at the respective encoder-decoder stage. Every linear layer is followed by 1D-batch normalization, LeakyReLU activation, and a dropout of 0.1. Finally, a linear layer transforms the high-dimensional feature space into a 128-dimensional representation and feeds it into the respective linear layer to classify or regress the available metadata entity.

3.2. Encoding MetaData Into a Metadata Tensor

The M&Ms-2 dataset was captured using MRI machines from three vendors and nine scanners under two magnetic field strengths, 1.5 and 3 Tesla, to create a highly heterogeneous dataset that reflects the diversity seen in real-world clinical practice. The training set has five diseases and instances of normal cases. Each metadata entity is mapped to a numerical representation; for example, vendors Philips, Siemens, and GE are mapped to numerical values (e.g., 1, 2, 3) using a predefined dictionary. A similar mapping procedure is adopted for scanner and disease categories. The field

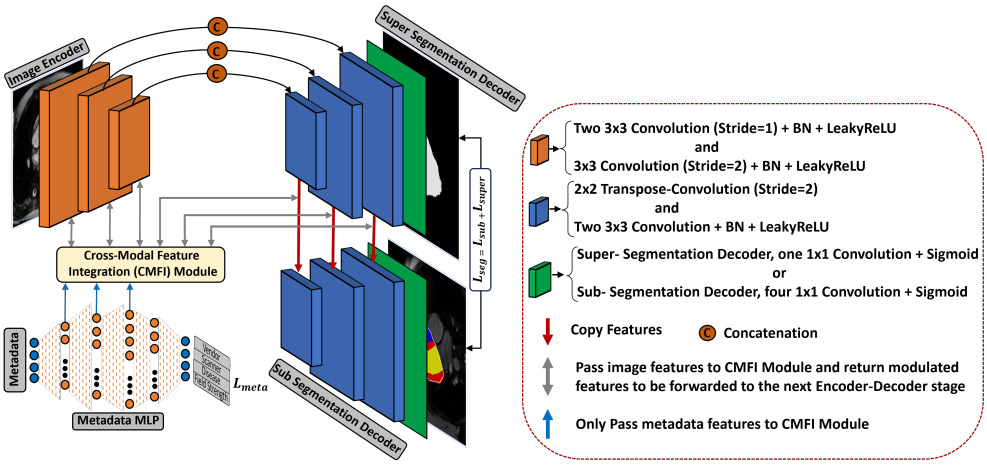

Figure 2. Overview of the proposed pipeline. The network has five encoder-decoder stages; only three are shown here for simplicity. The image encoder extracts features from the image; the two decoders perform super and sub-segmentation. The metadata is learned via MLP, followed by the interaction of image and metadata features using the CMFI module.

strength, i.e., (1.5 or 3 Tesla), is used directly.

The metadata extracted from the CAMUS dataset includes attributes such as end-systolic (ES) and end-diastolic (ED) frames, the total number of frames (NbFrame), patient sex and age, image quality, ejection fraction (EF), and frame rate. We normalized continuous metadata values by dividing them by a factor of 10 to scale input features appropriately. Categorical variables, such as sex and image quality, were mapped to numerical values; for example, sex:{Male, Female} mapped to 0,1 and image quality:{Good, Medium, Poor} to 0,1,2. All metadata encodings are released with our source code.

3.3. Cross-Modal Feature Integration (CMFI) module

The proposed CMFI module, shown in Figure 3 requires the Query ($\mathbf{Q}$) and Key ($\mathbf{K}$) interaction and is used at all stages of the segmentation network. At each encoder-decoder stage, the metadata feature matrix ($\mathbf{f}_M$) of shape $\mathbb{R}^{B\times C}$ is expanded to the size of the image feature matrix ($\mathbf{f}_I$), resulting in identical dimensions of $\mathbb{R}^{B\times C\times H\times W}$, where ($B$: Batch size, C: Number of channels, H: Height, W: Width). Finally, the features from both modalities are reshaped to $\mathbb{R}^{B\times N\times C}$, such that, $(\mathbf{f}_I, \mathbf{f}_M) \in \mathbb{R}^{B\times N\times C}$, and $N = H \times W$.

Each feature matrix $\mathbf{f}_M$ and $\mathbf{f}_I$ undergoes a projection to generate corresponding $\mathbf{Q}$ and $\mathbf{K}$ matrices, i.e., ($\mathbf{Q}_I$,$\mathbf{K}_I$: for image features, and $\mathbf{Q}_M$, $\mathbf{K}_M$: for metadata features), such that, $(\mathbf{Q}_I, \mathbf{K}_I, \mathbf{Q}_M, \mathbf{K}_M) \in \mathbb{R}^{B\times N\times C}$

The learnable attention weights for image modality ($\mathbf{w}_{aLI} \in \mathbb{R}^{B\times N\times 1}$) and metadata modality ($\mathbf{w}_{aLM} \in \mathbb{R}^{B\times N\times 1}$) are obtained by following steps.

1. Multiply the query matrix of the image modality ($\mathbf{Q}_I$) with its corresponding parameter vector ($\mathbf{w}_{aI} \in \mathbb{R}^{C\times 1}$) to get ($\mathbf{w}_{aLI}$).

2. Multiply the query matrix of the metadata modality ($\mathbf{Q}_M$) with its corresponding parameter vector ($\mathbf{w}_{aM} \in \mathbb{R}^{C\times 1}$) to get ($\mathbf{w}_{aLM}$).

3. Apply a scaling operation to these products.

Mathematically, these operations can be represented as:

$$\mathbf{w}_{aLI} = \frac{\mathbf{Q_I}\mathbf{w}_{aI}}{\sqrt{C}} \quad \text{and} \quad \mathbf{w}_{aLM} = \frac{\mathbf{Q_M}\mathbf{w}_{aM}}{\sqrt{C}} \tag{1}$$

Following this, the ($\mathbf{w}_{aLI}$,$\mathbf{w}_{aLM}$) are multiplied with ($\mathbf{Q}_I$,$\mathbf{Q}_M$) and summed along dimension N, to produce a single global attention query vector for image modality ($\mathbf{G}_I \in \mathbb{R}^{B\times C}$), and metadata modality ($\mathbf{G}_M \in \mathbb{R}^{B\times C}$):

$$\mathbf{G}_I = \sum_{n=1}^{N} (\mathbf{w}_{aLI})_N \odot \mathbf{Q}_{IN} \tag{2}$$

and,

$$\mathbf{G}_M = \sum_{n=1}^{n} (\mathbf{w}_{aLM})_n \odot \mathbf{Q}_{Mn} \tag{3}$$

where $\odot$ denotes element-wise multiplication, and global attention query vectors are expanded to the dimension of $\mathbb{R}^{B\times N\times C}$. The global and cross-global context for $\mathbf{f}_I$ is established through the interaction of $\mathbf{G}_M$ and $\mathbf{G}_I$ with both $\mathbf{K}_I$, and $\mathbf{K}_M$, followed by a linear transformation layer ($\boldsymbol{T_j}$) for the $\boldsymbol{j}$th pair of interactivity, where $\boldsymbol{j \in 1, 2, 3, 4}$.

$$\begin{aligned} \mathbf{f}_I^* = &\mathbf{T}_1(\mathbf{G}_I \odot \mathbf{K}_I) + \mathbf{T}_2(\mathbf{G}_I \odot \mathbf{K}_M) + \\ &\mathbf{T}_3(\mathbf{G}_M \odot \mathbf{K}_M) + \mathbf{T}_4(\mathbf{G}_M \odot \mathbf{K}_I) \end{aligned} \tag{4}$$

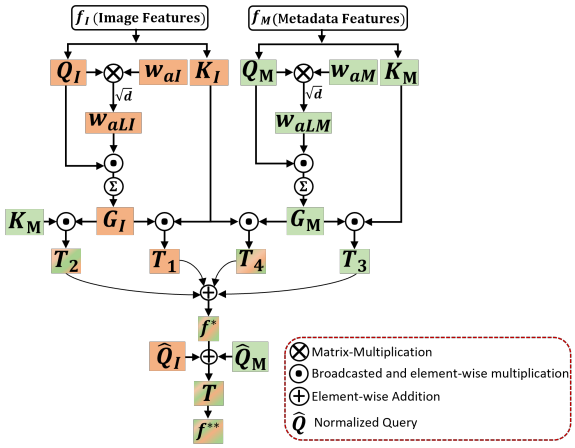

Figure 3. The proposed CMFI Module. Each block's subscripts, $\boldsymbol{I}$ and $\boldsymbol{M}$, represent the image and metadata features, respectively.

Finally, the normalized $\mathbf{Q_I}$, and $\mathbf{Q_M}$ are summed with $\mathbf{f}_I^*$, and another linear transformation layer ($\boldsymbol{T}$) is applied on the resultant vector to obtain the final modulated segmentation network features $\mathbf{f}_I^{**}$ which is passed to the next encoder-decoder stage:

$$\mathbf{f}_I^{**} = \boldsymbol{T}\Big(\mathbf{f}_I^* + \frac{\mathbf{Q}_I}{\|\mathbf{Q_I}\|_2} + \frac{\mathbf{Q}_M}{\|\mathbf{Q_M}\|_2}\Big) \tag{5}$$

where, $\|\cdot\|_2$ denotes the Euclidean norm.

3.4. Hierarchical Decoder Strategy

Our proposed hierarchical decoder strategy comprises two decoders: super and sub-segmentation decoders. Each decoder upsamples the features using 2×2 transpose convolutions, followed by two 3×3 convolutions, batch normalization, and LeakyReLU activation. The super-segmentation decoder utilizes skip connections from the encoder to segment all three ROI classes as a single binary segmentation map. The sub-segmentation decoder gets the features from the super-segmentation decoder and further segments the binary segmentation features into multiple classes. This feature-copying process from the super-segmentation decoder (shown by the red downward arrow in Figure 2) makes part-based segmentation of ROIs more efficient. Instead of directly looking at the entire image and finding the relevant features, our approach implicitly confines the search area based on super-segmentation. Empirical tests showed that adding skip connections to the sub-segmentation decoder did not improve its overall accuracy and added extra learnable weights. Therefore, the sub-segmentation decoder does not utilize skip connections.

The proposed multi-task network is trained end-to-end using the composite loss function below.

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{seg} + (1-\alpha)\mathcal{L}_{meta}, \tag{6}$$

where $\mathcal{L}_{seg}$ is the segmentation loss associated with the proposed compositional approach, and $\mathcal{L}_{meta}$ is the loss to learn the metadata entities, described below. The segmentation loss $\mathcal{L}_{seg}$ is composed of the sub-segmentation loss ($\mathcal{L}_{sub}$) and the super-segmentation loss ($\mathcal{L}_{super}$).

$$\mathcal{L}_{seg} = \mathcal{L}_{sub} + \mathcal{L}_{super}. \tag{7}$$

For both datasets, the segmentation losses, i.e, $\mathcal{L}_{sub}$ and $\mathcal{L}_{super}$ are the Dice losses. For the M&Ms-2 dataset, the $\mathcal{L}_{meta}$ loss is a cross-entropy loss for classification employed for the MLP network. For the CAMUS dataset, the $\mathcal{L}_{meta}$ is the sum of cross-entropy and L1-loss depending on the nature of metadata, i.e., for continuous variables, L1-loss and for categorical, the cross-entropy loss.

Based on empirical experiments, α is the balancing factor in equation (6) between the two loss terms and is set to $\alpha = 0.7$. Given the simpler nature of the classification task, which serves as an auxiliary component compared to segmentation, a higher weight is assigned to the segmentation task in the overall loss formulation.

Here, we emphasize that our proposed compositional approaches (super and sub-segmentation) and CMFI module for incorporating the metadata methods are general purposes and can be applied to other segmentation networks as well. In this paper, we have used the UNet architecture as a baseline with some modifications, including strided convolutions with stride = 2 to reduce the in-plane spatial dimensions of features in the encoder compared to 2 × 2 max-pooling operations of UNet, transpose convolutions in the decoder to upsample the features compared to UNet's bilinear interpolation method and replacing the ReLU activation with LeakyReLU.

4. Experimental Validation

In this section, we provide details of the datasets, implementation and our experimental validation results showing our approach's superior performance as compared to the state-of-the-art.

4.1. Datasets Description

We utilize the following two datasets for the experimental validation of our proposed methods:

M&Ms-2 data: This data comes from a challenge cohort hosted by MICCAI 2021 [3,27], consisting of RV blood pool segmentation across cardiac MRI imaging of SA and LA views. Segmentation labels are provided for three ROIs: (i) LV blood pools, (ii) RV blood pools, and (ii) LV myocardium (LV-MYO). In our work, we conduct LA view segmentation experiments from M&Ms-2 data. Following [26], we randomly shuffled the 160 training samples and evaluated all models using a 5-fold cross-validation split.

CAMUS data: This dataset [22] provides 2D echocardiographic images of two and four-chamber views for 500 patients. The CAMUS provides manual labels for the left

Table 1. Comparison of the results obtained from different methods using LA views of M&Ms-2 dataset using a five-fold cross-validation split. Methods indicated with a $*$ use multi-view inputs. The best results are shown in **Bold**.

Methods	Dice Score (%) ↑				HD (mm) ↓			
	LV	RV	Myo	Avg	LV	RV	Myo	Avg
UNet [30]	87.26	88.20	79.96	85.14	13.04	8.76	12.24	11.35
ResUNet [9]	87.61	88.41	80.12	85.38	12.72	8.39	11.28	10.80
InfoTrans* [24]	88.21	89.11	80.55	85.96	12.47	7.23	10.21	9.97
TransUNet [8]	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 [13]	86.93	89.07	80.48	85.49	11.47	6.35	10.02	9.28
UTNet* [13]	87.36	90.42	81.02	86.27	11.13	5.91	9.81	8.95
TransFusion* [26]	89.78	91.52	81.79	87.70	10.25	5.12	8.69	8.02
UNETR [15]	91.33	85.71	80.85	85.96	10.08	11.28	6.07	9.14
SWIN-UNETR [14]	92.21	86.93	81.77	86.96	8.84	9.73	5.60	8.05
nnUNet* [18]	94.08	90.70	86.41	90.39	5.91	6.61	5.98	6.16
Two-Stage [32]	94.20	89.15	85.51	89.62	4.40	6.29	3.90	4.86
Proposed (WO/ Super-Seg.)	90.57	88.05	82.19	86.93	6.94	8.41	5.22	6.85
Proposed (WO/ CMFI)	94.48	90.29	85.31	90.02	3.55	5.19	**2.64**	3.80
Proposed W/(Super-Seg.+CMFI)	**95.63**	**91.93**	**87.61**	**91.72**	**3.16**	**4.62**	2.95	**3.57**

ventricle endocardium (LV), the myocardium epicardium (MYO), and the left atrium (LA). In the proposed study, we have utilized the two-chamber views from 500 patients in a 5-fold cross-validation split.

4.2. Implementation Details

The proposed framework is implemented using PyTorch and an NVidia A100 GPU with 40GB RAM. All models are trained using Adam optimizer [21] for 500 epochs, learning rate = $1e^{-4}$ and batch size 16. The images are resampled to an in-plane resolution of 1.25×1.25 mm^2 for M&Ms-2 and 1×1 mm^2 for CAMUS data. Each image is normalized by its mean and standard deviation. Various geometric and intensity data augmentation strategies are utilized, including rotation, shift, scaling, elastic deformation, Gaussian noise, Gaussian blur, and random bias field.

4.3. Experimental Validation with M&Ms-2 Dataset

Our experimental validation results with the M&Ms-2 dataset, shown in Table 1, compare our proposed method with existing state-of-the-art methods, including UNet-based, transformer-based, and two-stage segmentation approaches. The proposed hierarchical segmentation method utilizes UNet and has a few extra trainable parameters compared to UNet. Figure 1 in the supplementary material depicts further insights into the performance and number of parameter methods. Our proposed method achieves state-of-the-art accuracy with the highest Dice and lowest Hausdorff Distance (HD) scores across all segmentation classes. Furthermore, Figure 4 shows qualitative results, showing the proposed compositional approach enables accurate delineation of segmentation while reducing false positives outside the heart region. This improved accuracy is mainly due to two aspects of our proposed: (i) hierarchical decoder where super-segmentation decoder, which confines the segmentation to the heart region and helps the sub-segmentation decoder by passing only relevant features, and (ii) incorporating additional metadata using the CMFI module that provides additional context resulting in specialized and accurate segmentation. We note that the accuracy of our hierarchical decoder is closer to a two-stage method; however, in contrast to the sequential execution of two-stage methods, where the first network localizes followed by segmentation, our compositional approach can simultaneously locate and segment heart and internal structures.

Strategically incorporating metadata alongside intensity images enhances the accuracy and reliability of the results. Metadata is motivated by its capacity to offer contextual information that image data alone cannot provide. For example, as shown in Figure 4, the first two rows come with FALL and Inter-atrial communication (CIA) diseases, which primarily impacts the RV outflow tract obstruction and volume overload, respectively. In the W/ CMFI Module column, where we supplied the model with this disease-specific information, it can prioritize the segmentation of the RV, ultimately leading to improved delineation and accuracy of the RV. Similarly, in the third row, metadata about the Dilated Left Ventricle (DLV) guides the model in accurately segmenting the LV and the MYO, acknowledging the expected dilation and MYO wall thinning associated with this condition. Furthermore, metadata such as vendor, scanner, and field strength have a crucial role in adapting the model to the underlying variations in image characteristics. The CMFI module is designed to merge metadata with images effectively, which enables the model to make more informed predictions by leveraging the additional data. This leads to improved segmentation performance.

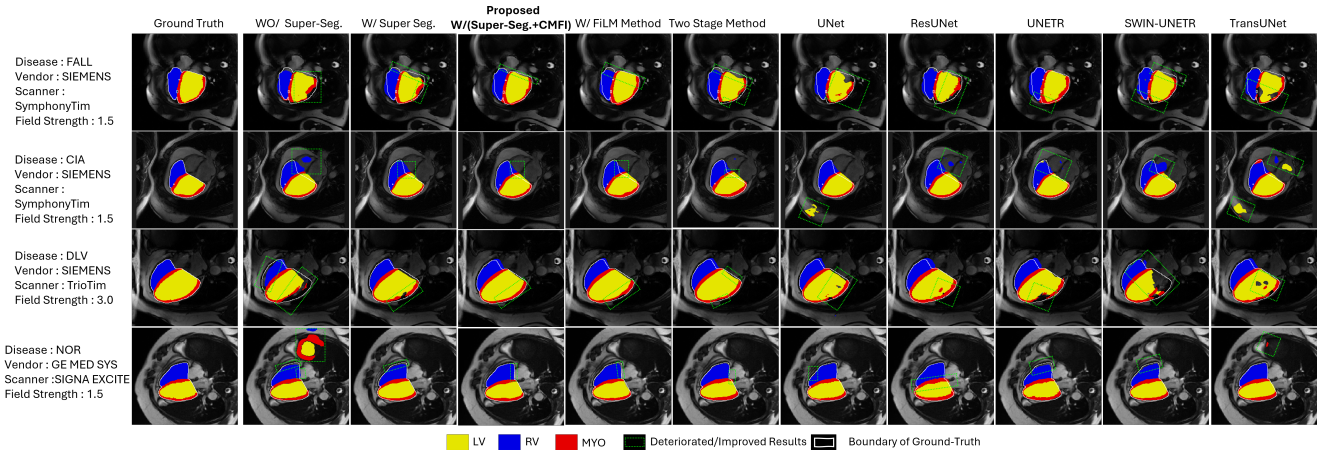

Figure 4. Visual comparison of our compositional approach with (W/) and without (WO/) the super-segmentation, metadata utilization strategy using FiLM [23], proposed CMFI module, and other comparative networks using M&Ms-2 dataset. Please zoom in for details.

Table 2. Quantitative results on five-fold cross-validation split of CAMUS data comparing performance with (W/) and without (WO/) super-segmentation decoder, metadata W/ CMFI module, and other comparative segmentation networks.

Methods	LV-Dice	Myo-Dice	LA-Dice	Avg-Dice	LV-HD	Myo-HD	LA-HD	Avg-HD
UNet [30]	91.52	84.70	86.44	87.55	19.78	21.22	32.02	24.34
ResUNet [9]	92.40	86.59	86.79	88.59	17.77	19.35	25.58	20.90
UNETR [15]	91.64	84.81	86.93	87.79	16.06	17.68	22.44	18.72
TransUNet [8]	88.58	80.79	81.65	83.67	31.54	42.18	34.89	36.20
SWIN-UNet [4]	92.06	85.64	87.47	88.38	14.94	16.40	18.78	16.70
SWIN-UNETR [14]	92.60	86.55	87.03	88.72	15.98	16.80	19.95	17.57
Proposed (WO/ Super-Seg.)	92.29	86.06	88.31	88.88	15.06	17.38	22.18	18.20
Proposed (WO/ CMFI)	93.44	87.58	89.08	90.03	13.44	15.52	18.19	15.71
Proposed W/(Super-Seg.+CMFI)	**93.48**	**88.70**	**89.90**	**90.69**	**12.17**	**15.01**	**17.89**	**15.02**

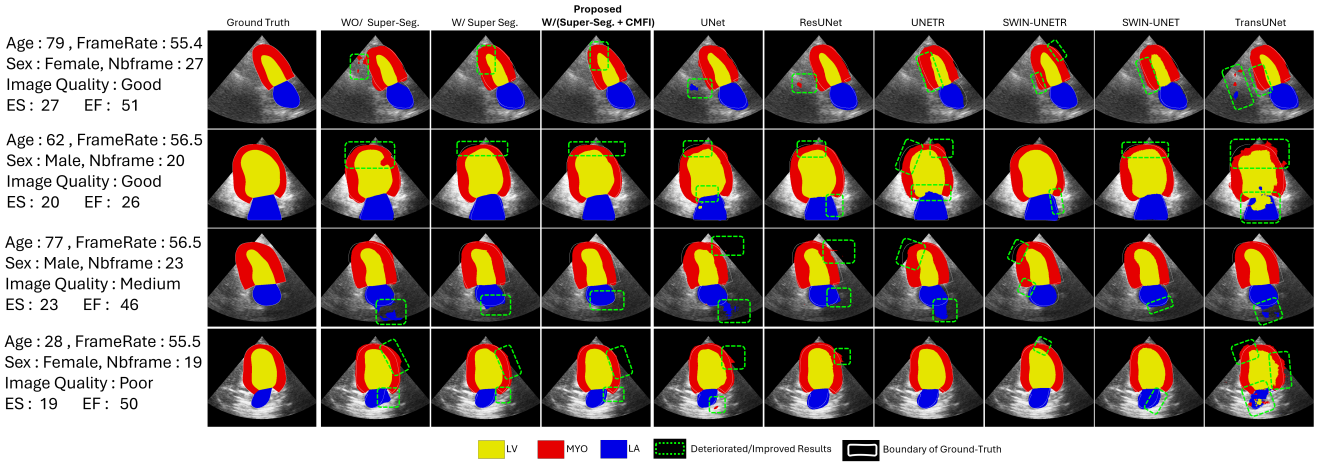

Figure 5. Qualitative comparison of proposed compositional approach with (W/) and without (WO/) the super-segmentation and metadata utilization strategy using CMFI module using CAMUS data. Please zoom in for details.

4.4. Experimental Validation with CAMUS Dataset

Our experiments on the CAMUS data [22] show the proposed methods' ability to work effectively in different modalities, where our proposed method achieves similar accuracy improvements as experiments with the M&Ms-2 dataset. Table 2 shows the quantitative results for CAMUS dataset, where the proposed super-segmentation de-

coder (without CMFI) yields notable improvements across all metrics for all three ROI. This is mainly due to better heart region localization as well as improved segmentation boundary delineation, as demonstrated in Figure 5 (W/ vs WO/ super segmentation). The performance metrics improve further when the super-segmentation decoder and CMFI module are employed together, as listed in the last row of Table 2. With the CMFI module and super segmentation decoder, the segmentation boundaries are further refined and accurately delineated, as shown in the fourth column of Figure 5. The CMFI module maintains performance with variable image quality across varying ages, genders and various settings for functional aspects of the cardiac images.

Despite the good image quality in the first two rows of Figure 5, the method without metadata (third column) results in extended LA boundaries and unclear separation between the MYO and LV, similarly, in rows three and four, where the image quality drops and results in blurred structures, the without metadata method further extends the MYO and LA due to unclear boundaries between the structures. However, when metadata is incorporated (fourth column), the model better understands the anatomical structures, leading to more precise boundaries and more explicit segmentation. Here, we speculate that the metadata, such as image quality, age, and sex, informs the model about the expected variations in size, motion, and function of the heart, making it more robust. For example, older individuals show thicker heart walls, reduced elasticity, and more tissue heterogeneity compared to younger. Also, males typically have larger heart chambers and thicker myocardial walls than females. The experimental settings can also help the network understand the patterns in the imaging data. For example, ES reflects the blood volume in the ventricle at the end of the contraction, and a higher value will indicate the larger size of LV, guiding the segmentation model to adjust the boundaries accordingly.

5. Ablation Studies

The proposed compositional approach's effectiveness and metadata utilization are evaluated through the following ablation studies using the M&Ms-2 dataset.

Ablation without using the super-segmentation decoder. In this ablation, we compare the effectiveness of our proposed method that utilizes a super-segmentation step to help localize the heart, shown in the first two rows of Table 3. We report a Dice score (>=95%) and $HD_{(mm)}$<=1.0 for super-segmentation (Not shown in tables as it's a pseudo-Dice score of three combined regions). The super-segmentation decoder improves the average Dice score of LV, RV, and MYO by 3.09% while reducing the average HD score by 3.05.

Figure 4 visualizes how the super-segmentation decoder improves the segmentation accuracy of different regions of interest. By localizing the heart, we confined the search area for the sub-segmentation decoder (identifying an overall heart topology) to segment the LV, RV, and MYO. We note that the super-segmentation enables our proposed method to localize the relevant region of interest, resulting in improved segmentation accuracy.

Ablation without using the CMFI module. In this experiment, we show the effectiveness of the CMFI module to condition segmentation on the metadata information. We also compare against FiLM [23] and Table 3 shows the accuracy comparison where we note that the network accuracy is improved by utilizing either of the approaches. However, the proposed CMFI module outperforms the FiLM method for all metrics with a 1.7% average improvement in the Dice score compared to 0.72% of the FiLM method. This advocates that the proposed CMFI module provides a better way of leveraging the metadata through a non-linear attention mechanism and integrates metadata more nuancedly.

Ablation without using the disease in metadata. This ablation provides clinical justification and how the proposed approach can help to integrate clinical knowledge into the segmentation process. Table 4 showcases the results W/ and WO/ using the disease information. The overall accuracy is improved by incorporating the disease into the segmentation network. This demonstrates that deep learning models can leverage clinical knowledge to guide the segmentation, especially where the goal is to prioritize the disease cases.

Ablation to handle unavailability of metadata. If *no metadata is available*, we can use our method without CMFI, which still performs well compared to the existing methods, shown in the last two rows of Table 1. If *some metadata entity is missing during training*, e.g., disease, we can still use it by excluding the particular information and utilizing the rest of metadata, shown in Table 4. If *some metadata entity is missing during inference*, e.g., vendor, we can run the model for each vendor Philips, Siemens, and GE used to train the model and then average the results to enhance the robustness by leveraging ensemble learning.

6. Conclusion

We proposed a compositional approach that simultaneously localizes the heart in cardiac images using a super-segmentation decoder and does part-based segmentation of different regions of interest through a sub-segmentation decoder. To leverage the image-specific metadata, we also propose a CMFI module to integrate metadata into the segmentation network, guiding it with patterns associated with intensity images to improve performance. Extensive ablation studies indicate the efficacy of each proposed approach and compare it against existing state-of-the-art methods. The experiments are performed on two different modalities, MRI and ultrasound, using M&Ms-2 and CAMUS datasets, respectively, to show that our approach works well in different domains thanks to the proposed compositional approach and

Table 3. Ablation studies comparing performance W/ and WO/ super-segmentation decoder and metadata with FiLM [23] or CMFI using a five-fold cross-validation split of the M&Ms-2 dataset. Our proposed method with the super-segmentation and CMFI is in the bottom row.

Super Segmentation	FiLM	CMFI	LV-Dice	RV-Dice	Myo-Dice	Avg-Dice	LV-HD	RV-HD	Myo-HD	Avg-HD
			90.57	88.05	82.19	86.93	6.94	8.41	5.22	6.85
✓			94.48	90.29	85.31	90.02	3.55	5.19	2.64	3.80
✓	✓		95.10	91.10	86.04	90.74	3.98	5.18	**2.15**	3.77
✓		✓	**95.63**	**91.93**	**87.61**	**91.72**	**3.16**	**4.62**	2.95	**3.57**

Table 4. Performance comparison WO/ and W/ disease inclusion in metadata. The experiments are conducted on a five-fold cross-validation split of the M&Ms-2 dataset.

	WO/ inclusion of disease				**W/ inclusion of disease**			
Methods	LA-Dice	RV-Dice	Myo-Dice	Avg-Dice	LA-Dice	RV-Dice	Myo-Dice	Avg-Dice
Metadata Utilization W/ FiLM Method	94.87	90.91	85.45	90.41	**95.10**	**91.10**	**86.04**	**90.74**
Metadata Utilization W/ CMFI module	95.61	91.01	86.72	91.11	**95.63**	**91.93**	**87.61**	**91.72**

CMFI module.

References

[1] Michael R Avendi, Arash Kheradvar, and Hamid Jafarkhani. A combined deep-learning and deformable model approach to fully automatic segmentation of the left ventricle in cardiac mri. *Medical image analysis*, 30:108–119, 2016. 2

[2] Christian F Baumgartner, Lisa M Koch, Marc Pollefeys, and Ender Konukoglu. An exploration of 2d and 3d deep learning techniques for cardiac mr image segmentation. In *8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers 8*, pages 111–119. Springer, 2018. 2

[3] 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. 2, 5

[4] 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. 7

[5] Jérôme Caudron, Jeannette Fares, Valentin Lefebvre, Pierre-Hugues Vivier, Caroline Petitjean, and Jean-Nicolas Dacher. Cardiac mri assessment of right ventricular function in acquired heart disease: factors of variability. *Academic radiology*, 19(8):991–1002, 2012. 1

[6] Jérôme Caudron, Jeannette Fares, Pierre-Hugues Vivier, Valentin Lefebvre, Caroline Petitjean, and Jean-Nicolas Dacher. Diagnostic accuracy and variability of three semi-quantitative methods for assessing right ventricular systolic function from cardiac mri in patients with acquired heart disease. *European radiology*, 21:2111–2120, 2011. 1

[7] 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

[8] 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. 2, 6, 7

[9] 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, 6, 7

[10] Ramazan Ozgur Dogan, Hulya Dogan, Coskun Bayrak, and Temel Kayikcioglu. A two-phase approach using mask r-cnn and 3d u-net for high-accuracy automatic segmentation of pancreas in ct imaging. *Computer Methods and Programs in Biomedicine*, 207:106141, 2021. 2, 3

[11] 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. 2

[12] Christoforos Galazis, Huiyi Wu, Zhuoyu Li, Camille Petri, Anil A Bharath, and Marta Varela. Tempera: Spatial transformer feature pyramid network for cardiac mri segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 268–276. Springer, 2021. 2

[13] 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. 2, 6

[14] 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. 6, 7

[15] 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. 6, 7

[16] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1998. 3

[17] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 3

[18] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 2, 6

[19] Sana Jabbar, Syed Talha Bukhari, and Hassan Mohy-ud Din. Multi-view sa-la net: A framework for simultaneous segmentation of rv on multi-view cardiac mr images. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 277–286. Springer, 2021. 2

[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 *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 326–336. Springer, 2021. 2, 6

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[22] Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019. 2, 5, 7

[23] Andreanne Lemay, Charley Gros, Olivier Vincent, Yaou Liu, Joseph Paul Cohen, and Julien Cohen-Adad. Benefits of linear conditioning with metadata for image segmentation. *arXiv preprint arXiv:2102.09582*, 2021. 3, 7, 8, 9

[24] Lei Li, Wangbin Ding, Liqin Huang, and Xiahai Zhuang. Right ventricular segmentation from short-and long-axis mris via information transition. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 259–267. Springer, 2021. 2, 6

[25] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 1

[26] 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 485–495. Springer, 2022. 2, 5, 6

[27] 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. 2, 5

[28] John A Onofrey, Dana I Casetti-Dinescu, Andreas D Lauritzen, Saradwata Sarkar, Rajesh Venkataraman, Richard E Fan, Geoffrey A Sonn, Preston C Sprenkle, Lawrence H Staib, and Xenophon Papademetris. Generalizable multi-site training and testing of deep neural networks using image normalization. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, pages 348–351. IEEE, 2019. 3

[29] Jordan Ringenberg, Makarand Deo, Vijay Devabhaktuni, Omer Berenfeld, Pamela Boyers, and Jeffrey Gold. Fast, accurate, and fully automatic segmentation of the right ventricle in short-axis cardiac mri. *Computerized Medical Imaging and Graphics*, 38(3):190–201, 2014. 2

[30] 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, 6, 7

[31] 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. *arXiv preprint arXiv:2303.15446*, 2023. 3

[32] Jie Xue, Kelei He, Dong Nie, Ehsan Adeli, Zhenshan Shi, Seong-Whan Lee, Yuanjie Zheng, Xiyu Liu, Dengwang Li, and Dinggang Shen. Cascaded multitask 3-d fully convolutional networks for pancreas segmentation. *IEEE Transactions on Cybernetics*, 51(4):2153–2165, 2019. 3, 6

[33] Yiwen Zhang, Haoran Lai, and Wei Yang. Cascade unet and ch-unet for thyroid nodule segmentation and benign and malignant classification. In *Segmentation, Classification, and Registration of Multi-modality Medical Imaging Data: MICCAI 2020 Challenges, ABCs 2020, L2R 2020, TN-SCUI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 23*, pages 129–134. Springer, 2021. 2

[34] Qiao Zheng, Hervé Delingette, Nicolas Duchateau, and Nicholas Ayache. 3-d consistent and robust segmentation of cardiac images by deep learning with spatial propagation. *IEEE transactions on medical imaging*, 37(9):2137–2148, 2018. 2