Article:
Vesal, S, Gu, M, Maier, A et al. (1 more author) (2020) Spatio-temporal Multi-task Learning for Cardiac MRI Left Ventricle Quantification. IEEE Journal of Biomedical and Health Informatics. ISSN 2168-2194

https://doi.org/10.1109/jbhi.2020.3046449

Spatio-temporal Multi-task Learning for Cardiac MRI Left Ventricle Quantification

Sulaiman Vesal, Mingxuan Gu, Andreas Maier *Member, IEEE*, and Nishant Ravikumar

***Abstract*—Quantitative assessment of cardiac left ventricle (LV) morphology is essential to assess cardiac function and improve the diagnosis of different cardiovascular diseases. In current clinical practice, LV quantification depends on the measurement of myocardial shape indices, which is usually achieved by manual contouring of the endo- and epicardial. However, this process subjected to inter and intra-observer variability, and it is a time-consuming and tedious task. In this paper, we propose a spatio-temporal multi-task learning approach to obtain a complete set of measurements quantifying cardiac LV morphology, regional-wall thickness (RWT), and additionally detecting the cardiac phase cycle (systole and diastole) for a given 3D Cine-magnetic resonance (MR) image sequence. We first segment cardiac LVs using an encoder-decoder network and then introduce a multitask framework to regress 11 LV indices and classify the cardiac phase, as parallel tasks during model optimization. The proposed deep learning model is based on the 3D spatio-temporal convolutions, which extract spatial and temporal features from MR images. We demonstrate the efficacy of the proposed method using cine-MR sequences of 145 subjects and comparing the performance with other state-of-the-art quantification methods. The proposed method obtained high prediction accuracy, with an average mean absolute error (MAE) of 129 mm^2, 1.23 mm, 1.76 mm, Pearson correlation coefficient (PCC) of 96.4%, 87.2%, and 97.5% for LV and myocardium (Myo) cavity regions, 6 RWTs, 3 LV dimensions, and an error rate of 9.0% for phase classification. The experimental results highlight the robustness of the proposed method, despite varying degrees of cardiac morphology, image appearance, and low contrast in the cardiac MR sequences.**

***Index Terms*—Left Ventricle Quantification, Cardiac MRI, Cardiac Segmentation, Deep Learning, Myocardial Infraction**

I. Introduction

Cardiovascular diseases (CVDs) and other cardiac pathologies are the leading cause of death worldwide [1], [2], [3]. Timely diagnosis is crucial for improving survival rates and delivering high-quality patient care. Cardiac magnetic resonance imaging (MRI) is a non-invasive imaging modality used to detect and monitor cardiovascular diseases. Quantitative assessment and analysis of cardiac-MR images are indispensable for diagnosis and devising suitable treatments. The reliability of quantitative metrics that characterize cardiac function such as myocardial deformation and ventricular ejection fraction are heavily dependent on the precision of ventricle quantification [4].

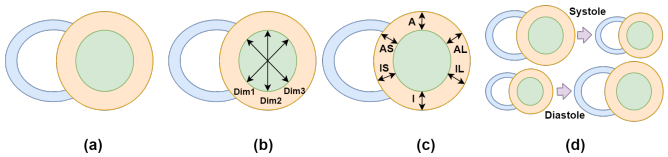


Fig. 1. A schematic representation of LV indices for short-axis cardiac Cine-MR image. The LV and Myo cavity areas are shown with green and blue colors in (a) and three LV cavity dimensions with black arrows in (b). (c) shows Six myocardial regional-wall thicknesses, namely anterolateral (AL), inferolateral (IL), inferior (I), inferoseptal (IS), anterior (A), inferoseptal (IS) and anteroseptal (AS). The cardiac phase (systole or diastole) is shown in (d).

In everyday clinical practice, evaluation of LV function is often conducted by visual assessment and semi-automatic tools to quantify dynamics in MRI [5], [6] [7]. Hence, clinical evaluation of regional LV function is mostly qualitative and by visually observing myocardial wall displacement and deformation. Naturally, this process can be error-prone either due to artifacts arising from cardiac, respiratory or patient motion, variations in image contrast, or human error. This may prevent an accurate evaluation of LV structures. On the other hand, LV assessment by cardiologists requires extensive expertise and experience [8], [9], [10]. A central part of morphological cardiac quantification involves manual/semi-automatic contouring of the endo- and epicardial walls of the left ventricular myocardium. It is time-consuming and often subjected to high intra and inter-observer variability. Moreover, the myocardium contouring process is performed on the end-systolic (ES), and end-diastolic (ED) frames that are inadequate for comprehensive analysis of heart function (across the full cardiac phase) [5]. Notwithstanding recent advances, LV segmentation is still a challenging problem due to limited contrast between tissue boundaries, and pathology-driven variability in shape and appearance in cine-MR sequences. [11], [12].

To address these challenges, we propose a deep learning approach to enable automatic full quantification of LV morphology in short-axis cardiac cine-MR images without any further information. We investigate the use of temporal and spatial information to estimate the cardiac phase, diameters of the LV blood pool (or cavity) along with different directions, regional wall thicknesses (RWTs) (as depicted in Fig. 1), and LV cavity and myocardial areas. Comprehensive assessment

S. Vesal, M. Gu, and A. Maier are with the Pattern Recognition Lab, Friedrich-Alexander-University Erlangen-Nuremberg, Germany. (E-mail: sulaiman.vesal@fau.de)

N. Ravikumar is with CISTIB, Centre for Computational Imaging and Simulation Technologies in Biomedicine, School of Computing, LICAMM Leeds Institute of Cardiovascular and Metabolic Medicine, School of Medicine, University of Leeds, United Kingdom.

The work described in this paper was partially supported by the project EFI-BIG-THERA: Integrative 'BigData Modeling' for the development of novel therapeutic approaches for breast cancer. The authors would also like to thank NVIDIA for donating a Titan X-Pascal GPU.

of these measures requires analysis of images from the entire $2D + t$ cine-MR sequence (covering the full cardiac cycle), thereby considering the temporal dynamics of the heart.

II. Related Work

In recent years, many studies have focused on automatic full cardiac LV morphological quantification. These methods are designed either as multi-stage or end-to-end approaches, in terms of their training strategy. Traditional multi-stage methods are mainly based on myocardium segmentation [13], [14], [15], [16], where first the LV endocardium and epicardium are segmented and then the desired LV indices are estimated based on segmentation masks. The latter takes advantage of machine learning algorithms [17], [18], where features are extracted automatically from cardiac MR images, and a regression model utilizes these features to estimate the LV indices. In comparison to the multi-stage methods, end-to-end methods [13], [19], [20] combine feature extraction and regression together using deep neural networks.

One of the earliest works for LV quantification based on manual segmentation proposed by Suinesiaputra *et al.* [14]. They asked seven cardiologists to manually delineate contours around myocardium and LV cavity volume at the ES and ED phases to evaluate cardiac function. As we know, manual segmentation is very time-consuming, subjective, and not very efficient. To tackle these limitations, several automatic segmentation algorithms [21], [22], [23], [24] have been proposed. Wang *et al.* [25] considered a Bayesian method for two-ventricular volume estimation that used a likelihood function for exploiting appearance features and a probability model to incorporate the area correlation between the cavities. Zhen *et al.* [18] initially extracted hierarchical profiles using multi-scale deep neural networks and then placed them in a random regression forest to estimate the left ventricle. In this type of two-step approach, there is only a forward linkage and no feedback from the second step. Therefore the features extracted in the first stage may not be closely related to the target tasks, and the estimated results in the second stage may not be very accurate.

Moreover, the direct regression-based methods have been also used to quantify LV indices either in an end-to-end or multi-stage fashion. Xue *et al.* [26] proposed an integrated model to present multiple LV criteria, including two regions and six RWTs per frame within the cardiac cycle. To model the temporal dynamics of cardiac sequences, they employed a Recurrent Neural Network (RNN) followed by a Convolutional Neural Network (CNN) module to regress six RWTs [19].Additionally, Xue *et al.* [20] focused on the quantification of complete LV measurements, which requires estimating the regions, orientation dimensions, and RWTs simultaneously for each MR image. To improve the prediction accuracy, they used both CNN and RNN modules and modeled the correlations between the different LV criteria with a multi-task loss. Nevertheless, these methods don't process the whole cardiac cycle as a whole for feature extraction, but rather an embedding of 3-5 neighboring MR frames to incorporate temporal information. Therefore, they do not guarantee the temporal dynamic consistency of the estimated volumes across the whole cardiac cycle. Wang *et al.* [13] proposed a cascaded segmentation and regression network, in which the segmentation component extracts left ventricular myocardial contours, and the regression component estimates the desired LV criteria. However, this method only computes the LV indices and not the cardiac phase cycle. Another study [22] processes stacks of adjacent slices($k = 5$) using 3D convolutional kernels to incorporate temporal information within the learned model. Tao *et al.* [12] in a more clinically adapted approach tested three different CNNs for fully automated quantification of LV on multi-centers and multi-vendors study. However, this work was performed on retrospective data and not cover a wide range of cardiovascular abnormalities, which is clinically more demanding. Recently, disentanglement representation learning methods also investigated [27] [28] to extract generalize features within a multi-task framework. The model encodes informative features for different tasks and employing the adversarial regularization to enforce the extracted features to be minimally informative about irrelevant tasks.

Inspired by previous works, and to address the challenges as yet unmet by current methods for LV quantification, we propose a novel end-to-end multi-task learning framework based on $2D + t$ spatio-temporal convolutions to simultaneously tackle multiple tasks and allow them mutually learn from each other [29], [30],[31]. The proposed approach permits accurate quantification of standard LV indices and provides 3D segmentation for the blood pool and myocardium of the LV for further morphological analysis. Introducing a single model that is capable of solving multiple tasks at the same time can be clinically very relevant and reduce the overhead for the cardiologists to have individual models for each sub-task including ventricle segmentation and quantification.

In summary, our main contributions are three folds:

- First, we propose an end-to-end deep learning model that directly learns temporal and spatial features using 3D spatio-temporal convolutions from the estimated 3D cine-MR segmentation masks. The proposed model takes the full temporal cine-MR sequence into account, to quantify the LV, rather than a single 2D image or by concatenating a few 2D image slices from adjacent time frames (i.e. 2.5D).
- Second, a multi-task network is introduced to leverage the shared information useful for LV segmentation, LV indices regression, and cardiac phase cycle classification tasks, by jointly optimise all three. We further demonstrate with empirical evidence that the temporal information and volumetric quantification improves prediction accuracy significantly compared to 2D and 2.5D deep learning models.
- Third, we validate our proposed method using the publicly available LVQuan 2018 benchmark dataset, which provides short-axis cine-MR sequences with annotations for the above indices (for the whole cardiac cycle). The proposed method achieved better robustness and interpre-

tation for LV quantification and morphology assessment in comparison to state-of-the-art methods. The code and model are available at: https://github.com/sulaimanvesal/CardiacQuanNet

III. Materials and Methods

In this section, we first describe the details about the dataset. Further, we introduce our proposed Spaito-temporal multi-task learning network pipeline and its component for LV segmentation, regression, and classification. Eventually, the objective function and training settings are described.

A. Dataset

We validated our proposed framework on the STACOM LVQuan 2018 challenge training dataset [26]. The dataset collected from 3 different hospitals and in collaboration with two healthcare centers, namely London Healthcare Center and St. Josephs Healthcare [19], [24]. It consists of 2D Cine-MR images of $n_S = 145$ patients with an average age of 58.9 years. The Cine-MR images have a pixel spacing ranges between 2.0833 mm/pixel and 0.6836 mm/pixel. The dataset has a set of various pathologies like myocardial hypertrophy, regional-wall motion abnormalities, atrial septal defect, mildly enlarged LV, LV dysfunction, etc. Each Cine-MR sequence has 20 frames per cycle, resulting in a total of 2900 images in the training dataset. Following the standard American Heart Association (AHA) recommendation [32], each frame in the dataset includes only mid-cavity regions, which is perpendicular to the long axis of the heart.

All cardiac images annotated manually to obtain the epicardium and endocardium boundaries, which are double-checked by two experienced cardiologists. The ground truth values of LV indices are computed based on these delineations. The RWTs and LV blood pool dimensions indices are normalized with respect to the image size, while the areas normalized by the number of pixels (2900). After the training step, the computed indices are converted back to physical thickness (mm) and area (mm^2) by changing the resizing procedure and multiplying each subject with their corresponding pixel spacing. To evaluate and compare model performance, we employed five-fold cross-validation similar to other studies. Offline data augmentation was used by randomly rotating, flipping horizontally/vertically, and applying elastic deformation to the training images. This process increased the number of training samples in each fold by a factor of 8.

B. Network Architecture

As discussed previously, our LV quantification framework consists of two modules: a segmentation network $\mathbf{G}$ and the multi-task classification/regression network $\mathbf{D}$. During training, we first provide an MR image sequence $\mathcal{I}_S \in \mathbb{R}^{t\times h\times w\times 1}$ (with annotations) to the segmentation network for optimizing $\mathbf{G}$. Then, we convert the soft probabilities of the *softmax* layer to hard probabilities, and provide the LV and Myo segmentation predictions ($\mathcal{P}_S \in \mathbb{R}^{t\times h\times w\times 2}$) as the input to $\mathbf{D}$. Here, 2 refers to image channels corresponding to the LV and Myo segmentation masks, and we discard the background channel as the indices are computed from LV and Myo channels only. The network propagates gradients from $\mathbf{D}$ to $\mathbf{G}$, which in turn encourages $\mathbf{G}$ to optimize its weights with respect to both the segmentation labels (tissue boundaries) and the LV indices of interest. Fig. 2 shows an overview of the proposed algorithm. In this section, we first describe the left ventricle segmentation module and subsequently, the multi-task network for regression and classification of 11 indices and cardiac phase recognition, respectively. The $\mathbf{G}$ and $\mathbf{D}$ modules are trained using two strategies: (1) multi-stage and (2) end-to-end. In an end-to-end fashion, we optimize both networks simultaneously.

C. Left Ventricle Segmentation

To segment the LV blood pool and myocardium of the LV, we employ a fully convolutional network architecture inspired by [33] called Dilated Residual-UNet (DR-UNet), which is depicted in Fig. 3. The segmentation network $\mathbf{G}$ has an encoder and decoder paths that are connected by a bottleneck block. Every block in encoder and decoder paths has two 2D convolution layers followed by a Rectified Linear Unit (ReLU), batch-normalization, and a 2D max-pooling layer to reduce the dimensions of feature maps. To improve the flow of gradients, and enforcing the encoder to extract more discriminative features, a residual connection [34] added in each encoder block. The last layer of the network has a *softmax* activation function to produce probability segmentation maps for each class. In DR-UNet, the normal convolution layers are replaced with dilated convolutions to permit the network to capture both global and local contextual information by increasing the respective field. A sequence of dilated convolutions can introduce gridding effects (different output nodes use disjoint subsets of input nodes) if dilation rates are not selected properly [35]. As a consequence, we have created a block of stacked dilated convolutions whose outputs are summed together. This way, each subsequent layer has full access to previous features learned using different dilation rates, addressing the issue of gridding artefacts. In our network settings, we used four dilated convolutions with a dilation rate of $1-8$ in the network bottleneck. In the case of the end-to-end training strategy, we adapted the segmentation network $\mathbf{G}$ from 2D to 3D by replacing all the 2D convolution operations with 3D convolution layers, while the rest of the network remained the same. It is because we consider the full temporal sequence as the input for the regression and classification task, resulting in an input tensor size of $20\times80\times80\times2$. On the other hand, the 3D spatio-temporal module for classification and regression have 3D kernels which required 3D input. Therefore, we selected DR-UNet 3D as the segmentation backbone.

D. Left Ventricle Quantification

To quantify the cardiac LV, we propose a multi-task CNN architecture, which is trained both in an end-to-end and multi-stage fashion. This network employs spatio-temporal convolutions to consider not only spatial but also temporal

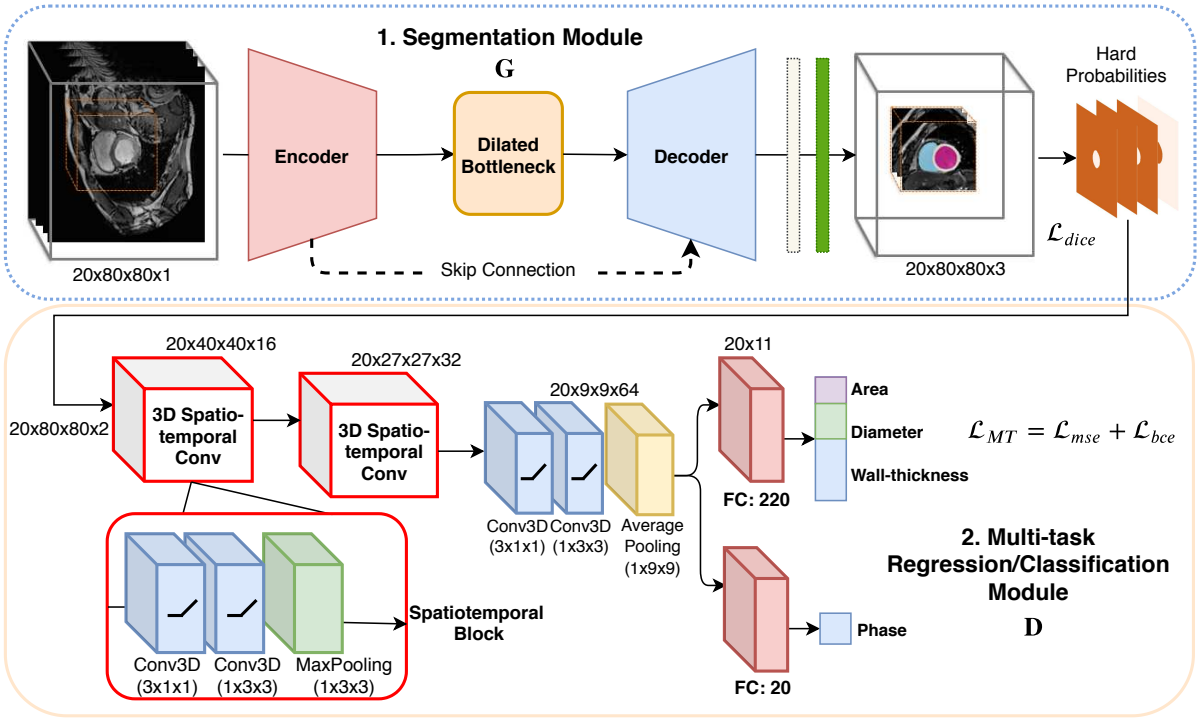


Fig. 2. Network architecture overview. Given Cine-MR volumes with the size $t \times h \times w$ from the training dataset as input, we pass it through the segmentation network to obtain output segmentation masks for the LV cavity and Myo. A segmentation loss is computed based on the ground truth. To make predictions for 11 indices and cardiac phase detection, we utilize a multi-task spatio-temporal network with two parallel branches. Then a multi-task loss is calculated on the target prediction for both regression and classification tasks and is back-propagated to the segmentation network.

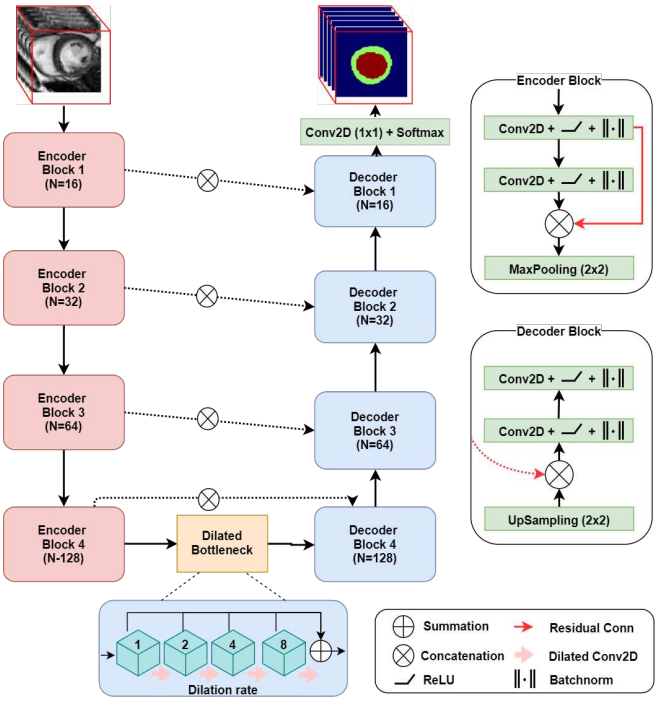


Fig. 3. The schematic illustration of our left ventricle segmentation network equipped with dilated convolutions(4 dilation rate 1-8) in the bottleneck to capture multi-scale features and residual blocks in the encoder path.

information. 3D CNNs applied to $2D + t$ image frames to preserve temporal information and propagate it through the layers of the network [36], [37]. The cardiac Cine-MR dataset has a short-term temporal dynamic between neighboring slices in the sequence over one cardiac cycle. For this reason, we consider 3D information as features for temporal modeling on the $2D + t$ MR images. The 3D convolution addresses each image with assistance from adjacent slices, and it can represent more robust structural features as well as temporal information [38].

Fig. 2 illustrates our proposed spatio-temporal network architecture. The proposed network consists of three spatio-temporal blocks and two task-specific parallel branches. The first branch computes 11 indices for LV areas, LV dimensions, and RWTs, and the second branch classifies the cardiac phase across entire sequences. Cardiac phase-detection is normally considered as a sequence modeling task as the temporal dynamics are important for determining the cardiac phase. RNN blocks are widely used for this type of task [24], but these models are difficult to optimize. The proposed spatio-temporal blocks on the other hand, already take into account both the spatial and temporal dynamics of slices, which removes the need for RNN blocks.

The input to the encoder part has a size of $t \times h \times w$, where t is the temporal axis, and $h \times w$ are the spatial axes. Each spatio-temporal block has two 3D convolution layers and a subsequent 3D MaxPooling layer. The first 3D convolution layer has a kernel size of $3 \times 1 \times 1$ that captures temporal features across the time axis. The second convolution layer extracts spatial features with a kernel size of $1 \times 3 \times 3$ and strides of 1. The MaxPooling layer has a window size of $1 \times 3 \times 3$ as we only want to downsample the inputs along

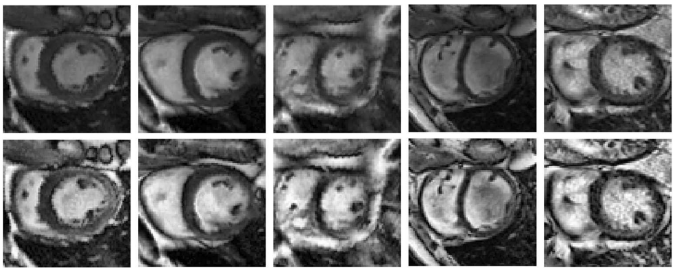


Fig. 4. LV MR image slices before(first row) and after pre-processing and normalisation (second row).

the spatial dimensions while keeping the number of frames fixed (as we would like to compute LV indices for each frame). In comparison to fully 3D convolution operation, this decomposition offers two advantages as highlighted in [36]. First, in this setup, the number of parameters is not changed. However, it increases the number of nonlinearities in the network due to the additional ReLU between the two convolution layers in each block. Doubling the number of nonlinearities enhances the complexity of functions, which approximate the effect of a big filter by applying multiple smaller filters with additional nonlinearities in between. The second benefit is that forcing the 3D convolution into separate spatial and temporal components makes the optimization easier [36]. This in turn helps reduce the error rate compared to conventional 3D CNNs of the same capacity [36]. For all convolution layers, we initialize the kernels with the He initializer [39] and employ $\mathcal{L}_2$ weight regularization to reduce the overfitting of the proposed model on the training data.

E. Pre-processing

All cardiac images were preprocessed by the challenge organizer, including landmark labeling to find the ROI, rotation to align the volumes, ROI cropping, and resizing. The resulting images are 80 $\times$ 80 in size. The LVQuan 2018 dataset images vary a lot in terms of contrast and brightness. The variability results from different acquisition parameters and scanners are always a challenge for designing a robust and generalized neural network model. Contrast limited adaptive histogram equalization (CLAHE) is applied to further improve the contrast of Cine-MR images as well as reducing the variability across the dataset, particularly for those images with low contrast [40]. Subsequently, the images were normalized by subtracting the mean and dividing by the standard deviation for each sequence. These prepossessing steps significantly improved the segmentation accuracy for DR-UNet. Fig. 4 presents a few sample images before and after employing CLAHE and image normalization. From these images, it is evident that variability in brightness and contrast is largely diminished following contrast enhancement.

Target Labels Normalisation: In the LVQuan 2018 dataset, there is a large amount of variation in the magnitude of the various indices. Different distributions of different target indices can cause unbalanced and unstable training. To tackle this issue, we normalize the label indices using the z-score for all the 11 indices (2 LV and Myo areas, 3 LV blood pool dimensions, and 6 RWTs) across the whole dataset. After training for the final evaluation, we scale the indices back to the original value by multiplying the standard deviation and adding the mean value. It should be noted that this normalization step is only for quantification labels, while the image normalization process described in the preprocessing section.

F. Loss Functions

Segmentation Loss: The segmentation network $\mathbf{G}$ is trained with a multi-class soft-Dice loss, which shown to be less sensitive when there is a huge class imbalance within the dataset in comparison to binary cross-entropy loss. Many recent studies [33], [41] also used this objective function for medical image segmentation. We first compute the Dice loss for every class individually and then average it over the number of available classes. To segment a Cine-MR image $\mathcal{I}_S \in \mathbb{R}^{t\times h\times w\times 1}$ with having LV, Myo and background as labels, the output of *Softmax* layer is three probability maps for classes $k = 0, 1, 2$ where for each pixel $\sum_c \boldsymbol{y}_{n,k} = 1$. Given the ground-truth label $\hat{\boldsymbol{y}}_{n,k}$ for that identical pixel, the multi-class soft Dice loss is computed as follows:

$$\mathcal{L}_{dice}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = 1 - \frac{1}{K}(\sum_k \boldsymbol{w}_k \frac{\sum_n \boldsymbol{y}_{nk}\hat{\boldsymbol{y}}_{nk}}{\sum_n \boldsymbol{y}_{nk} + \sum_n \hat{\boldsymbol{y}}_{nk}}) \quad (1)$$

where $\boldsymbol{w}_k$ is the weight factor to tackle class imbalance as Myo region has fewer pixels compared to the other two classes. We empirically set the weights for each class: $\{BG : 0.2, LV : 0.3, Myo : 0.5\}$. We achieved better segmentation performance by weighting the Myo class higher.

Classification Loss: Given the class probability output $\mathcal{P}_{phase} = \mathbf{D}(\mathcal{I}_S)$ from the cardiac phase classification branch, the cross-entropy loss $\mathcal{L}_{bce}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ for the two classes (i.e., ED and ES) can be written as:

$$\mathcal{L}_{bce}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{y}_i).log(\hat{\boldsymbol{y}}_i) + (1 - \boldsymbol{y}_i).log(1 - \hat{\boldsymbol{y}}_i) \quad (2)$$

where, $\boldsymbol{y}$ is the label (1 for ED phase and 0 for ES phase) and $\hat{\boldsymbol{y}}$ is the predicted probability.

Multi-Task Loss: To train the multi-task regression and classification module $\mathbf{D}$, we optimized the model using a joint loss, combining $\mathcal{L}_{bce}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ and Mean Squared Error (MSE), for cardiac phase classification and LV indices regression, respectively. The loss is formulated as:

$$\mathcal{L}_{mse}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\frac{1}{N}\sum_{s=1}^{11}\sum_{i=1}^{N}||\boldsymbol{y}_{s,i} - \hat{\boldsymbol{y}}_{s,i}||_2^2 \quad (3)$$

$$\mathcal{L}_{mt}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \underset{\hat{y}}{\operatorname{argmin}}\, \lambda_1.\mathcal{L}_{mse}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \lambda_2.\mathcal{L}_{bce}(\boldsymbol{y}, \hat{\boldsymbol{y}}) \quad (4)$$

where, λ_1 and λ_2 are the weights to control the influence of the individual tasks on the combined loss. We empirically set $\lambda_1 = 1$ and $\lambda_2 = 4$. This is due to the fast convergence of

$\mathcal{L}_{mse}$, which necessitated higher weights for the classification task in order to stabilize the training process. Equation (4) can be extended further to train the entire pipeline including the segmentation network $\mathbf{G}$, in an end-to-end fashion:

$$\mathcal{L}_{mt}(\boldsymbol{y},\hat{\boldsymbol{y}}) = \underset{\hat{y}}{\operatorname{argmin}}\, \lambda_1.\mathcal{L}_{dice}(\boldsymbol{y},\hat{\boldsymbol{y}}) + \lambda_2.\mathcal{L}_{mse}(\boldsymbol{y},\hat{\boldsymbol{y}}) + \lambda_3.\mathcal{L}_{bce}(\boldsymbol{y},\hat{\boldsymbol{y}}) \tag{5}$$

Here, we combine the segmentation loss $\mathcal{L}_{dice}$ with the multi-task loss for regression and classification tasks. However, to have control over the influence of different losses and gradients in $\mathcal{L}_{mt}$, we have empirically set $\lambda_1 = 10$, $\lambda_2 = 1$ and $\lambda_3 = 1$ as weights. Since the regression and classification part depends on the output of the segmentation network, we have given more weights to this task. This enforces the method to produce more accurate myocardial segmentation. Fig. 5 illustrates a simplified view of our framework, which shows the flow of gradients between different modules and how we train all the three tasks end-to-end.

G. Network Training

As discussed previously, we trained our LV quantification pipeline using both multi-stage and end-to-end strategies. In each training batch, the MR sequence $\mathcal{I}_S$ is provided to the segmentation network $\mathbf{G}$ which is trained by optimizing $\mathcal{L}_{dice}(\boldsymbol{y},\hat{\boldsymbol{y}})$ in (Eq. 1), to generate the output probability map P_s. These soft probability masks are subsequently converted to hard probabilities and passed on to the multi-task network $\mathbf{D}$. The latter in turn is trained by optimizing $\mathcal{L}_{mt}$ in (Eq 4). The predicted normalized LV indices are subsequently converted back to physical values before evaluation.

The performance of our framework evaluated across 5-fold cross-validation experiments. In each fold, there are 29 subjects. Four folds used to train the spatio-temporal multi-task learning model and the fifth to test. We repeated the same procedure five times until the LV indices of all subjects were obtained. The proposed deep learning model designed and developed in Keras and TensorFlow [42], which is an open-source deep-learning library for Python. All the networks trained on an NVIDIA Titan X-Pascal GPU with 12GB memory. In the multi-stage strategy, we trained the segmentation network $\mathbf{G}$ using the ADAM optimizer [43]. A fixed learning rate of 0.0001 with exponential decay rates of the 1^{st} is used, and Adam momentum parameters were set to 0.9 and 0.999, respectively. The multi-task network $\mathbf{D}$ was also trained using Adam optimizer with a learning rate of 0.004, and with similar decay rate and momentum as $\mathbf{G}$. For the end-to-end training strategy, we employed the same optimizer and hyperparameters to make both the models comparable.

IV. Experiments and Results

A. Evaluation Metrics

To evaluate the performance of the methods quantitatively, we used the Pearson correlation coefficient (PCC) and Mean Absolute Error (MAE) metrics. For the cardiac LV indices MAE and PCC are computed as follows:

$$MAE_{ind} = \frac{1}{N}\sum_{i=1}^{N} |\hat{\boldsymbol{y}}^i_{ind} - \boldsymbol{y}^i_{ind}| \tag{6}$$

$$PCC_{ind} = \frac{\sum_{i=1}^{N}(\hat{\boldsymbol{y}}^i_{ind} - \bar{\hat{\boldsymbol{y}}}^i_{ind})(\boldsymbol{y}^i_{ind} - \bar{\boldsymbol{y}}^i_{ind})}{\sqrt{\sum_{i=1}^{N}(\hat{\boldsymbol{y}}^i_{ind} - \bar{\hat{\boldsymbol{y}}}_{ind})^2(\boldsymbol{y}^i_{ind} - \bar{\boldsymbol{y}}_{ind})^2}} \tag{7}$$

where, $ind \in (A1, A2, D1...D3, RWT1...RWT6)$, $\hat{\boldsymbol{y}}_{ind}$ is the estimated value by the model and $\boldsymbol{y}_{ind}$ is the ground-truth value provided by the rater. Here, $\bar{\boldsymbol{y}}_{ind}$ and $\bar{\hat{\boldsymbol{y}}}_{ind}$ are their mean values, respectively.

To evaluate and assess the model performance for cardiac phase classification, we used the Error Rate (ER), which is defined as:

$$ER_{phase} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\boldsymbol{y}}^i_{phase} \neq \boldsymbol{y}^i_{phase})100\% \tag{8}$$

where $\boldsymbol{y}_{phase}$ and $\hat{\boldsymbol{y}}_{phase}$ are the ground-truth and estimated classes for the cardiac phase, respectively. To evaluate the accuracy of the segmentation results, we used the well-known metrics in the medical image segmentation field, the Dice coefficient (Dice) score, and Hausdorff distance (HD). [33].

B. Comparison With State-of-the-art Methods

LV Segmentation Performance: All segmentation networks were evaluated with and without the connected component analysis (CCA) as postprocessing step and weighted-class loss function, respectively. Table I summarizes the results for our segmentation module $\mathbf{G}$ under different settings. We observe that DR-UNet 2D and DR-UNet 3D achieved high segmentation accuracy for the LV blood pool and background classes in terms of Dice score and HD value. However, it was less successful for the Myo, primarily due to the presence of noise, low contrast tissues, and different pathologies. The 2D DR-UNet with eight filters without CCA and weighted class average loss achieved an average Dice of 85.0% on the validation set for Myo. However, by including both operations and increasing the number of filters to 16, the average Dice for Myo improved to 89.0%. HD value also reduced to 4.87 mm, respectively. Accurate segmentation of Myo is crucial since most LV functional indices are computed based on the endo- and epicardial contours. A similar performance gain was achieved by DR-UNet 3D. DR-UNet 3D with CCA and weighted-loss achieved an average Dice score of 88.7% for Myo, but a substantially higher HD value of 3.71 mm. Fig. 6 depicts the segmentation output for the complete cardiac cycle of a patient and segmentation error w.r.t the ground truth. It can be seen that the predicted contours for LV and Myo are precise and have a very low error-rate at tissue boundaries. Table I also demonstrates the intermediate segmentation results for DR-UNet 3D and DR-UNet 2D when the models were trained end-to-end along with the regression and classification tasks (rows 8-11). Interestingly, DR-UNet 3D and 3D spatio-temporal network achieved a dice score of

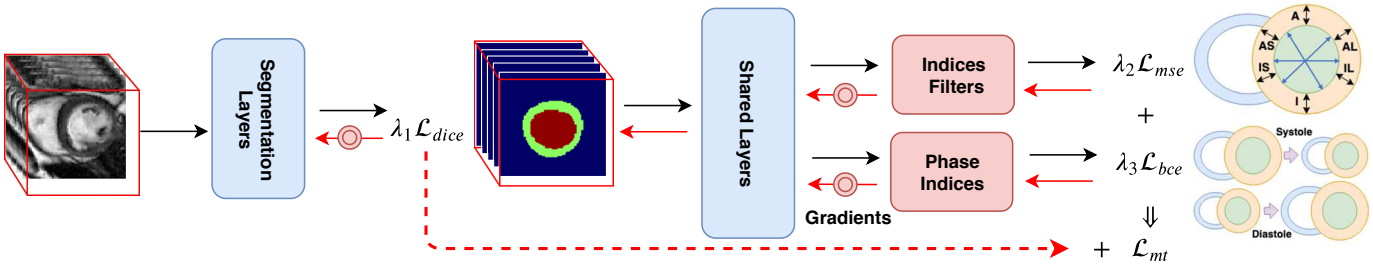


Fig. 5. Representation of gradient flow between segmentation, regression and classification tasks in our cardiac quantification framework.

TABLE I
SEGMENTATION ACCURACY USING DIFFERENT EVALUATION METRICS AND TRAINING STRATEGIES. AS AN ABLATION STUDY, THE NUMBER OF FILTERS, POST-PROCESSING AND LOSS FUNCTION HAVE BEEN CHANGED TO EVALUATE THE PERFORMANCE.HERE, CONNECTED COMPONENT ANALYSIS IS ABBREVIATED AS CCA.

Experiments	Filters	CCA	Weighted-loss	Dice Score ↑			HD [mm] ↓			Parameters
				Bg	LV	Myo	Bg	LV	Myo	
CSRNet [13]	-	-	-	0.989	0.959	0.886	4.88	3.55	5.43	0.3 M
DR-UNet 2D	8	×	×	0.978	0.945	0.854	4.54	4.65	11.92	0.9 M
	16	×	×	0.982	0.950	0.868	4.03	5.76	11.53	3.6 M
	8	✓	×	0.981	0.951	0.857	4.30	4.33	11.92	0.9 M
	16	✓	×	0.983	0.954	0.871	3.67	4.30	10.51	3.6 M
	16	✓	✓	**0.990**	0.959	0.888	3.34	3.63	4.87	3.6 M
DR-UNet 3D	16	✓	✓	0.989	0.958	0.887	4.57	3.71	5.04	3.6 M
End-to-End Training										
DR-UNet 2D & 3D CNN	16	✓	✓	0.990	**0.959**	0.889	4.01	3.60	3.90	3.6 M
DR-UNet 2D & 3D spatio-temporal	16	✓	✓	0.989	0.957	0.881	4.05	3.50	4.50	3.6 M
DR-UNet 3D & 3D CNN	16	✓	✓	0.990	0.959	0.883	3.80	3.35	4.03	3.6 M
R-UNet 3D & 3D spatio-temporal	16	✓	✓	0.990	0.957	**0.889**	**3.29**	**2.85**	**3.40**	3.6 M

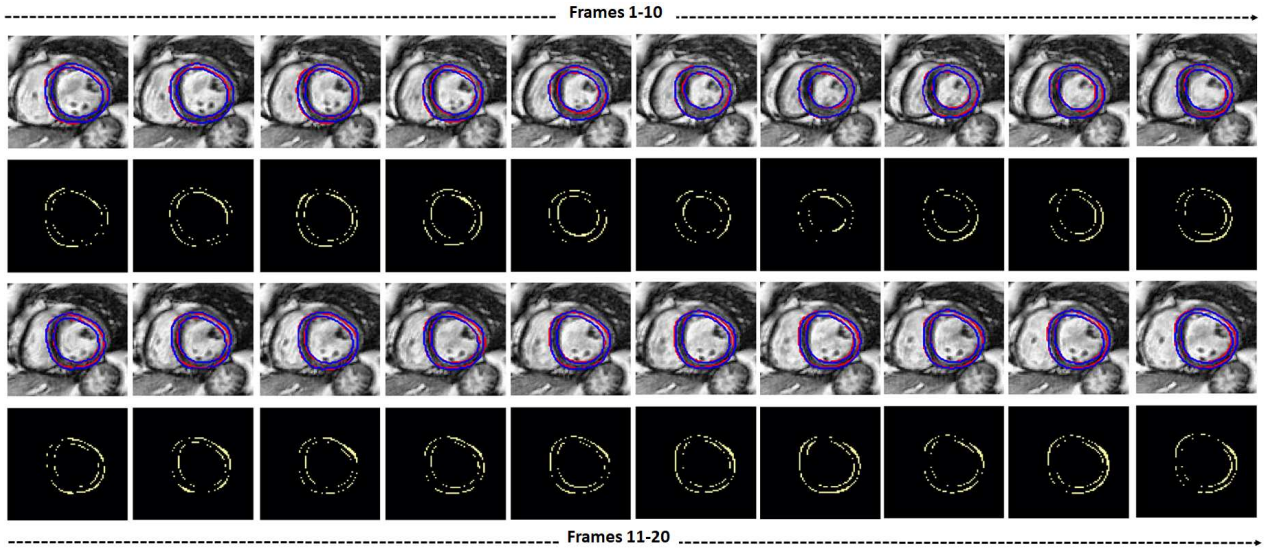


Fig. 6. Qualitative segmentation results of DR-UNet 2D. The red contour shows the ground truth segmentation contour, and the blue color is overlaid as the proposed model prediction output. The row 1-2 shows the first 10 slices with their segmentation error. The row 2-4 illustrate frames 11 to 20 with segmentation error respectively. It can be seen, that segmentation error is quite low for both endocardium and epicardium contours in most of the frames.

89.0% and the lowest HD values of 2.85 mm and 3.40 mm for LV and Myo. It can also confirm the advantage afforded by jointly optimizing three tasks together under a multi-task learning scenario.

LV Quantification Performance: In order to highlight the gain in performance for LV quantification afforded by our approach relative to the state-of-the-art, we compared our approach with six recent methods - Xue *et al.* [19] (Indices-Net) , Xue *et al.* [26] (FullLVNet), Xue *et al.* [24] (DMTRL), Li *et al.* [23] (DLA), and Wang *et al.* [13] (CSRNet) under the same experiment settings. Xue *et al.* introduced Indices-Net to estimate multiple cardiac indices at the same time. The author uses two closely coupled networks: a deep convolutional auto-encoder for feature extraction from cardiac images and a multiple-output CNN for index regression. Xue *et al.* extended

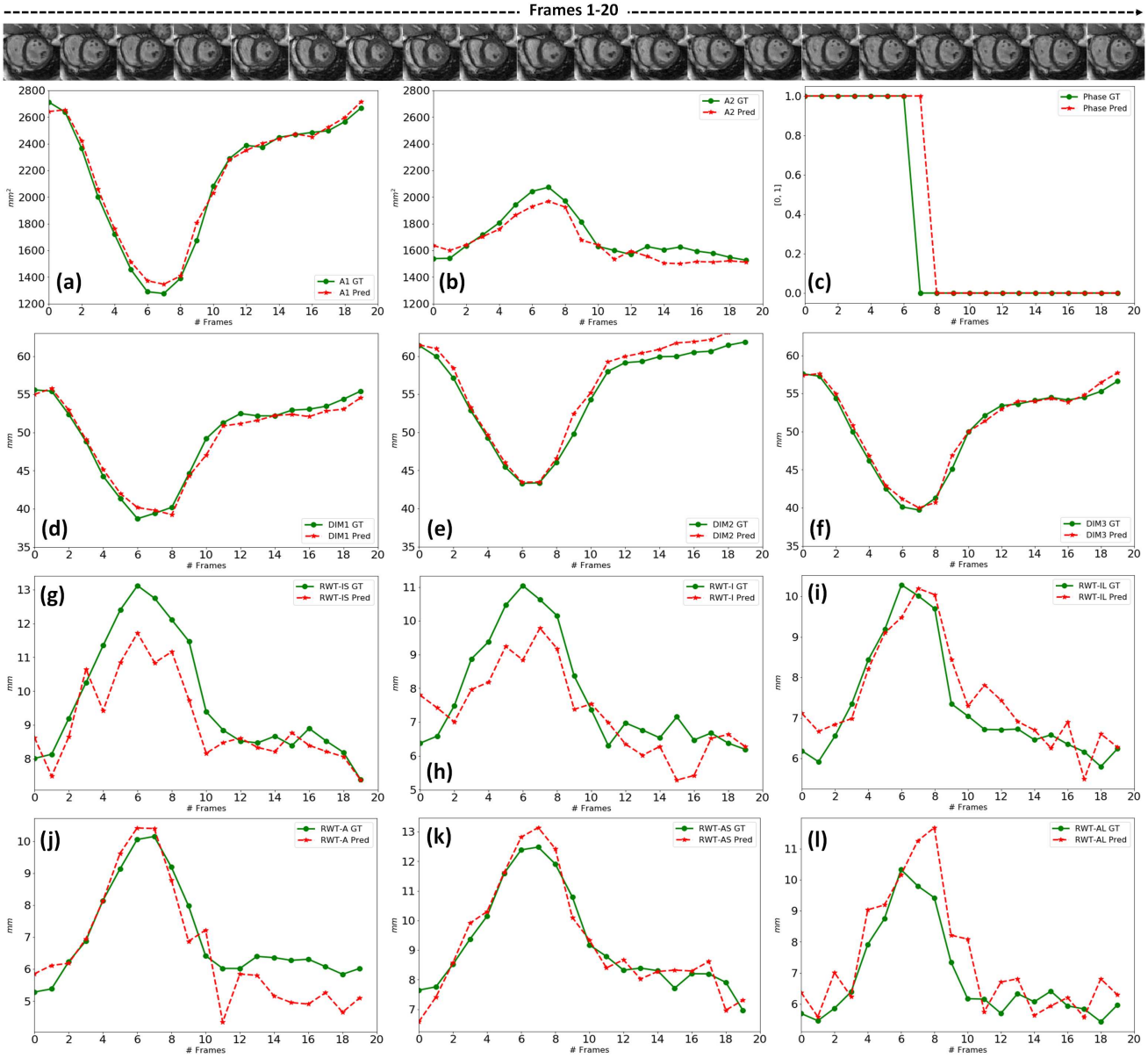


Fig. 7. Example of LV indices and cardiac phase estimation by our proposed framework for a median Cine-MR case across the whole cardiac cycle. The estimated results (dashed-line in red) are very close to their ground truth values (solid line in green color) for the three types of LV indices: areas (a-b), diameters (d-f), wall-thickness (g-l) and phase cycles (c). The proposed method captures the temporal variation pattern of all indices precisely. The corresponding MR images are shown in the top row for visual comparison.

their algorithm [24] to first learn cardiac representations with a deep CNN, and subsequently, the temporal dynamics of the cardiac sequence with two parallel RNN modules. Li *et al.* [23] proposed a method based on deep learning that includes 11 indices of regression and cardiac phase detection. The authors use deep layer aggregation (DLA) as the backbone to perform 11 index regressions simultaneously on 2D single images and derive the cardiac phase by searching for the maximum and minimum frames from the polynomial LV cavity region. In the most recent attempt for LV quantification, Wang *et al.* [13] proposed CSRNet as an end-to-end framework that computes the LV indices based on the segmentation mask similar to our model. However, most of these only use a single 2D image or embedding of 5 images for feature extraction.

Our proposed spatio-temporal multi-task learning approach outperformed the majority of these state-of-the-art methods for estimating LV indices, evaluated in terms of mean absolute error (MAE) and the Pearson correlation coefficient (PCC). These metrics were evaluated with respect to the ground truth values and are reported in TABLE II. We can see that our methods yield the lowest average MAE of 129 ± 115 mm^2 and PCC of 0.964 for the LV blood pool and Myo areas, compared to all other methods. Moreover, it achieved an average MAE of 1.76 ± 1.44 mm and 1.24 ± 1.01 mm for the LV dimensions and RWTs, which are very close to the state-of-the-art results reported by Wang *et al.* [13] on this benchmark dataset. However, in terms of PCC values, we outperformed the model proposed by Wang *et al.*, achieving values of 0.975

TABLE II
PERFORMANCE OF FULL LV QUANTIFICATION FOR EXISTING STATE-OF-THE-ART METHODS AND OUR PROPOSED METHODS. MAE AND PCC ARE SHOWN IN EACH CELL. OUR METHOD OUTPERFORMS ALL EXISTING METHODS FOR THE THREE TYPES OF LV INDICES AND CARDIAC PHASE IN TERMS OF AVERAGE MAE, PCC AND ER.

Indices	Metric	Indices-Net [19]	FullLVNet [26]	FullLVNet (intra/inter) [26]	DMTRL [24]	DLA [23]	CSRNet [13]	Ours multi-stage	Ours end-to-end
Area (mm^2)									
A-cav	MAE	185±162	205±182	181±155	172±148	135	107±98	106±87	**101±92**
	PCC	0.953	0.926	0.94	0.943	\	0.982	0.985	**0.986**
A-myo	MAE	223±193	204±195	199±174	189±159	177	162±127	165±132	**158±128**
	PCC	0.853	0.925	0.935	0.947	\	0.928	0.935	**0.940**
Average	MAE	204±133	205±145	190±128	180±118	156	134±115	135±29	**129±115**
	PCC	0.903	0.925	0.937	0.945	\	0.955	0.960	**0.964**
Dimension (mm)									
Dim1	MAE	\	2.87±2.23	2.62±2.09	2.47±1.95	2.04	**1.57±1.42**	1.76±1.43	1.85±1.49
	PCC	\	0.938	0.952	0.957	\	0.974	**0.975**	0.973
Dim2	MAE	\	2.96±2.35	2.64±2.12	2.59±2.07	2.02	**1.48±1.36**	1.80±1.49	1.76±1.44
	PCC	\	0.864	0.881	0.894	\	**0.979**	0.977	0.977
Dim3	MAE	\	2.92±2.48	2.77±2.22	2.48±2.34	2.05	**1.56±1.33**	1.72±1.41	1.82±1.43
	PCC	\	0.924	0.935	0.943	\	0.979	**0.978**	0.975
Average	MAE	\	2.92±1.89	2.68±1.64	2.51±1.58	2.03	**1.54±1.37**	1.76±1.44	1.81±1.46
	PCC	\	0.901	0.917	0.925	\	0.978	0.977	0.975
RWT (mm)									
IS	MAE	1.39±1.13	1.42±1.21	1.32±1.09	1.26±1.04	1.39	**1.06±0.87**	1.15±0.93	1.16±0.921
	PCC	0.824	0.806	0.84	0.856	\	0.895	0.908	**0.910**
I	MAE	1.51±1.21	1.53±1.25	1.38±1.10	1.40±1.10	1.41	1.33±1.14	**1.24±1.01**	1.25±1.01
	PCC	0.701	0.678	0.751	0.747	\	0.812	**0.856**	0.855
IL	MAE	1.65±1.36	1.74±1.43	1.57±1.35	1.59±1.29	1.48	**1.33±1.09**	1.42±1.13	1.47±1.17
	PCC	0.671	0.618	0.691	0.693	\	0.788	**0.836**	0.825
AL	MAE	1.53±1.25	1.59±1.31	1.60±1.36	1.57±1.34	1.46	**1.32±1.09**	1.37±1.08	1.38±1.06
	PCC	0.698	0.657	0.651	0.659	\	0.77	**0.829**	**0.831**
A	MAE	1.30±1.12	1.36±1.17	1.34±1.11	1.32±1.10	1.24	**1.08±0.92**	1.13±0.97	1.14±0.99
	PCC	0.781	0.754	0.768	0.777	\	0.84	**0.875**	0.870
AS	MAE	1.28±1.00	1.43±1.24	1.26±1.10	1.25±1.01	1.31	**0.97±0.80**	1.05±0.84	1.03±0.83
	PCC	0.871	0.821	0.864	0.877	\	0.919	0.928	**0.933**
Average	MAE	1.44±0.71	1.51±0.81	1.41±0.72	1.39±0.68	1.38	**1.16±0.097**	1.23±1.01	1.24±1.01
	PCC	0.758	0.723	0.761	0.768	\	0.868	**0.872**	0.871
Phase (%)									
ES/DS	ER	\	13	10.4	8.2	**8.1**	\	10.8	9.0

and 0.871. On the other hand, there is a reduction of 38.8% MAE value for the cavity area and 38.6% for the Myo area when compared to the IndiceNet, FullLVNet, DMTRL, and DRL methods. The higher average PCC value means that there is a better linear relationship between the estimation results from our model and the ground truth, which is illustrated for a test case in Fig. 7. It can be seen in the figure that RWTs are more difficult to estimate in comparison to LV dimensions and area of the cavity. This is because RWTs estimation involves both the endocardium and epicardium contours (small region), which is usually difficult to segment due to available noise and low contrast within the image. Furthermore, the improvements afforded by our approach to estimates the myocardial indices and RWTs, in terms of MAE are less prevalent as the segmentation results for the LV blood pool were more accurate than for the Myo.

Our model is trained both in an end-to-end and multi-stage fashion. The end-to-end model unified all three modules (segmentation, regression, and classification), and based on evaluation metrics achieved overall better performance compared to the former.

Ablation Studies: We conducted ablation experiments to evaluate the effectiveness of spatio-temporal convolutions in our proposed spatio-temporal multi-task framework. The results are presented in Table III. Our baseline network uses only 3D convolution layers to incorporate temporal information for better LV indices regression and phase classification in a multi-stage manner (row 1). It can be seen from the table, that this configuration achieved only an average MAE value of 2.06 mm for LV blood pool dimensions, 154 mm^2 for LV and Myo areas, and 1.35mm for RWTs. The error rate for phase classification is also quite high close to 11.0%. Training this network with the same configuration in an end-to-end manner improved the LV areas, dimensions and RWT quantification slightly (row 2). Next, we replaced the 3D convolution blocks with proposed $2D + t$ spatio-temporal layers with a kernel size of $3 \times 1 \times 1$ and $1 \times 3 \times 3$. This model trained both in a multi-stage and end-to-end fashion (rows 5-8). We can see that the MAE and ER values improved almost for all the indices in comparison to 3D convolution configuration. The model with multi-stage training and spatio-temporal CNN achieved the lowest MAE value of 1.76 mm and 1.23 mm for LV dimensions and RWTs (row 7). However, the 3D spatio-temporal CNN with an end-to-end training strategy achieved the lowest MAE value for LV and Myo areas and reduced the phase classification error-rate to 9.0% (row 8). Moreover, we have also trained two models based on our DR-UNet 2D segmentation model in multi-stage and end-to-end fashion (rows 3 & 5). Here, we can see again that DR-UNet 2D and 3D spatio-temporal model outperformed DR-UNet 2D and 3D CNN. These results can confirm the effectiveness of spatio-temporal layers, which encode jointly global temporal

information and local spatial information.

C. Statistical Analysis

To statistically measure the effectiveness of our proposed multi-task approach compare to manual ground-truths, we employ Bland-Altman analysis [44]. This statistical technique determines the agreement between two quantitative measurements by constructing limits of agreement (LoA).

The Bland-Altman plots for differences in LV indices (Areas, DIMs, and RWTs) obtained using manual and proposed methods are shown in Fig. 8. The areas per patient are expressed in mm^2 and for dimensions and RWTs in mm respectively. It can be observed that in terms of LV area estimation (Fig. 8(a)) the agreement between our proposed method and manually generated ground truth is high with a bias (mean signed difference) of 14.98 mm^2 and limits of agreement of $\pm$342.45 mm^2. This is also the same for the dimensions and RWTs with LoR of $\pm$4.51 mm^2 and $\pm$2.99 mm^2, respectively. These results suggest that the proposed method has a small bias to overestimate RWTs and that the variation between automated and manual estimates of the LV area is only slightly greater than the expert manual annotation. There are some outlier cases (refer to plots a-c in Fig. 8), regarded as hard-examples to measure due to the presence of low contrast and noise in the scans or not a precise segmentation of LV and Myo cavity areas. Overall, our methods produced accurate indices quantification resulting in a significantly lower mean difference in most of the cases.

V. Discussion

We proposed a novel multi-task end-to-end method for full LV quantification in cardiac cine-MR images in this study. This approach leverages spatial and temporal information contained within the cine-MR sequences using 3D spatio-temporal convolutions, to quantify the LV, unlike most existing methods that utilize just 2D spatial convolutions. Additionally, our model exploits the information contained in the estimated segmentation masks rather than the raw images, to combine both spatial and temporal features during model training, and inference. Thus, features are learned from the full cine-MR sequence to estimate the various LV indices of interest. This framework can be considered as an efficient tool for cardiac LV functional analysis that tackles three different related tasks simultaneously, namely - LV blood pool and Myo segmentation, regression of 11 LV morphological indices, and classification of cardiac phase.

A multi-task learning framework enables specialized modules to tackle different tasks simultaneously while benefiting from one another. By sharing the same feature extraction backbone, this framework allows information synergy between various tasks and presents a mutual influence process that can further obtain performance gains from different tasks. In our network, the regression and classification tasks are optimized in an end-to-end paradigm, together with LV segmentation. The presented results indicate that methodically consolidating multiple but interrelated tasks with mutual information sharing and considering the task relationship using a suitable weighting strategy, yields better performance. Moreover, our method is similar to the commensal correlation network proposed by Luo *et al.* [45], where feature extraction performed in parallel to the segmentation for LV quantification. However, this method computes the LV indices using a single 2D MR image and temporal information of the cardiac cycle discarded. In contrast, our proposed method takes advantage of the full cardiac cycle and the quantification indices computed in a cascade manner.

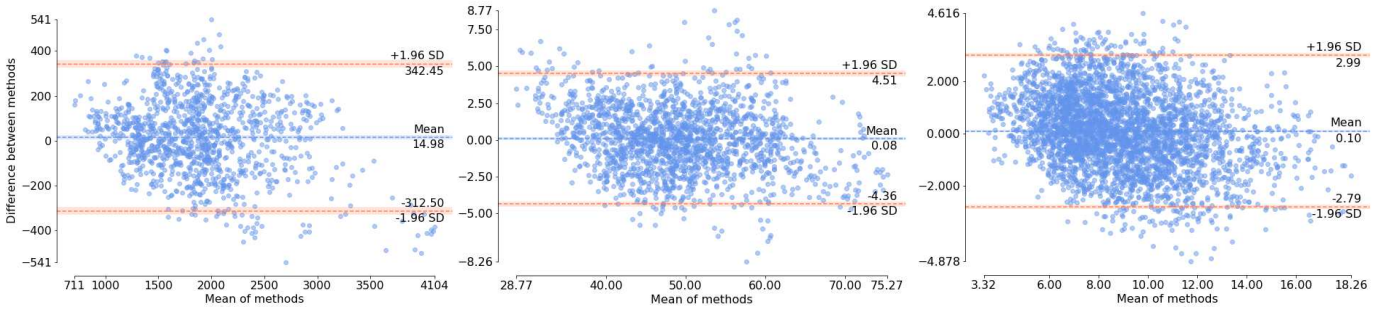
Extensive experiments on the LVQuan 2018 benchmark dataset have highlighted the effectiveness of our approach. By integrating myocardial segmentation with multi-task classification and regression learning framework. The method combines the advantages of two-step methods based on segmentation with end-to-end learning approaches. The segmentation module of the framework can remove task-independent structures so that the following regression and classification network can extract discriminating features from the segmented masks only. The myocardial contours only guide the regression task, but do not fully determine the accuracy of the quantification results like the two-step procedures based on segmentation. The results for DR-UNet with post-processing and weighted-loss outperformed the other segmentation methods. Additionally, our segmentation network has fewer than 3.6 million trainable parameters and takes less than 20 minutes to train with 300 epochs, which is considered modest in size/complexity, compared with other relevant SOTA approaches [24]. In the validation phase, LV metrics are estimated in each fold for 28 subjects with 580 images in just $\sim$2.2 seconds on a machine with 4GB GPU memory. This demonstrates the real-time characteristic of the pipeline, which could be integrated into MR acquisition systems to triage patients into high and low-risk categories of CVDs for improved efficiency and clinical decision making.

The proposed method achieved comparable results to CSR-Net and even for LV and Myo areas, our model achieved better results. On the other hand, for 3 LV diameters and 6 RWT, our model could achieve very close results. However, It should be noted that CSRNet doesn't consider cardiac phase detection tasks (systolic or diastolic) and only quantify the indices based on raw segmentation. On the other hand, our method simultaneously performs three tasks including segmentation, indices regression and phase classification. In this end-to-end training, every learning step is directed at the final goal, encoded by the overall objective function. There is no need for training modules on an auxiliary objective. Our end-to-end multi-task learning framework is nicely consistent with the general approach of machine learning to take the human expert out of the loop and to solve problems in a purely data-driven manner.

Although the segmentation accuracy of our approach was high for the LV blood pool, there is still room for improvement, especially for the myocardium. Moreover, in the end-to-end training strategy, combining the losses of different tasks is a critical issue, because different tasks converge at different rates. To balance task importance during optimization, we empirically set the hyperparameter λ for each task in the

TABLE III
EFFECTIVENESS OF 3D SPATIO-TEMPORAL CNN LAYERS IN OUR PROPOSED FRAMEWORK IN COMPARISON TO 2D AND 3D CONVOLUTION OPERATIONS. MULTI-STAGE AND END-TO-END DENOTES THE TYPE OF TRAINING STRATEGY. THE VALUES SHOW THE AVERAGE MAE FOR LV BLOOD POOL AND MYO AREAS, 3 LV DIMENSIONS AND 6 RWTS.

Methods	**Multi-stage**	**End-to-End**	**Area** (mm^2)	**DIM** (mm)	**RWT** (mm)	**Phase** (%)
DR-UNet 3D & 3D CNN	✓	×	154±134	2.06±1.60	1.35±1.09	11.0
DR-UNet 3D & 3D CNN	×	✓	148±122	2.00±1.32	1.32±1.02	11.2
DR-UNet 2D & 3D CNN	✓	×	157±127	2.01±1.64	1.30±1.03	11.3
DR-UNet 2D & 3D CNN	×	✓	146±116	2.16±1.73	1.32±1.01	10.6
DR-UNet 2D & 3D spatio-temporal	✓	×	136±115	1.77±1.44	1.23±1.01	10.7
DR-UNet 2D & 3D spatio-temporal	×	✓	142±113	2.13±1.71	1.25±1.05	10.6
DR-UNet 3D & 3D spatio-temporal	✓	×	135±29	**1.76±1.44**	**1.23±1.01**	10.8
DR-UNet 3D & 3D spatio-temporal	×	✓	**129±115**	1.81±1.46	1.24±1.01	**9.0**



(a) (b) (c)

Fig. 8. Bland-Altman analysis plots show that the LV indices for areas, dimensions, and RWTs estimated using our model is very close to the ground truth. Plot (a) illustrates the LV cavity and Myo areas (A1 and A2). The plots for LV blood pool dimensions and RWTs are shown in b-c. The Bland-Altman plots are computed using a confidence interval of 95%. The blue line indicates the mean and the dashed red lines indicate the level of agreement.

loss function after a greedy search. We believe incorporating a more methodical multi-task loss weighting strategy could improve the performance of our pipeline even further. Furthermore, the LVQuan benchmark dataset is preprocessed *a priori*, and LV ROIs are extracted from Cine-MR sequences, which is not common in a real clinical scenario. We aim to extend our method to include LV detection within the pipeline and design a fully automated computer-aided diagnosis system that eliminates the need for any pre-processing step. In this study, all quantification indices are 2D, while the most important cardiac functional/morphological indices are 3D metrics like ejection fraction and LV volumes. As future work, we also aim to extend our method to include these metrics as well.

VI. CONCLUSION

In this study, we proposed a robust, efficient, and light-weight network architecture for fully automatic LV quantification in Cine-MR images. The proposed network first segments the LV and Myo blood pool. Subsequently, the segmented structures fed to a spatio-temporal multi-task regression and classification component to estimate cardiac LV indices. It includes the LV cavity and Myo areas, LV dimensions, myocardial RTWs, and the cardiac cycle phase (diastolic or systolic). Although the LV anatomical shape and appearance are highly variable across different subjects and the training subjects acquired in various hospitals, the proposed method successfully learned robust representations from the MR sequences and estimated the LV indices of interest with high accuracy. We evaluated our method on 145 subjects, and the experimental results highlighted the advantage afforded by our approach in comparison to the state-of-the-art methods. The proposed method can be a promising contribution having clinical importance both during diagnosis, where the cardiac MR volume needs to be analyzed and during treatment planning when the quantification of the anatomical structure of LV needs to be accurate and fast.

REFERENCES

[1] E. S. D. Group, A. Timmis, E. Wilkins, L. Wright, N. Townsend *et al.*, "European Society of Cardiology: Cardiovascular Disease Statistics 2017," *European Heart Journal*, vol. 39, no. 7, pp. 508–579, 11 2017.

[2] E. J. Benjamin, M. J. Blaha, S. E. Chiuve, M. Cushman, S. R. Das *et al.*, "Heart disease and stroke statistics—2017 update: A report from the american heart association," *Circulation*, vol. 135, no. 10, pp. e146–e603, 2017.

[3] J. Stewart, G. Manmathan, and P. Wilkinson, "Primary prevention of cardiovascular disease: A review of contemporary guidance and literature," *JRSM Cardiovascular Disease*, vol. 6, p. 2048004016687211, 2017.

[4] M. R. Avendi, A. Kheradvar, and H. Jafarkhani, "Automatic segmentation of the right ventricle from cardiac mri using a learning-based approach," *Magnetic Resonance in Medicine*, vol. 78, no. 6, pp. 2439–2448, 2017.

[5] H. W. Kim, A. Farzaneh-Far, and R. J. Kim, "Cardiovascular magnetic resonance in patients with myocardial infarction: Current and emerging applications," *Journal of the American College of Cardiology*, vol. 55, no. 1, pp. 1 – 16, 2009.

[6] M. Cantinotti and M. Koestenberger, "Quantification of left ventricular size and function by 2-dimensional echocardiography: So basic and so difficult," *Circulation: Cardiovascular Imaging*, vol. 10, no. 11, p. e007165, 2017.

[7] W. Bai, M. Sinclair, G. Tarroni, O. Oktay, M. Rajchl *et al.*, "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks," *Journal of Cardiovascular Magnetic Resonance*, vol. 20, no. 1, p. 65, Sep 2018.

[8] T. Kurzendorfer, C. Forman, M. Schmidt, C. Tillmanns, A. Maier *et al.*, "Fully automatic segmentation of left ventricular anatomy in 3-DLGE-MRI," *Computerized Medical Imaging and Graphics*, vol. 39, no. 59, pp. 13–27, 2017.

[9] M. Afshin, I. B. Ayed, A. Islam, A. Goela, T. M. Peters *et al.*, "Global assessment of cardiac function using image statistics in mri," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 535–543.

[10] R. M. Lang, L. P. Badano, V. Mor-Avi, J. Afilalo, A. Armstrong *et al.*, "Recommendations for Cardiac Chamber Quantification by Echocardiography in Adults: An Update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging," *European Heart Journal - Cardiovascular Imaging*, vol. 16, no. 3, pp. 233–271, 02 2015.

[11] M. Afshin, I. B. Ayed, K. Punithakumar, M. Law, A. Islam *et al.*, "Regional assessment of cardiac left ventricular myocardial function via mri statistical features," *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 481–494, Feb 2014.

[12] Q. Tao, W. Yan, Y. Wang, E. H. M. Paiman, D. P. Shamonin *et al.*, "Deep learning–based method for fully automatic quantification of left ventricle function from cine mr images: A multivendor, multicenter study," *Radiology*, vol. 290, no. 1, pp. 81–88, 2019, pMID: 30299231.

[13] W. Wang, Y. Wang, Y. Wu, T. Lin, S. Li *et al.*, "Quantification of full left ventricular metrics via deep regression learning with contour-guidance," *IEEE Access*, vol. 7, pp. 47 918–47 928, 2019.

[14] A. Suinesiaputra, D. A. Bluemke, B. R. Cowan, M. G. Friedrich, C. M. Kramer *et al.*, "Quantification of lv function and mass by cardiovascular magnetic resonance: multi-center variability and consensus contours," *Journal of Cardiovascular Magnetic Resonance*, vol. 17, no. 1, p. 63, 2015.

[15] B. Ruijsink, E. Puyol-Antón, I. Oksuz, M. Sinclair, W. Bai *et al.*, "Fully automated, quality-controlled cardiac analysis from cmr: Validation and large-scale application to characterize cardiac function," *JACC: Cardiovascular Imaging*, 2019.

[16] R. Attar, M. Pereañez, A. Gooya, X. Albà, L. Zhang *et al.*, "Quantitative cmr population imaging on 20,000 subjects of the uk biobank imaging study: Lv/rv quantification pipeline and its evaluation," *Medical Image Analysis*, vol. 56, pp. 26 – 42, 2019.

[17] I. B. Ayed, H. mei Chen, K. Punithakumar, I. Ross, and S. Li, "Max-flow segmentation of the left ventricle by recovering subject-specific distributions via a bound of the bhattacharyya measure," *Medical Image Analysis*, vol. 16, no. 1, pp. 87 – 100, 2012.

[18] X. Zhen, Z. Wang, A. Islam, M. Bhaduri, I. Chan *et al.*, "Multi-scale deep networks and regression forests for direct bi-ventricular volume estimation," *Medical Image Analysis*, vol. 30, pp. 120 – 129, 2016.

[19] W. Xue, A. Islam, M. Bhaduri, and S. Li, "Direct multitype cardiac indices estimation via joint representation and regression learning," *IEEE Transactions on Medical Imaging*, vol. 36, no. 10, pp. 2057–2067, Oct 2017.

[20] W. Xue, I. B. Nachum, S. Pandey, J. Warrington, S. Leung *et al.*, "Direct estimation of regional wall thicknesses via residual recurrent neural network," in *Information Processing in Medical Imaging*, 2017, pp. 505–516.

[21] Chenyang Xu and J. L. Prince, "Snakes, shapes, and gradient vector flow," *IEEE Transactions on Image Processing*, vol. 7, no. 3, pp. 359–369, March 1998.

[22] A. Debus and E. Ferrante, "Left ventricle quantification through spatio-temporal cnns," in *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges*, 2019, pp. 466–475.

[23] J. Li and Z. Hu, "Left ventricle full quantification using deep layer aggregation based multitask relationship learning," in *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges*, 2019, pp. 381–388.

[24] W. Xue, G. Brahm, S. Pandey, S. Leung, and S. Li, "Full left ventricle quantification via deep multitask relationships learning," *Medical Image Analysis*, vol. 43, pp. 54 – 65, 2018.

[25] Z. Wang, M. B. Salah, B. Gu, A. Islam, A. Goela *et al.*, "Direct estimation of cardiac biventricular volumes with an adapted bayesian formulation," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 4, pp. 1251–1260, April 2014.

[26] W. Xue, A. Lum, A. Mercado, M. Landis, J. Warrington *et al.*, "Full quantification of left ventricle via deep multitask learning network respecting intra- and inter-task relatedness," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, 2017, pp. 276–284.

[27] Q. Meng, N. Pawlowski, D. Rueckert, and B. Kainz, "Representation disentanglement for multi-task learning with application to fetal ultrasound," in *Smart Ultrasound Imaging and Perinatal, Preterm and Paediatric Image Analysis*. Cham: Springer International Publishing, 2019, pp. 47–55.

[28] A. Chartsias, T. Joyce, G. Papanastasiou, S. Semple, M. Williams *et al.*, "Disentangled representation learning in cardiac image analysis," *Medical Image Analysis*, vol. 58, p. 101535, 2019.

[29] J. Zhang, Y. Xie, Q. Wu, and Y. Xia, "Medical image classification using synergic deep learning," *Medical Image Analysis*, vol. 54, pp. 10 – 19, 2019.

[30] M. Osadchy, M. L. Miller, and Y. L. Cun, "Synergistic face detection and pose estimation with energy-based models," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. MIT Press, 2005, pp. 1017–1024.

[31] A. K. Maier, C. Syben, B. Stimpel, T. Würfl, M. Hoffmann *et al.*, "Learning with known operators reduces maximum error bounds," *Nature Machine Intelligence*, vol. 1, no. 8, pp. 373–380, 2019.

[32] M. D. Cerqueira, N. J. Weissman, V. Dilsizian, A. K. Jacobs, S. Kaul *et al.*, "Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart," *Circulation*, vol. 105, no. 4, pp. 539–542, 2002.

[33] S. Vesal, N. Ravikumar, and A. Maier, "Dilated convolutions in neural networks for left atrial segmentation in 3d gadolinium enhanced-mri," in *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges*, 2019, pp. 319–328.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[35] Z. Wang and S. Ji, "Smoothed dilated convolutions for improved dense prediction," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 2486–2495.

[36] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun *et al.*, "A closer look at spatiotemporal convolutions for action recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 6450–6459.

[37] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *Computer Vision – ECCV 2010*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 140–153.

[38] S. S. Yoon, E. Hoppe, M. Schmidt, C. Forman, P. Sharma *et al.*, "Automatic Cardiac Resting Phase Detection for Static Cardiac Imaging Using Deep Neural Networks," in *Proceedings of the Joint Annual Meeting ISMRM-ESMRMB (27th Annual Meeting & Exhibition)*, I. S. for Magnetic Resonance in Medicine, Ed., 2019.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 1026–1034.

[40] K. Zuiderveld, "Contrast limited adaptive histogram equalization," *Graphics Gems IV*, pp. 474–485, 1994.

[41] F. Milletari, N. Navab, and S. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, Oct 2016, pp. 565–571.

[42] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 2016, pp. 265–283.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

[44] J. M. Bland and D. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *The Lancet*, vol. 327, no. 8476, pp. 307 – 310, 1986.

[45] G. Luo, S. Dong, W. Wang, K. Wang, S. Cao *et al.*, "Commensal correlation network between segmentation and direct area estimation for bi-ventricle quantification," *Medical Image Analysis*, vol. 59, p. 101591, 2020.