Long-Range Decoder Skip Connections: Exploiting Multi-Context Information for Cardiac Image Segmentation

Nicolás Gutierrez-Castilla[1], Ricardo da S. Torres[2], Alexandre X. Falcão[2], Sebastian Kozerke[3], Jürg Schwitter[4], Pier-Giorgio Masci[5], and Javier A. Montoya-Zegarra[3,1]
[1]Department of Computer Science, Universidad Católica San Pablo, Arequipa, Perú
[2]Institute of Computing, University of Campinas, Campinas, SP, Brazil
[3]Institute for Biomedical Engineering, ETH Zurich, Zurich, Switzerland
[4]Center for Cardiac Magnetic Resonance, Lausanne University Hospital, Lausanne, Switzerland
[5]Rayne Institute School of Bioengineering and Imaging Sciences, King's College London, London, United Kingdom

***Abstract*—The heart is one of the most important organs in our body and many critical diseases are associated with its malfunctioning. To assess the risk for heart diseases, Magnetic Resonance Imaging (MRI) has become the golden standard imaging technique, as it provides to the clinicians stacks of images for analyzing the heart structures, such as the ventricles, and thus to make a diagnosis of the patient's health. The problem is that examination of these stacks, often based on the delineation of heart structures, is tedious and error prone due to inter- and intra-variability among manual delineations. For this reason, the investigation of fully automated methods to support heart segmentation is paramount. Most of the successful methods proposed to solve this problem are based on deep-learning solutions. Especially, encoder-decoder architectures, such as the U-Net [1], have demonstrated to be very effective architectures for medical image segmentation. In this paper, we propose to use long-range skip connections on the decoder-part to incorporate multi-context information onto the predicted segmentation masks and also to improve the generalization of the models. In addition, our method obtains smoother segmentations through the combination of feature maps from different stages onto the final prediction layer. We evaluate our approach in the ACDC [2] and LVSC [3] heart segmentation challenges. Experiments performed on both datasets demonstrate that our approach leads to an improvement on both the total Dice score and the Ejection Fraction Correlation, when combined with state-of-the-art encoder-decoder architectures.**

I. Introduction

Cardiovacular diseases are one of the leading cause of death in the world [4]. A way to prevent those diseases to expand until a critical point relies on early examination. An imaging technique that is considered the golden standard to visualize and record the heart is Magnetic Resonance Imaging (MRI), which gives clinicians a 3D temporal data of the heart in a cardiac cycle. From these 3D temporal data, clinicians examine two special phases of the heartbeat cycle – the End-Diastole (ED) and the End-Systole (ES) – and then segment the most important structures of the heart: the Left Ventricle (LV), Right Ventricle (RV), and the Left Ventricle Myocardium (MYO). For the segmentation of these structures, clinicians often use semi-automatic methods. However, this task is still time consuming, and is prone to intra- and inter-observer variability. The creation of fully automatic heart segmentation approaches is, therefore, of paramount importance.

Deep learning has been shown to lead state-of-the-art results on several highly complex computer vision problems [5]. The same is true for the semantic segmentation problem [6], where the most common architecture is a fully convolutional network composed of an encoder and a decoder [7]. The encoder usually consists of a sequence of convolution, non-linear activations, and pooling operations that obtain a hierarchical representation of the input data and that also reduce the dimensions of the input images. The decoder learns more complex feature maps that are upsampled in order to obtain predictions with the same or similar dimensions as the input data. For medical image segmentation, one of the most popular architectures is the U-Net [1], which serves as the basis for several successful solutions [8] [9]. The main characteristic of the U-Net architecture is that it uses skip connections to aggregate information from the encoder to the decoder using concatenation. Also, this approach uses multiple transposed convolutions to learn how to proper upsample the data, often leading to improved segmentation results.

In fact, in the context of heart structure segmentation, U-Net-based approaches have been demonstrated to yield very effective results [10] [11]. For example, in the recent ACDC heart segmentation challenge [2], most of the successful competing approaches relied on the use of U-Net-like architectures [12] [13].

In this paper, we introduce the use of long skip connections on the decoder part of encoder-decoder architectures, such as the U-Net, as a way to aggregate multi-context information from different levels onto the final predictions and to refine them. Additionally, our module acts as a regularizer, adds few extra parameters onto the final model, and helps the model to converge faster. Those long skip connections are encapsulated in a module, referred to as *Dense-Decoder skip connections module* (or simply *Dense-Decoder module*), as it looks similar to a Dense block [14].

It is worth to mention that skip connections are also a good way to improve deep learning architectures for many reasons: (i) they simplify the loss dimensional space making it easier to find a good minimum which can generalize better [15]; (ii) they eliminate singularities on the training of the network [16]; (iii) they improve the propagation of the gradients [17]; and (iv) they reuse previous learned features [14].

We evaluate our "Dense-Decoder skip connections" on two datasets for cardiac segmentation – the ACDC [2] and LVSC [3] datasets – using different configurations of U-Net presented in the ACDC competition. Experimental results demonstrate that the incorporation of the Dense-Decoder module improves the segmentation of U-Net based approaches.

II. Related Work

Cardiac image segmentation has been addressed for a long time. The first methods relied on image-processing techniques [18], pixel classification methods [19], deformable models [20], and graph-based approaches [21]. Other initiatives were based on strong geometrical priors of the cardiac structures including shape-based deformable models [22], active shape and appearance models [23], and atlas-based methods [24]. Most of these last initiatives required a training dataset with manual annotations.

More recently, deep learning approaches have shown effective results on semantic segmentation [25]. Especially for medical image segmentation, the U-Net is the most popular approach [1]. For instance, in the recent ACDC cardiac segmentation challenge [2], eight out of the ten participants relied on U-Nets or modified versions.

Some of these successful initiatives include the methods of Baumgartner et al. [12] and Isensee et al. [13]. Baumgartner et al. [12] evaluated different 2D and 3D encoder-decoder architectures for heart segmentation. Their motivation for using a 3D architecture relied on the fact that the segmentation of slices near to the apex and to the base of the heart requires spatial information. They also explored more compact models by adapting the upsampling path of the U-Net which lead to an improvement on their segmentation results. Isensee et al. [13], in turn, implemented an ensemble of 2D and 3D U-Net architectures. When combining the predictions of the 2D and 3D architectures, they achieved a better performance on the Right Ventricle. In both architectures, they created low resolution segmentations from early upsampled feature maps and added this information to the final prediction. They referred to this strategy as Deep supervision.

Another work that tested the suitability of 3D networks over 2D networks was the method of Patravali et al. [26]. In their experiments, the best performance was obtained using a 2D U-Net combined with a Dice loss. Finally, the last method based on a encoder-decoder architecture was the one of Yang et al. [27]. In their method, they used a 3D U-Net. The weights of the 3D U-Net's encoder were initialized with the weights of a network trained for a video classification task.

In summary, the research initiatives that assessed the use of 3D networks concluded that such architectures led to less effective segmentation results when compared with 2D networks. This is probably caused by the low resolution of the data in the z-dimension (having between $8-10$ slices), which makes it difficult to use a 3D network efficiently without losing information on z-dimension (see Figure 1) and also by the fact that less training samples are then available.

Contributions: Previous works for cardiac image segmentation have explored the use of 2D or 3D encoder-decoder architectures such as the U-Net. One main difference with previous approaches is that our method obtains smoother and refined segmentations by combining feature maps from different stages of the decoder directly onto the final prediction layer. In addition, the combination of feature maps from different stages adds implicitly multi-context/scale information onto the final segmentation. Also, since the feature maps are added in the form of long-range skip connections from the decoder layers onto the final prediction layer, the training process is also benefited as the gradients can flow directly from the final outputs to the decoder layers during back-propagation. Finally, since the new module relies on skip-connections, the size of the model remains constant as practically no extra parameters are added.

III. Dense-Decoder Skip Connection

In this section, we first introduce the U-Net architecture. Later on, we present our proposed Dense-Decoder skip connection module. Finally, we present the final network architecture, which takes advantage of this module to achieve state-of-the-art results.

A. U-Net architecture

The U-Net model consists of an *encoder-decoder* architecture with skip connections between them. In the *encoder* part, the convolutional layers use 3×3 filters and follow two design rules at each stage: (i) the number of output channels is equal to the number of input filters and (ii) if the size of the feature map is halved (e.g., after max pooling), then the number of filters at that given layer is doubled. In the *decoder* part, the convolutional layers also use 3×3 filters and at each stage: (i) the number of output channels of a given feature map is equal to the number of input filters and (ii) if the feature map size is upsampled (e.g., by using 2×2 transposed convolutions), the number of filters is halved. The skip connections are added symmetrically by concatenating features maps from the encoder to the decoder path. Figure 2b illustrates the base U-Net model. Because of the skip connections, a hierarchical feature representation is obtained by combining the low level features of the shallow layers with the high level features of the dense layers. Also as mentioned in [28], the skip connections play a fundamental role in this model, as they affect the final prediction by creating a direct flow for feature maps from early layers (encoder stage) in the network to later ones (decoder stage).

B. Dense-Decoder Skip Connections

As stated before, skip connections have many advantages such as: (i) simplifying the loss dimensional space by mak-

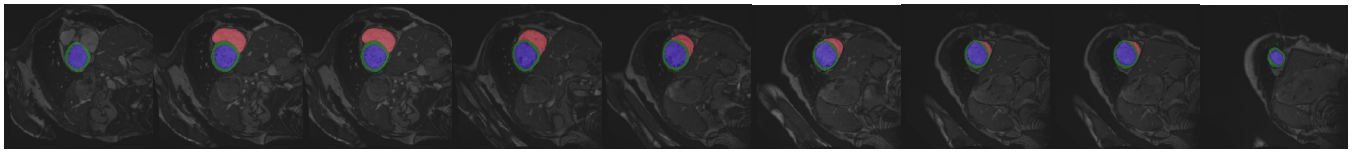

Fig. 1. An example of a stack of MRI slices from a patient for the ED phase, containing 9 slices from base to apex. We can see that the shape of the right ventricle changes more than the other two structures (endocardium, myocardium) and that the area of the myocardium on average is the smallest one.

ing it easier to find a good minimum [15], (ii) eliminating singularities (gradient ambiguity) during the training of the network [16], (iii) improving the propagation of gradients [17], and (iv) reusing learned features [14]. Inspired by those advantages, we propose, in this paper, to use long-range skip connections on the decoder to both refine the final segmentation result and also to incorporate multi-scale/context information into the final prediction.

The proposed module reuses the learned features from the decoder part and jointly aggregates them directly onto the final prediction layer (see Figure 2c). To do so, we first select which feature maps from the decoder should be used. Next, the selected feature maps are upsampled to the same size as the final output map. The upsampling process can be done in two ways: (i) using deconvolutions with n_{class} filters or (ii) using convolutions with n_{class} filters followed by an upsampling operation to match the final prediction size. Note that the n_{class} variable denotes the number of classes to be predicted. Finally, the upsampled feature maps are combined together with the last convolutional layer of the decoder either by using concatenation or addition operations.

By combining feature maps from different feature levels, multi-scale context information is added onto the final prediction and smoother predictions are obtained. In addition, the proposed module also helps the model to converge faster because the gradients are directly propagated across the different levels of the network onto their corresponding encoder part. Finally, our multi-context design also makes the decoder part of the network more adaptable to a specific problem, e.g., if a network does not need a complex decoder, the network could learn to use the simplest decoder path to the final prediction; but if a more complex decoder is needed then the network could learn to use all of the skip connections.

C. Adapted Network Architecture

By using our proposed module, we can easily extend the functionality of existing state-of-the-art architectures to incorporate multi-context information. The proposed Dense-Decoder U-Net is illustrated in Fig 2. Given an input MRI slice (Fig. 2a), we use an Encoder-Decoder architecture such as the U-Net to obtain feature maps at different scales (Fig. 2b). Next, we plug our Dense-Decoder module (Fig. 2c) to incorporate the multi-context information. Finally, we combine the upsampled feature maps to obtain the final predictions (Fig. 2d).

To refer to which of the features of the U-Net's decoder are used in the Dense-Decoder Module, we use the notation in Table I.

Name	UNet's features from Figure 2
Dense-Decoder Module1add	features(red)
Dense-Decoder Module2add	features(red, green)
Dense-Decoder Module3add	features(red, green, yellow)
Dense-Decoder Module4add	features(red, green, yellow, orange)
Dense-Decoder Module1concat	features(red)
Dense-Decoder Module3concat	features(red, green, yellow)

TABLE I
SETTINGS FOR THE DENSE-DECODER MODULE.

IV. EVALUATION PROTOCOL

A. Datasets

To validate our experiments, we used two datasets: the ACDC dataset [2] and the Sunnybrook dataset [3].

*a) **ACDC Dataset**:* The ACDC Dataset was first made publicly available in the Automatic Cardiac Diagnosis Challenge ACDC (2017). This dataset comprises short-axis cine-MRIs of 150 patients acquired at the University Hospital of Dijon. Each cine-MRI was manually annotated by two medical experts. The patients are classified into five evenly distributed subgroups (4 pathological plus 1 healthy subject groups). The considered categories are: Normal (NOR), Dilated Cardiomyopathy (DCM), Hypertrophic Cardiomyopathy (HCM), Myocardial Infarction (MINF), and Right Ventricular Abnormality (RVA). The cine MRIs were acquired using two MR scanners of different magnetic strengths (1.5T and 3.0T) with resolutions ranging from (0.70×0.70mm – 1.92×1.92 mm) in-plane and (5mm-10mm) through-plane. Each time series is composed of 28 to 40 3D volumes, which partially or completely cover the cardiac cycle. Each 3D volume covers the LV from base to the apex. Examples of some images with their corresponding ground-truth labels are given in Figure 1. For the partition of the data into training and validation set, we used the same protocol as in [12] and [13].

*b) **Sunnybrook Cardiac Dataset**:* The Sunnybrook Cardiac Dataset (SCD), also known as the 2009 Cardiac MR Left Ventricle Segmentation Challenge datasets, consists of 45 cine-MRI images from a mixed of patients and pathologies: healthy, hypertrophy, heart failure with infarction and heart failure without infarction. Each subset contains 15 cases of which 4 heart failure with infraction (HF-I), 4 heart failure without infraction (HF), 4 LV hypertrophy (HYP) and 3 healthy subjects. In all 45 samples, LV endocardial contours were drawn by an experienced cardiologist by taking 2D slices at both the end-systolic (ES) and end-diastolic phase (ED), and then independently confirmed by a second annotator. Each sequence has been acquired during a 10-15 second breath-holds, with a temporal resolution of 20 cardiac phases over thee heart cycle, starting from the ED cardiac phase, and containing 6 to 12

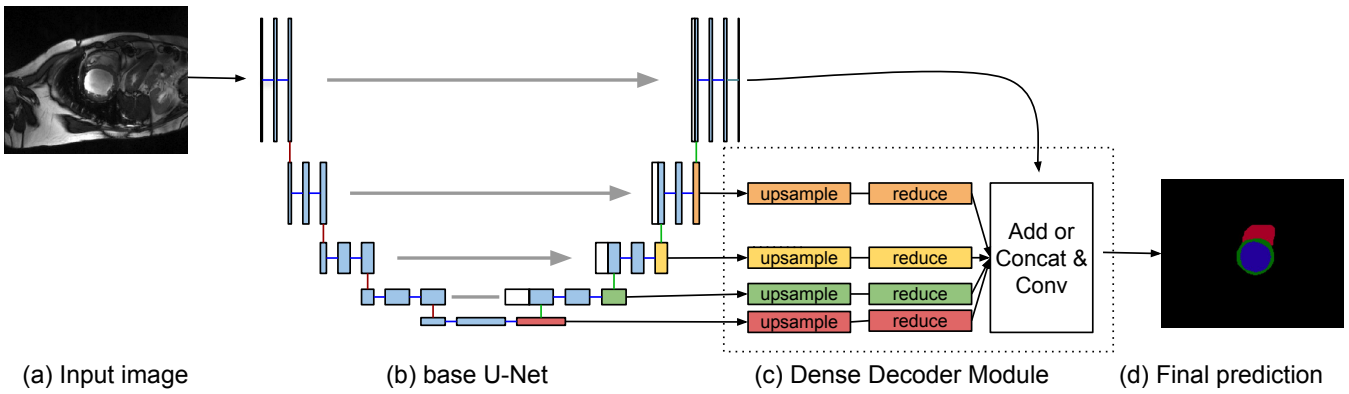

Fig. 2. Overview of our proposed Dense-Decoder U-Net. Given an input MRI image (a), an Encoder-Decoder architecture (b) is used to obtain feature maps at different levels (color blocks) that are upsampled and reduced in size along their depth dimension. The upsampling is done by a transposed convolution operation with n_{class} filters or by a bilinear upsampling operation followed with a convolution of n_{class} filters. The upsampled feature maps are aggregated together to the final prediction map of the model by adding them or concatenating them (c). A final convolutional operation is used to match the dimensions of the number of predicted classes (d).

SAX images obtained from the atrioventricular ring to the apex (thickness= $8mm$, gap= $8mm$, FOV= $320mm \times 320mm$, matrix= 256×256, in-plane resolution= $1.3mm - 1.4mm$). This dataset is already divided into training, validation and online set. The reported results were obtained using the validation set.

B. Evaluation Measures

To evaluate our results, we considered four different measurements: Dice metric [29], Hausdorff Distance [30], Ejection Fraction Correlation [31], and Performance Analysis.

a) ***Dice metric****:* The dice index measures the overlap between two areas (2D Dice index) or two volumes (3D Dice index). More formally, it is defined as:

$$D(A, B) = 2\frac{A \cap B}{A + B}$$

A and B are defined as the two areas or two volumes. The Dice index varies from 0 (complete mismatch) to 1 (perfect match).

b) ***Hausdorff Distance****:* The Hausdorff distance measures the distance between two areas (2D Hausdorff distance) or two volumes (3D Hausdorff distance). It is defined as:

$$H(A, B) = \max(\max_{a \in A}(\min_{b \in B} d(a, b)), \max_{b \in B}(\min_{a \in A} d(a, b)))$$

where d denotes the Euclidean distance. A smaller Hausdorff distance implies a better match. The Hausdorff distance is computed in millimeter with spatial resolution obtained from the PixelSpacing DICOM field of the MRI metadata.

c) ***Ejection Fraction Correlation****:* The Ejection fraction is an important metric that it is used by doctors at Hospitals and measures the quantity of blood pumped out of the heart in each beat as a percentage. A reduced EF is a common symptom in many cardio-vascular diseases. It is calculated using the volumes of two phases of the cardiac cycle, End-Diastole and End-Systole, and is calculated as:

$$EF = \frac{(Vol_{ED} - Vol_{ES})}{Vol_{ED}}$$

where:

Vol_{ED} is the calculated volume for an specific heart structure at the ED phase;

Vol_{ES} is the calculated volume for an specific heart structure at the ES phase.

To measure the similarity of Ejection Fraction between the ground-truth and the predictions, we use the Pearson correlation coefficient as in [2] and use this coefficient to report results.

d) ***Performance Analysis****:* For the performance analysis, we considered the convergence time and total number of parameters of the models.

V. Experiments

To evaluate our Dense-Decoder module, we chose as baselines the two best Encoder-Decoder architectures from the last ACDC competition [2], namely, the methods of Isensee [13] and Baumgartner [12]. These approaches ranked first and third, respectively. We further evaluated different configurations for our Dense-Decoder module combined with different number of stage feature maps from the decoder. For a fair comparison, the parameters of the network and the training procedure are the same as described in both baselines.

A. Pre-processing

The pre-processing procedures are the same as the ones used in [12] and [13]. All the MRI images are normalized to zero mean and unit variance. Next, both the MRI images and their ground-truth masks are re-sampled onto a new in-plane resolution. Lastly, these images are center-cropped to a specific size.

1) Training:

a) ***Baumgarter et al.'s model [12] with Dense-Decoder skip connection module****:* Similar to [12], our decoder is based on transposed convolutions. We trained our model from scratch by initializing its weights using the method of [32].

We optimize the standard pixel-wise cross entropy, which is defined as:

$$L_{crossentropy} = -\sum_{n=1}^{N}\sum_{k=1}^{K} y_{n,k} log(p_{n,k}) \quad (1)$$

where:

K is the number of classes;

N is the number of pixels/voxels;

$y_{n,k}$ is the k^{th} position in a one-hot vector encoding of the true label for pixel on n^{th} location;

$p_{n,k}$ is the k^{th} position in a softmax output vector encoding for pixel on n^{th} location.

To minimize the loss function, we used the ADAM optimizer [33] with a learning rate of 0.01, $B_1 = 0.9$ and $B_2 = 0.999$.

*b) **Isensee et al.'s model [13] with Dense-Decoder skip connection module**:* As in the work of [13], our decoder uses bilinear upsampling operations. To add the dense-decoder module, we first remove the deep supervision layers in their model. Next, we attach our module into the corresponding layers of the decoder. We initialize the model using a He initialization [32] with a normal distribution. We optimize the network using a multi-class dice loss [13], defined as:

$$L_{dc} = -\frac{2}{|K|}\sum_{k\in K}\frac{\sum_i u_i^k v_i^k}{\sum_i u_i^k + \sum_i v_i^k} \quad (2)$$

where:

u is the softmax output of the network;

v denotes a one hot encoding of the ground-truth segmentation map;

$k \in K$ being the classes.

To minimize the loss function, we used the ADAM optimizer with an initial learning rate of 5×10^{-4} and a learning rate decay of 0.98 per epoch. An epoch is set to 100 batches of 5 images each one. Since the training procedure is the same as in [13], we also used data augmentation for each batch: mirroring along the x and y axes, random rotations, gamma-correction, and elastic deformations.

B. Post-processing:

Since spurious predictions of heart structures might appear in implausible locations, we kept for every cardiac structure only the largest connected component from their predictions [12]. With this simple technique, we are able to reduce the number of false-positives.

C. Experiments on the ACDC Dataset

*a) **Ablation Study for the DenseDecoder U-Net**:* To find the best configuration for our DenseDecoder U-Net, we conducted experiments with different settings, e.g. the number and which features maps from the decoder, how to upsample and aggregate the selected feature maps. Combining the model of Baumgartner with the Dense-Decoder Modul3add (Unet Baum + Dense-Decoder Modul3add for short) has shown to produce effective results in terms of Dice metric except for the Right Ventricle on the End-Systole Phase, where U-Net Baum + Dense-Decoder Module1add has the best performance. In any case, using any configuration of the Dense-Decoder module outperforms the proposed architecture of Baumgartner. In addition, we note that when using concatenation instead of addition for combining the upsampled feature maps, the Hausdorff distance is also improved (see Table II). We found out that the best parameters used the three lowest feature maps from the decoder, and addition operations to combine the feature maps onto the final prediction map. From now on, we fix this configuration and use this setting to report results.

When plugging the Dense-Decoder module onto Isensee's model, we first removed the deep supervision part from their architecture and then add the Dense-Decoder module. The results are equivalent to Isensee's model with improvements on the End-Diastole phase for the Right Ventricle (see Table III).

Visual Analysis: We provide examples were the differences between the ground-truth and the proposed Dense-Decoder module on Baumgartner's model were higher (see Figures 3 and 4). We separate these cases in two types: the first in which the images are relatively easy to segment using deep learning (see Figure 3); and the second in which the images are more difficult to predict because of the size of the heart structure or the shape of the Right Ventricle (see Figure 4). From these two cases, we can see that using the Dense-Decoder module leads to more complete and smoother predictions, especially for the right ventricle.

Clinical Measurement: We report results on Ejection Fraction Correlation on Table IV. The Ejection Fraction Correlation is improved by our Dense-Decoder Module. Especially, when the Dense-Decoder Module3Add configuration is used, the improvement is about 0.9% over Baumgartner et al. [12] for the left ventricle. A major gain is obtained in the right ventricle, in which by using the Dense-Decoder Module1Add configuration, we obtain a gain of 3.9% over Baumgartnert's results. In this sense, our proposed Dense-Decoder Module demonstrates an improvement between 0.9% - 3.9% on clinical measurements for both ventricles.

D. Experiments on the Sunnybrook Cardiac Dataset

For the Sunnybrook Cardiac Dataset, we considered two baselines, a Fully Convolutional Network (FCN) with 15 layers from [10] and the Baumgartner's model, reporting the results on Table V. The FCN presented in [10] was one of the best methods for this dataset. From the results, we can see that the U-Net model outperforms the Fully Convolutional Network(FCN) in terms of Dice score but is worse in Average Distance than the FCN from [10]. This behaviour means that the predictions from the U-Net contain spurious structures which make the error higher in this metric.

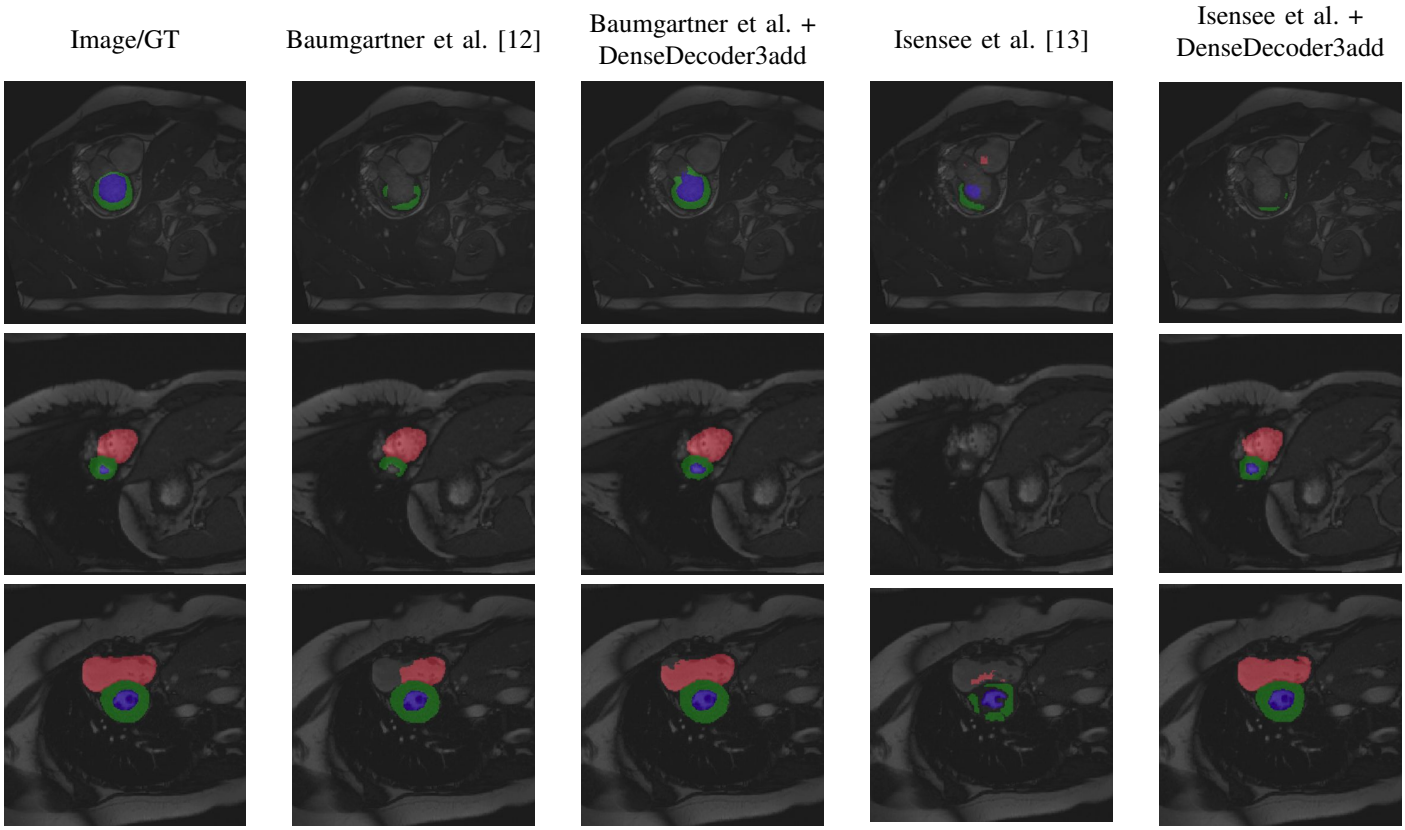

Fig. 3. Qualitative results on the ACDC [2] dataset for the right ventricle, myocardium and endocardium. Left to right columns: Ground-truth, Baumgartner [12], Baumgartner [12] + DenseDecoder3add, Isensee [13], Isensee [13] + DenseDecoder3add. Using the DenseDecoder module leads to more accurate predictions.

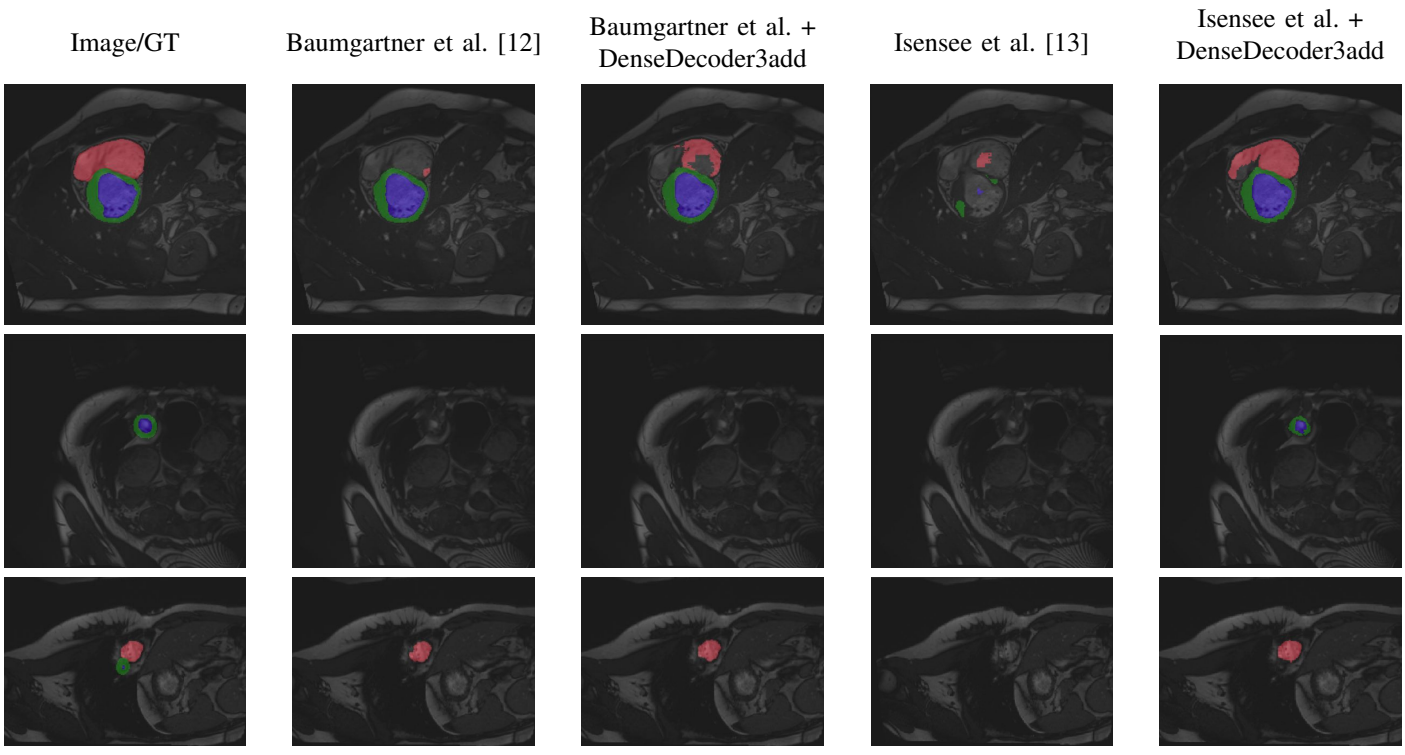

Fig. 4. Qualitative results on the ACDC [2] dataset for the right ventricle, myocardium and endocardium. Left to right columns: Ground-truth, Baumgartner [12], Baumgartner [12] + DenseDecoder3add, Isensee [13], Isensee [13] + DenseDecoder3add. Adding the Dense-Decoder module to Baumgartner's model improves the segmentation of the right ventricle.

Models	ED						ES					
	LV		RV		Myo		LV		RV		Myo	
	D ↑	d_H ↓	D ↑	d_H ↓	D ↑	d_H ↓	D ↑	d_H ↓	D ↑	d_H ↓	D ↑	d_H ↓
	val.	mm	val.	mm	val.	mm	val.	mm	val.	mm	val.	mm
Baumgartner et al. [12]	0.966	5.735	0.939	12.458	0.888	8.982	0.930	7.286	0.858	14.458	0.905	**8.677**
Baumgartner et al. [12] + DDM1Concat	0.966	5.053	0.939	11.924	0.886	**7.751**	0.932	6.895	0.831	15.771	0.901	9.203
Baumgartner et al. [12] + DDM1Add	0.967	5.557	0.939	13.736	**0.895**	8.002	0.941	7.047	0.858	**13.319**	0.906	9.605
Baumgartner et al. [12] + DDM2Add	0.964	5.871	0.936	13.353	0.884	9.340	0.940	6.326	0.843	14.064	0.904	10.103
Baumgartner et al. [12] + DDM3Add	**0.968**	**4.855**	**0.943**	**11.592**	0.891	8.865	**0.944**	**6.254**	**0.861**	14.276	**0.907**	8.716
Baumgartner et al. [12] + DDM4Add	0.965	5.640	0.939	12.115	0.882	9.029	0.930	6.927	0.847	14.344	0.897	9.188

TABLE II
ANALYSIS OF THE SEGMENTATION RESULTS ON THE ACDC DATASET COMPARING THE U-NET BY BAUMGARTNER ET AL. AND U-NET BAUMGARTNER ET AL. + DENSE-DECODER SKIP CONNECTION MODULE IN TERMS OF THE DICE COEFFICIENT (D) AND HAUSDORFF DISTANCE (d_H).

Models	ED				ES			
	LV ↑	RV ↑	Myo ↑	Avg. ↑	LV ↑	RV ↑	Myo ↑	Avg. ↑
Isensee et al. [13]	**0.961**	0.913	**0.882**	0.958	**0.903**	0.800	**0.896**	0.928
Isensee et al. [13] + DDM3Add	0.956	**0.921**	0.882	**0.959**	0.884	0.795	0.884	0.924
Isensee et al. [13] + DDM3Concat	0.955	0.919	0.880	0.958	0.895	**0.800**	0.888	**0.928**

TABLE III
ANALYSIS OF THE SEGMENTATION RESULTS ON THE ACDC DATASET BY COMPARING THE U-NET BY ISENSEE AND U-NET ISENSEE + DENSE-DECODER SKIP CONNECTION MODULE IN TERMS OF THE DICE COEFFICIENT.

Models	Correlation	
	Left Ventricle (EF) ↑	Right Ventricle (EF) ↑
Baumgartner et al. [12]	0.983	0.909
Baumgartner et al. [12] + DDM1Add	0.988	**0.948**
Baumgartner et al. [12] + DDM3Add	**0.992**	0.914

TABLE IV
EJECTION FRACTION CORRELATION.

Models	Dice ↑	Average Distance (mm) ↓
Tran's model [10]	0.904	**1.799**
Poudel et al.'s model [11]	0.900	2.050
Baumgartner et al.'s model [12]	0.917	1.856
Baumgartner et al. [12] + DDM3Add	**0.921**	1.879

TABLE V
SEGMENTATION RESULTS FOR LEFT VENTRICLE ENDOCARDIUM ON THE VALIDATION SET ON THE SUNNYBROOK CARDIAC DATA.

Models	Parameters	Model size (bytes)
Baumgartner et al. [12]	25.27M	101097888
Baumgartner et al. [12] + DDM1Concat	25.40M	101622400
Baumgartner et al. [12] + DDM1Add	25.40M	101622240
Baumgartner et al. [12] + DDM2Add	25.43M	101753344
Baumgartner et al. [12] + DDM3Add	25.45M	101818912
Baumgartner et al. [12] + DDM4Add	25.46M	101851712

TABLE VI
HARDWARE REQUIREMENTS.

E. Performance Analysis

Besides the analysis of the predicted segmentations, we also compared the training, convergence, and inference time among the different models. As we want to facilitate an automatic analysis of the heart-structures and to measure the EF for both LV and RV based on the predictions, the methods need to be fast during inference time. The fastest method in terms of inference time is Baumgartner's model. Nonetheless, as segmentation accuracy is a key factor, clinicians would also prefer to have a bit slower model but with better EF correlation, which is in this case the Baumgartner + Dense-Decoder Module3Add model.

An advantage of the Baumgartner's model over the standard U-Net is that the authors used much less feature maps in the upsampling path of the decoder by reducing the number of parameters by more than 90%. In addition, their method performs better than the standard U-Net for most of the predicted cardiac structures. Our proposed model performs better than Baumgartner's model itself without adding many additional parameters (See Table VI).

VI. CONCLUSIONS AND FUTURE WORK

In this work, we introduced the *Dense-Decoder Module* which can be easily added to state-of-the-art encoder-decoder architectures. It has been shown that our approach can lead to an improvement on the total Dice score for the segmentation of the heart on two challenging datasets, namely the the ACDC [2] and LVSC [3] heart segmentation challenges.

The main benefits of our approach, include: (i) exploiting different levels of context from the decoder part by combining the corresponding feature maps directly onto the final predictions, (ii) obtaining a smoother loss landscape and better convergence as the gradients can flow directly from the final outputs to the decoder layers during back-propagation, and (iii) finally, the size of the model remains constant as practically no extra parameters are added. As future work, we are planning to exploit the geometrical properties of the heart structures. More precisely, since each heart structure presents a specific shape, we are planning to add shape information

into our network and train an end-to-end model. Our initial results have shown that this is a promising strategy.

ACKNOWLEDGMENTS

The authors are grateful to CNPq (grants #307560/2016-3 and #303808/2018-7), São Paulo Research Foundation – FAPESP (grants #2014/12236-1, #2015/24494-8, #2016/50250-1, and #2017/20945-0), the FAPESP-Microsoft Virtual Institute (grants #2013/50155-0 and #2014/50715-9), Swiss National Science Foundation – SNSF (grant # 32003B_159727), a Google Cloud Research Award, and a Titan Xp GPU donation from NVIDIA Corporation. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. The present work was also supported by grant 234-2015-FONDECYT (Master Program) from Cienciactiva of the National Council for Science, Technology and Technological Innocation (CONCYTEC-PERU).

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[2] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester *et al.*, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE transactions on medical imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[3] P. Radau, Y. Lu, K. Connelly, G. Paul, A. Dick, and G. Wright, "Evaluation framework for algorithms segmenting short axis cardiac mri," *The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge*, vol. 49, 2009.

[4] E. J. Benjamin, S. S. Virani, C. W. Callaway, A. M. Chamberlain, A. R. Chang, S. Cheng, S. E. Chiuve, M. Cushman, F. N. Delling, R. Deo *et al.*, "Heart disease and stroke statistics-2018 update: a report from the american heart association," *Circulation*, vol. 137, no. 12, p. e67, 2018.

[5] S. S. Islam, S. Rahman, M. M. Rahman, E. K. Dey, and M. Shoyaib, "Application of deep learning to computer vision: A comprehensive study," in *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*. IEEE, 2016, pp. 592–597.

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[7] Z. Wojna, V. Ferrari, S. Guadarrama, N. Silberman, L.-C. Chen, A. Fathi, and J. Uijlings, "The devil is in the decoder: Classification, regression and gans," *International Journal of Computer Vision*, pp. 1–13, 2019.

[8] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8798–8807.

[9] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, "No new-net," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 234–244.

[10] P. V. Tran, "A fully convolutional neural network for cardiac segmentation in short-axis mri," *arXiv preprint arXiv:1604.00494*, 2016.

[11] R. P. Poudel, P. Lamata, and G. Montana, "Recurrent fully convolutional neural networks for multi-slice mri cardiac segmentation," in *Reconstruction, segmentation, and analysis of medical images*. Springer, 2016, pp. 83–94.

[12] C. F. Baumgartner, L. M. Koch, M. Pollefeys, and E. Konukoglu, "An exploration of 2d and 3d deep learning techniques for cardiac mr image segmentation," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2017, pp. 111–119.

[13] F. Isensee, P. F. Jaeger, P. M. Full, I. Wolf, S. Engelhardt, and K. H. Maier-Hein, "Automatic cardiac disease assessment on cine-mri via time-series segmentation and domain specific features," in *International workshop on statistical atlases and computational models of the heart*. Springer, 2017, pp. 120–129.

[14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[15] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," in *Advances in Neural Information Processing Systems*, 2018, pp. 6389–6399.

[16] A. E. Orhan and X. Pitkow, "Skip connections eliminate singularities," *arXiv preprint arXiv:1701.09175*, 2017.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] H. Liu, H. Hu, X. Xu, and E. Song, "Automatic left ventricle segmentation in cardiac mri using topological stable-state thresholding and region restricted dynamic programming," *Academic radiology*, vol. 19, no. 6, pp. 723–731, 2012.

[19] J. Ulén, P. Strandmark, and F. Kahl, "An efficient optimization framework for multi-region segmentation based on lagrangian duality," *IEEE transactions on medical imaging*, vol. 32, no. 2, pp. 178–188, 2013.

[20] T. Chen, J. Babb, P. Kellman, L. Axel, and D. Kim, "Semiautomated segmentation of myocardial contours for fast strain analysis in cine displacement-encoded mri," *IEEE Transactions on Medical Imaging*, vol. 27, no. 8, pp. 1084–1094, 2008.

[21] I. B. Ayed, H.-m. 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.

[22] S. Queirós, D. Barbosa, B. Heyde, P. Morais, J. L. Vilaça, D. Friboulet, O. Bernard, and J. Dhooge, "Fast automatic myocardial segmentation in 4d cine cmr datasets," *Medical image analysis*, vol. 18, no. 7, pp. 1115–1131, 2014.

[23] S. C. Mitchell, J. G. Bosch, B. P. Lelieveldt, R. J. Van der Geest, J. H. Reiber, and M. Sonka, "3-d active appearance models: segmentation of cardiac mr and ultrasound images," *IEEE transactions on medical imaging*, vol. 21, no. 9, pp. 1167–1178, 2002.

[24] W. Bai, W. Shi, C. Ledig, and D. Rueckert, "Multi-atlas segmentation with augmented features for cardiac mr images," *Medical image analysis*, vol. 19, no. 1, pp. 98–109, 2015.

[25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[26] J. Patravali, S. Jain, and S. Chilamkurthy, "2d-3d fully convolutional neural networks for cardiac mr segmentation," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2017, pp. 130–139.

[27] X. Yang, C. Bian, L. Yu, D. Ni, and P.-A. Heng, "Class-balanced deep neural network for automatic ventricular structure segmentation," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2017, pp. 152–160.

[28] B. Kayalibay, G. Jensen, and P. van der Smagt, "Cnn-based segmentation of medical imaging data," *arXiv preprint arXiv:1701.03056*, 2017.

[29] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[30] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[31] P. Peng, K. Lekadir, A. Gooya, L. Shao, S. E. Petersen, and A. F. Frangi, "A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 29, no. 2, pp. 155–195, 2016.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.