City, University of London Institutional Repository

3D HIGH-RESOLUTION CARDIAC SEGMENTATION RECONSTRUCTION FROM 2D VIEWS USING CONDITIONAL VARIATIONAL AUTOENCODERS

Carlo Biffi⋆,†, *Juan J. Cerrolaza*⋆, *Giacomo Tarroni*⋆, *Antonio de Marvao*†,
Stuart A. Cook†, *Declan P. O'Regan*†, *Daniel Rueckert*⋆

⋆Biomedical Image Analysis Group, Imperial College London, UK.
†MRC London Institute of Medical Sciences, Imperial College London, UK.

ABSTRACT

Accurate segmentation of heart structures imaged by cardiac MR is key for the quantitative analysis of pathology. High-resolution 3D MR sequences enable whole-heart structural imaging but are time-consuming, expensive to acquire and they often require long breath holds that are not suitable for patients. Consequently, multiplanar breath-hold 2D cines sequences are standard practice but are disadvantaged by lack of whole-heart coverage and low through-plane resolution. To address this, we propose a conditional variational autoencoder architecture able to learn a generative model of 3D high-resolution left ventricular (LV) segmentations which is conditioned on three 2D LV segmentations of one short-axis and two long-axis images. By only employing these three 2D segmentations, our model can efficiently reconstruct the 3D high-resolution LV segmentation of a subject. When evaluated on 400 unseen healthy volunteers, our model yielded an average Dice score of 87.92 ± 0.15 and outperformed competing architectures (TL-net, Dice score = 82.60 ± 0.23, $p = 2.2 \cdot 10^{-16}$).

Index Terms— Cardiac MR, Variational Autoencoder, 3D Segmentation Reconstruction, Deep Learning.

I. INTRODUCTION

Cardiac magnetic resonance (CMR) is the gold-standard technique for assessment of cardiac morphology. Conventional practice is to acquire a stack of breath-hold 2D image sequence in the left ventricular (LV) short axis supplemented by long axis image sequence in prescribed planes to enable reproducible volumetric analysis and diagnostic assessment [1]. Disadvantages of this approach for whole-heart segmentation are low through-plane resolution, misalignment between breath-holds and lack of whole-heart coverage. High-resolution 3D image sequences address some of these issues, but also have disadvantages in terms of long acquisition times, relatively low in-plane resolution and lack of clinical availability. However, high-resolution 3D segmentations proved to be crucial for the construction of integrative statistical models of cardiac anatomy and physiology and disease characterization [2], [3]. For these reasons, a method to reconstruct a 3D high-resolution segmentation from routinely-acquired 2D cines could be highly beneficial - offering high resolution phenotyping robust to artefact in large clinical populations with conventional imaging.

The reconstruction of 3D anatomical structures from a limited number of 2D views has been previously studied via deformable statistical shape models [4]. However, these methods require complex reconstruction procedures and are very computationally-intensive. In recent years, with the advent of learning-based approaches, and in particular of deep learning, a number of alternative strategies have been proposed. The TL-embedding network (TL-net) consists of a 3D convolutional autoencoder (AE) which learns a vector representation of the 3D geometries, whereas a second convolutional neural network attached to the latent space of the AE maps 2D views of the same object to the same vector representation [5]. More recently, [6] proposed a convolutional conditional variational autoencoder (CVAE) architecture for the 3D reconstruction of the fetal skull from 2D ultrasound standard planes of the head. Finally, [7] showed how a convolutional variational autoencoder (VAE) can learn a shape segmentation model of left ventricular (LV) segmentations and how the learned latent space can be exploited to accurately identify healthy and pathological cases and generate realistic segmentations unseen during training.

In this work, we present a CVAE architecture that reconstructs a high-resolution 3D segmentation of the LV myocardium from three segmentations of 2D standard cardiac views (one short-axis and two long-axis). Moreover we show how the proposed model naturally produces confidence maps associated to each reconstruction, unlike deterministic models, thanks to its generative properties.

II. MATERIALS AND METHODS

II-A. 3D Cardiac Image Acquisition and Segmentation

A high-spatial resolution 3D balanced steady-state free precession cine MR image sequence was acquired from 1,912 healthy volunteers of the UK Digital Heart Project at Imperial College London using a 1.5-T Philips Achieva system (Best, the Netherlands) [3]. Left and right ventricles
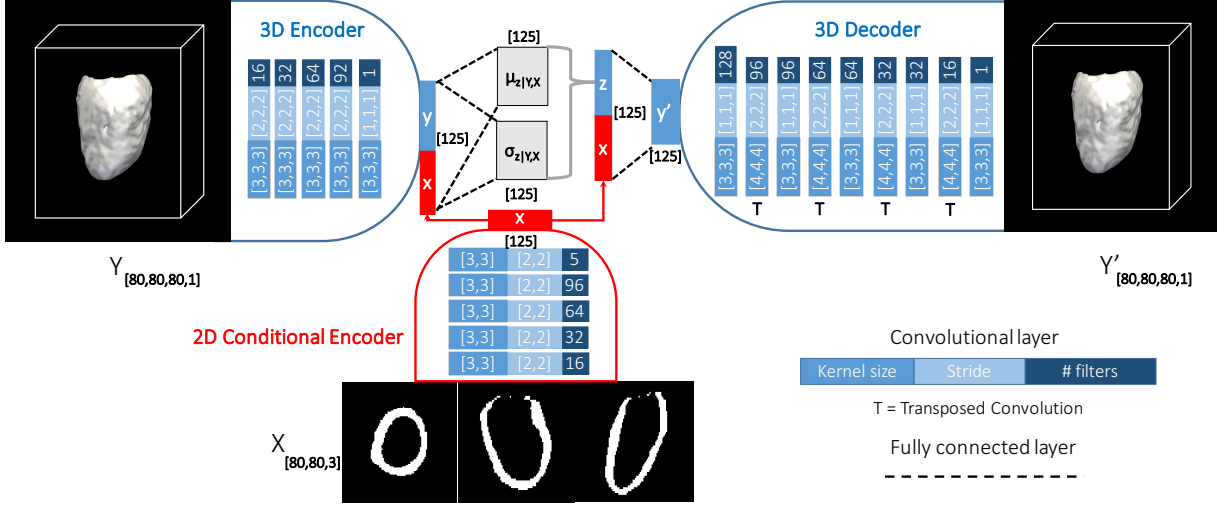
Fig. 1. The proposed conditional variational autoencoder (CVAE) architecture.

were imaged in their entirety in a single breath-hold (60 sections, repetition time 3.0 ms, echo time 1.5 ms, flip angle 50°, field of view $320 \times 320 \times 112$ mm, matrix 160×95, reconstructed voxel size $1.2 \times 1.2 \times 2$ mm, 20 cardiac phases, temporal resolution 100 ms, typical breath-hold 20 s). For each subject, a 3D high-resolution segmentation of the LV was automatically obtained using a previously reported technique employing a set of manually annotated atlases [3]. In this work, only the end-diastolic (ED) frame was considered.

II-B. Conditional Variational Autoencoder Architecture

The outline of the CVAE architecture we propose is shown in Fig. 1. We aim at reconstructing a 3D high-resolution LV segmentation $\mathbf{Y}$ from i segmentations obtained in as many 2D views $\mathbf{X} = \{X_i \mid i = 1, 2, 3\}$. We aim to learn from the training data a conditional generative model $P(\mathbf{Y}|\mathbf{X})$ by means of a d-dimensional latent distribution $\mathbf{z}$ and a low-dimensional representation $\mathbf{x}$ of the views $\mathbf{X}$. In this work we use a single 2D convolutional neural network (CNN) to encode the 2D views $\mathbf{X}$ in a low-dimensional representation $\mathbf{x}$. An alternative encoding strategy was proposed in [6], using a separate branch for each conditional input of the model. However, whilst this latter approach proved efficient when the views suffer from large inconsistencies or variability (e.g., free-hand ultrasound scans), we can notably reduce the model complexity by combining the views $\mathbf{X}$ as a unique three-channel input as these are consistently acquired in clinical routine.

Directly inferring $P(\mathbf{Y}|\mathbf{X})$ is impractical as it would require sampling a large number of $\mathbf{z}$ values. However, variational inference allows us to approximate $P(\mathbf{Y}|\mathbf{X})$ by introducing a high-capacity function $\mathbf{Q}(\mathbf{z}|\mathbf{Y}, \mathbf{X})$ which gives us a distribution over $\mathbf{z}$ values that are likely to produce $\mathbf{Y}$.

Hence we can learn $P(\mathbf{Y}|\mathbf{X})$ by minimizing the following objective:

$$log(P(\mathbf{Y}|\mathbf{X})) - \mathbf{D_{KL}}[\mathbf{Q}(\mathbf{z}|\mathbf{Y}, \mathbf{X})||\mathbf{P}(\mathbf{z} \mid \mathbf{Y}, \mathbf{X})] =$$

$$E_{\mathbf{z} \sim \mathbf{Q}}[log P(\mathbf{Y}|\mathbf{z}, \mathbf{X})] - \mathbf{D_{KL}}[\mathbf{Q}(\mathbf{z}|\mathbf{Y}, \mathbf{X})||\mathbf{P}(\mathbf{z} \mid \mathbf{X})] \quad (1)$$

where D_{KL} represents the Kullback-Leibler (KL) divergence of two distributions (full mathematical derivation of the equation can be found in [8]). The encoding function $\mathbf{Q}(\mathbf{z}|\mathbf{Y}, \mathbf{X})$ can be modelled as a Gaussian distribution parametrized by $\mu_{\mathbf{z}|\mathbf{Y},\mathbf{X}}$ and $\sigma_{\mathbf{z}|\mathbf{Y},\mathbf{X}}$ vectors. These two vectors can be learned by encoding the input 3D segmentation $\mathbf{Y}$ we want to reconstruct via a 3D CNN to a set of features $\mathbf{y}$, which are then concatenated together with the lower dimensional representation $\mathbf{x}$ of the views $\mathbf{X}$. By concatenating $[\mathbf{x}, \mathbf{y}]$ with a fully connected neural network to $\mu_{\mathbf{z}|\mathbf{Y},\mathbf{X}}$ and $\sigma_{\mathbf{z}|\mathbf{Y},\mathbf{X}}$ we can thus learn $\mathbf{Q}(\mathbf{z}|\mathbf{Y}, \mathbf{X})$.

If $\mathbf{Q}(\mathbf{z}|\mathbf{Y}, \mathbf{X})$ is modelled by a sufficiently expressive function, then this function will match the real $P(\mathbf{z} \mid \mathbf{Y}, \mathbf{X})$ and the $D_{KL}[Q(\mathbf{z}|\mathbf{Y}, \mathbf{X})||\mathbf{P}(\mathbf{z} \mid \mathbf{Y}, \mathbf{X})]$ term in (1) will be zero. Therefore optimizing the right side of (1) will correspond to optimizing $P(\mathbf{Y}|\mathbf{X})$. In this work, the first term of the right side of (1) is computed as the Dice score (DSC) between $\mathbf{Y}$ and its reconstruction $\hat{\mathbf{Y}}$, which is the output of the generative model. The second term in (1) can be computed in a closed form if we assume its prior distribution to be $\mathcal{N}(\mathbf{0}, \mathbf{1})$, a d-dimensional normal distribution with zero mean and unit-standard deviation, and where d is the number of dimensions of the latent space. Therefore the loss function we optimize becomes $\mathcal{L} = DSC(\mathbf{Y}, \hat{\mathbf{Y}}) + \alpha \, \mathbf{D_{KL}}[\mathbf{z}||\mathcal{N}(\mathbf{0}, \mathbf{1})]$.

II-C. Experimental Setup and Network Training

In this work, we mimicked the two long-axis and the one short-axis views acquired in a routine acquisition with the

following steps: (1) we rigidly aligned all the ground truth 3D high-resolution segmentations by performing landmark-based and subsequent intensity-based rigid registration; (2) we kept only the LV myocardium label and we cropped and padded the segmentations to [x = 80, y = 80, z = 80, t = 1] dimension using a bounding box centered at the centre of mass of the LV myocardium; (3) we sampled three orthogonal views passing through the centre of each segmentation (an example is shown in Fig. 1). Thanks to this process we extracted three 2D views showing the same three LV sections consistently for all subjects. In the following experiments, the ground truth 3D high-resolution segmentations and their corresponding 2D views were kept all in the same reference space. Inter-subject pose variability will be addressed in future work, potentially with a simple data augmentation strategy.

The dimension d of the latent space was fixed to 125 as values smaller than 100 provided less accurate results, while above 125 no further improvements were observed. The dimensionality of the low dimensional representation $\mathbf{x}$ was kept equal to the dimensionality of $\mathbf{z}$ to guarantee a balanced contribution to the generative model. Simulations for different values of the parameter α in the loss function were performed: low values of α ($\alpha < 0.5$) provided better reconstruction results on the training data at the expenses of a strong deviation from normality of the latent space distribution (KL term not converging) causing overfitting. Higher values of α ($\alpha > 2$) penalized the reconstruction term in favour of a strictly normal latent space, hence providing poorer reconstruction accuracy. In this work we set $\alpha = 1$ as this provided good reconstruction accuracy and convergence of the KL term.

Experiments were performed with different numbers of views X_i as conditions for the proposed model. In particular, referring to the first long-axis view as 1, the second long-axis view as 2 and the short-axis view as 3, we performed the training using either only one view (which we will indicate as CVAE_1), or a combination or two views (CVAE_12, CVAE_23, CVAE_13), or all the three views (CVAE_123). We have also studied the feasibility of training a 2D AE to reconstruct the 3 views and used its encoder as a pre-trained conditional encoder (pCVAE_123). Moreover, the reconstruction capability of the proposed architecture was compared with the one of the TL-net [5]. Finally, we compared the reconstruction obtained by a VAE with $\mathbf{z}$=0 (VAE_0) to all our test segmentations, as this represents the best segmentation that the generative model can reconstruct when no information is provided to it. Results obtained with an autoencoder (AE) are also reported since this model yielded better results than different VAEs with distinct α values as it only optimizes the reconstruction accuracy. All the models share the same 3D encoder and decoder architectures.

The dataset was split into training, evaluation and testing

| Model	DSC	Hausd. [mm]	MassDiff [%]
VAE_0	65.48 ± 0.38	9.32 ± 0.06	35.37 ± 0.70
CVAE_1	78.08 ± 0.33	5.29 ± 0.04	3.94 ± 0.38
CVAE_23	82.90 ± 0.21	4.43 ± 0.04	3.93 ± 0.19
CVAE_12	85.21 ± 0.20	4.46 ± 0.04	3.73 ± 0.19
CVAE_13	83.18 ± 0.18	4.77 ± 0.04	3.69 ± 0.19
CVAE_123	**87.92 ± 0.15**	**3.99 ± 0.03**	**2.70 ± 0.14**
pCVAE_123	87.63 ± 0.16	4.04 ± 0.04	3.05 ± 0.16
TL_net	82.60 ± 0.23	4.66 ± 0.04	3.85 ± 0.19
AE	90.45 ± 0.12	3.46 ± 0.03	1.50 ± 0.10

Table I. Reconstruction metrics together with their standard error of the mean for all the studied models.

sets consisting of 1362, 150 and 400 subjects respectively. Data augmentation included rotation around the three orthogonal axis with rotation angles randomly extracted from a normal distribution $\mathcal{N}(0, 6°)$ and random closing and opening morphological operations. All the networks were implemented in Tensorflow and training was stopped after 300k iterations, when the total validation loss function had stopped improving (approximately 42 hours per network on an NVIDIA Tesla K80 GPU), using stochastic gradient descent with momentum (Adam optimizer, learning rate = 10^{-4}) and batch size of 8. During testing, the 3D encoder branch was disabled and the reconstruction were obtained by setting the latent variables $\mathbf{z} = \mathbf{0}$.

III. RESULTS AND DISCUSSION

III-A. Accuracy of 3D Reconstruction

Table 1 shows the reconstruction accuracy in terms of 3D Dice score, 2D slice-by-slice Hausdorff distance and LV mass difference between 3D high-resolution segmentations (ground truth and reconstructed ones) for all the studied architectures. LV mass is an important clinical biomarker, therefore we have estimated for each reconstruction its percentage difference in mass with the ground truth. The results indicate that the reconstruction accuracy decreases when views are removed. From the experiments with two views we can also infer how different views have different importance. In particular, the short-axis view seems to have the smallest impact on the reconstruction accuracy. This could be motivated by the fact that the long-axis views contain more information about the regional changes in curvature of the LV, which strongly influences the Dice Score. The results reported in Table 1 also show how our architecture significantly outperforms the TL-net by a large amount ($p = 2.2 \cdot 10^{-16}$), and how the pre-training of the 2D CNN encoder network did not help to achieve better results. Finally, we can observe that the mass difference is systematically overestimated by a small amount that decreases with the number of views provided. We believe this a consequence of using the Dice score in the loss function. On the other hand, models trained using cross entropy in the loss function
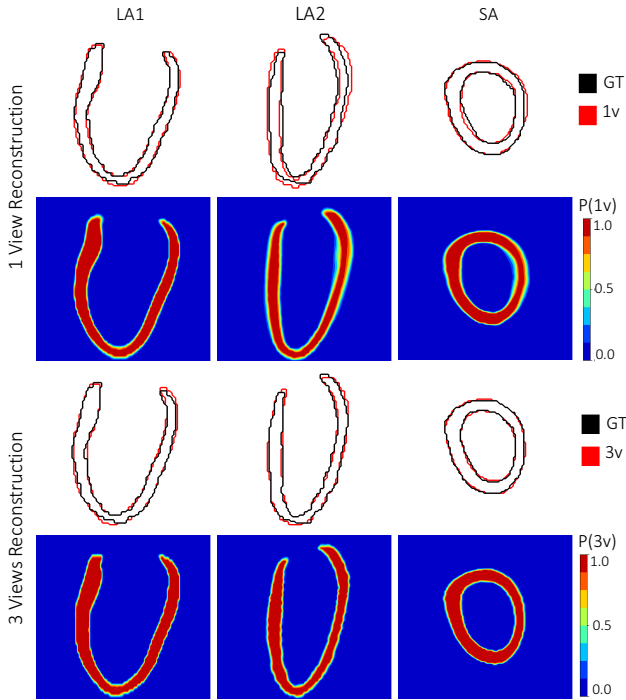
Fig. 2. First and third rows, reconstructed segmentation obtained with one and three views (in red, 1v and 3v) overlaid onto the ground truth segmentation (in black, GT) for one random subject. Second and fourth rows, confidence maps for the reconstruction with one and three views - $P(1v)$ and $P(3v)$. First and second columns, long-axis views (LA1 and LA2). Third column, short-axis (SA) view.

yielded a systematic underestimation of the mass, often with reconstructions with missing LV apex, as this loss term tends to favour the background instead of the myocardium.

III-B. Visualisation and Uncertainty Estimation

In the first and third rows of Fig.2 we report the reconstructed segmentations obtained with one and three views (in red) overlaid onto the ground truth segmentation (in black) for one subject of the testing dataset (with DSC 0.80 and 0.89, respectively). In the second and fourth rows we instead report the confidence maps obtained for the reconstruction with one and three views - $P(1v)$ and $P(3v)$. These maps have been obtained by sampling N times ($N = 1,000$) $\mathbf{z}$ from $\mathcal{N}(\mathbf{0}, \mathbf{1})$ to reconstruct N segmentations from the same set of views $\mathbf{X}$. Unlike deterministic architectures (such as the TL-net), by averaging these maps we can compute the probability of each voxel to be labelled as LV myocardium, providing to clinicians a richer and more intuitive interpretation of the reconstruction. It can be seen in Fig. 2 how the confidence map obtained with only 1 view has greater uncertainty than the one obtained with 3 views,

which instead shows lower variability. Moreover, the amount of uncertainty in the $P(1v)$ map for the long-axis view 1 is less than for the other two views, as this view was the one provided to the network as condition. Interestingly, in the reconstruction with one view the areas with more uncertainty correspond to the areas where there is less overlap with the ground truth, i.e. the areas where the network is less accurate in predicting the shape.

IV. CONCLUSIONS

In this paper we present the first deep conditional generative network for the reconstruction of 3D high-resolution LV segmentations from three segmentations of 2D orthogonal views. The reported results show the potential of this class of models to provide better quantitative cardiac models from sparse data. Future work will focus on using real standard long-axis views (instead of the simulated ones in this work), on reconstructing multiple structures and on extending the proposed framework to pathological datasets, for which acquiring breath-hold sequences is even more challenging.

V. REFERENCES

[1] K. Alfakih et al., "Assessment of ventricular function and mass by cardiac magnetic resonance imaging," *European radiology*, vol. 14, no. 10, pp. 1813–1822, 2004.

[2] C. Biffi et al., "Three-dimensional cardiovascular imaging-genetics: a mass univariate framework," *Bioinformatics*, vol. 34, no. 1, pp. 97–103, 2018.

[3] W. Bai et al., "A bi-ventricular cardiac atlas built from 1000+ high resolution mr images of healthy subjects and an analysis of shape and motion," *Medical image analysis*, vol. 26, no. 1, pp. 133–145, 2015.

[4] T. Whitmarsh et al., "Reconstructing the 3d shape and bone mineral density distribution of the proximal femur from dual-energy x-ray absorptiometry," *IEEE transactions on medical imaging*, vol. 30, no. 12, pp. 2101–2114, 2011.

[5] R. Girdhar et al., "Learning a predictable and generative vector representation for objects," in *European Conference on Computer Vision*. Springer, 2016, pp. 484–499.

[6] J.J. Cerrolaza et al., "3d fetal skull reconstruction from 2dus via deep conditional generative networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 383–391.

[7] C. Biffi et al., "Learning interpretable anatomical features through deep generative models: Application to cardiac remodeling," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 464–471.

[8] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.