Medical Image Analysis 77 (2022) 102354

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Learning to complete incomplete hearts for population analysis of cardiac MR images

Yan Xia[a,b,*], Nishant Ravikumar[a,b], Alejandro F. Frangi[a,b,c,d,*]

[a] Centre for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), School of Computing, University of Leeds, Leeds, UK
[b] Leeds Institute for Cardiovascular and Metabolic Medicine (LICAMM), School of Medicine, University of Leeds, Leeds, UK
[c] Medical Imaging Research Center (MIRC), University Hospital Gasthuisberg, and Cardiovascular Science and Electronic Engineering Departments, KU Leuven, Leuven, Belgium
[d] Alan Turing Institute, London, UK

ARTICLE INFO

Article history:
Received 18 December 2020
Revised 10 November 2021
Accepted 3 January 2022
Available online 13 January 2022

Keywords:
Cardiac MRI
Incomplete heart coverage
Data imputation
Conditional generative adversarial net
Convlstm

ABSTRACT

Cardiac MR acquisition with complete coverage from base to apex is required to ensure accurate subsequent analyses, such as volumetric and functional measurements. However, this requirement cannot be guaranteed when acquiring images in the presence of motion induced by cardiac muscle contraction and respiration. To address this problem, we propose an effective two-stage pipeline for detecting and synthesising absent slices in both the apical and basal region. The detection model comprises several dense blocks containing convolutional long short-term memory (ConvLSTM) layers, to leverage through-plane contextual and sequential ordering information of slices in cine MR data and achieve reliable classification results. The imputation network is based on a dedicated conditional generative adversarial network (GAN) that helps retain key visual cues and fine structural details in the synthesised image slices. The proposed network can infer multiple missing slices that are anatomically plausible and lead to improved accuracy of subsequent analyses on cardiac MRIs, e.g., ventricle segmentation, cardiac quantification compared to those derived from incomplete cardiac MR datasets. For instance, the results obtained when compensating for the absence of two basal slices show that the mean differences to the reference of stroke volume and ejection fraction are only -1.3 mL and -1.0%, respectively, which are significantly smaller than those calculated from the incomplete data (-26.8 mL and -6.7%). The proposed approach can improve the reliability of high-throughput image analysis in large-scale population studies, minimising the need for re-scanning patients or discarding incomplete acquisitions.

1. Introduction

Cardiovascular diseases (CVDs) remain the leading cause of deaths worldwide. They are responsible for a large proportion of premature mortality, e.g., 30.4% and 25.3% death rates for men and women in Europe, before age 65. Cardiac Magnetic Resonance (CMR) cine imaging is the gold standard in cardiovascular medicine, providing key diagnostic information for various clinical applications through morphological and functional left ventricular (LV) quantification (Pennell, 2003).

Due to its excellent reproducibility of quantitative measurements compared with other modalities, CMR cine imaging is a robust and attractive technique for large-scale population studies to identify new cardiovascular biomarkers for improved diagnosis of CVDs at early stages. Many initiatives have been launched, including the UK Biobank (UKBB) (Petersen et al., 2013), the German National Cohort (Bamberg et al., 2015), and the Canadian Alliance for Healthy Hearts and Minds (CAHHM) (Anand et al., 2016).

Cardiac MR acquisition with complete ventricular coverage from base to apex is required to ensure any subsequent analyses are reliable, such as ventricle segmentation, quantification of cardiac functional and morphological indices, strain and deformation analysis through non-rigid registration, cardiac shape modelling and computational mesh generation for use in mechanical and flow simulations, etc. The critical point for the acquisition is to find the direction of the LV long axis defined by the line going through from the apex to the centre of the mitral valve, and to acquire the SAX stack encompassing both landmarks with slices perpendicular to the long-axis. However, due to insufficient radiographer experience during scan acquisition planning, natural cardiac muscle contraction, breathing motion, and imperfect triggering, the acquired SAX

[*] Corresponding author at: Centre for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), School of Computing and School of Medicine, University of Leeds, Leeds, UK.

E-mail addresses: y.xia@leeds.ac.uk (Y. Xia), a.frangi@leeds.ac.uk (A.F. Frangi).

https://doi.org/10.1016/j.media.2022.102354
1361-8415/© 2022 Published by Elsevier B.V.

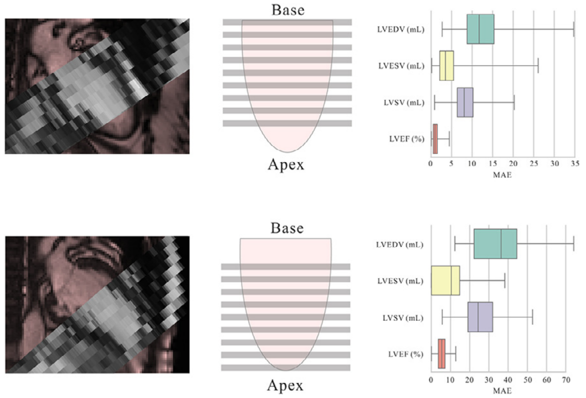

Fig. 1. Illustration of incomplete cardiac coverage present in CMR images. The first column shows 2 subjects with the incomplete SAX stack overlaid on top of LAX 2-chamber view. The rightmost column shows the box plots of absolute differences for 4 typical clinical parameters, namely LVEDV, LVESV, LVSV and LVEF, derived between incomplete (two slices missing) and complete coverage data. The CMR images were reproduced with the permission of UK Biobank.

stack may display a certain degree of sub-optimal cardiac coverage with an insufficient number of slices, as illustrated in Fig. 1. Such clinically acquired data with incomplete ventricular coverage not only hinder visual interpretation, but also pose challenges to downstream analyses. For instance, ventricular volume measurements, and associated Ejection Fraction (EF) and Stroke Volume (SV), are important in the management of various CVDs because they are strong predictors of clinical outcomes. However, accurate volume and functional measurements cannot be guaranteed if the base and apex of the ventricular chambers are not fully covered in short-axis (SAX) image stacks, which poses further challenges to quantitative LV characterisation and accurate diagnosis. Incomplete coverage can also adversely impact 3D cardiac shape reconstruction algorithms or non-rigid registration techniques developed for assessing myocardial strain across the cardiac cycle.

Quality assessment is challenging in large-scale population studies, where CMR images are acquired across different imaging centres before core lab analysis. Not only repetitive quality assurance tasks cannot maintain consistency and reliability for large amounts of imaging data, but also large volumes of data may be stored without being qualitatively checked by experienced radiographers before analysis Ferreira et al. (2013). Consequently, sub-optimal coverage may occur at any time point throughout the cardiac cycle, leading to unreliable CMR image data.

In post-processing, a common strategy to account for insufficient cardiac coverage, if identified, is to discard incomplete samples in the cohort without providing feedback to the acquisition team and correcting/re-acquiring the data (Klinke et al., 2013). Excluding incomplete data, however, not only reduces statistical power aggregated over the cohort and causes bias, but is also of ethical and financial concern as partially acquired subject data remains unused, and limits the application of post-processing methods used to analyse the data. Thus, developing robust and generic techniques to compensate for incomplete cardiac coverage in CMRs can have a transformative impact on population imaging applications by preventing incomplete data from being disregarded when analysing any cohort.

So far, most previous studies have focused solely on detecting incomplete heart coverage in CMR cine images, without providing the means to mitigate this issue. One group of such methods aims to identify the absence of the basal and apical slices in the SAX image stack using convolutional neural networks (CNN) or generative adversarial networks (GANs), to learn a feature representation that in turn can be used for accurate classification (Zhang et al., 2017; 2018b). Another category of heart coverage estimation is based on the automated detection of specific cardiac landmarks (i.e. the apex and the centre of the mitral valve) from the acquired images. For instance, (Tarroni et al., 2018) proposed a decision forest method to detect the landmarks on long-axis 2-chamber view images, and used landmark positions to evaluate the space encompassed by the acquired SAX stack and to estimate the coverage.

Although automated learning-based image quality control (QC) techniques have been studied, to our knowledge, no data completion methods have been proposed for incomplete coverage in CMR images in the literature. Few studies for medical data imputation exist but focus on interpolation only. One strategy is to fill in the missing slices by identifying redundant, relevant detail in a scan and re-synthesising high-frequency information (Manjón et al., 2010; Plenge et al., 2013). However, medical images are often sparsely acquired and hence it is hard to accurately estimate functional representations without prior knowledge. To obtain enough fine-scale information to recover the missing data, (Dalca et al., 2018) proposed a probabilistic generative model that captures repetitive anatomical structure across subjects in clinical image collections and derived an algorithm for filling in the middle slices in scans with large through-plane spacing. There also exist methods that attempt to exploit the temporal aspect of dynamic CMR data, to recover important image features and render a high-resolution sequence (Basty and Grau, 2018; Guo et al., 2020).

On the other hand, various medical image synthesis tasks have significantly improved using GANs (Zhuang and Shen, 2016; Han et al., 2018; Yang et al., 2018; Sánchez and Vilaplana, 2018). GAN-based image translation techniques are closely related to image imputation since they can estimate the missing data by modelling the intrinsic manifold of the image data. One successful application involves casting the image imputation problem as a cross-domain images-to-image translation task so that a GAN network can estimate the missing data using the other available datasets, for instance, generating MR images from the other contrast inputs (Yurt et al., 2019; Lee et al., 2019; Dar et al., 2019). (Xia et al., 2020) proposed a conditional GAN to learn key features of SAX slices near the missing slice, and used them as conditioning variables to infer missing slices in the query volumes. However, the work only focused on interpolation to mitigate the issue of missing intermediate slices in SAX image stacks, without addressing the problem of missing apical and/or basal slices, which in turn requires extrapolation to ensure complete heart coverage.

Our goal in this study is to detect and impute (or extrapolate) the missing apical and/or basal slices from single SAX stacks, such that the accuracy of standard quantitative analyses conducted subsequently, is retained. There are several challenging aspects to this task: First, large inter-slice spacing (typically ranging from 8 to 10 mm) and variations in anatomical structures across CMR slices, pose a significant challenge to any data imputation and completion approaches. Second, there exist cases where multiple slices are missing in the cardiac apical and basal area of a single stack. Hence, image synthesis errors tend to accumulate when imputing the outermost slice(s). Finally, cardiac images exhibit a large degree of variability due to cardiac and respiratory motion, compared with other anatomical regions such as the brain.

In this paper, we propose an automated, learning-based pipeline to resolve the problem of sub-optimal heart coverage, that is non-negligible in CMR population studies. To the best of our knowledge, this is the first study to tackle the problem of automatic coverage detection and data completion. The main contributions of our approach are:

(1) A cascaded conditional GAN architecture is proposed to accurately synthesise missing apical and basal slices for CMR images. Particularly, we consider a challenging multi-slice scenario, where more than one slice may be missing in the apical and basal re-

gions. Conditioned on features extracted image slices adjacent to the detected missing slice, the designed network is able to impute missing slices in an anatomically plausible manner, consistent with the acquired image slices. Additionally, we demonstrate that using our approach, the accuracy of subsequent quantitative analyses can be improved (relative to using incomplete SAX image stacks).

(2) As a practical consideration, we introduce an effective detection strategy for incomplete coverage identification for a given CMR volume stack. The model employs several dense blocks comprising convolutional long short-term memory (ConvLSTM) layers, to leverage the rich discriminative capacity of features learned by considering the sequential change in cardiac anatomy observed across multiple adjacent slices, and achieve reliable classification outcomes. With this detection model, a two-stage data completion pipeline can be automatically deployed on large-scale population CMR imaging studies, such as UKBB.

(3) We comprehensively assess the proposed method across a large cohort of subjects. We systemically compare the differences between quantitative cardiac measurements calculated from the native CMR images and the imputed ones, showing the accuracy and robustness of the proposed approach in scenarios with multiple missing slices. Following analysis of 37,000+ UKBB subjects, we demonstrate that the pipeline can be used to compensate for the incomplete heart coverage in CMR population imaging.

The paper is organised as follows. Section 2 introduces the proposed pipeline, key model components and the learning algorithm. Section 2 describes the experiments conducted to validate the proposed models, and Section 4 presents the results of qualitative and quantitative analyses conducted to evaluate model performance. The essential characteristics of our model and its relevance in real clinical scenarios is then discussed in Section 5, before providing concluding remarks in Section 6.

2. Method

In this section, we present an image imputation network that leads to full cardiac coverage. The pipeline consists of two stages: incomplete coverage detection and missing slice imputation. For the detection, we integrated the ConvLSTM module into a densely connected convolutional network to encode both the global sequential through-plane spatial information and local spatial information of the input image stacks. We present the ConvLSTM and the integrated detection model in Section 2.1 and 2.2, respectively. To synthesise visually appealing cardiac MR images and facilitate accurate subsequent measurements, we present a GAN-based imputation framework with a dedicated generator and discriminator in Section 2.3-2.5. The generator contains residual blocks, where all normalisation layers are conditioned and modulated with input images to ensure that relevant texture details are effectively propagated through the network. A multi-scale discriminator was employed to ensure recovery of both global and local spatial features. We also propose a cascaded architecture stacking on multiple networks to handle cumulative errors when imputing multiple slices in Section 2.6 and present the final combined objective in Section 2.7.

The overall workflow comprises the following steps: First, three consecutive slices (cropped to a cardiac ROI) are extracted at both the apex and base and fed to the coverage detection model to identify whether the apical slice and basal slice are present in the volume. If not, the generative model takes the full-sized 3-slice stack as conditioning input and recursively synthesises a realistic CMR cine slice at the corresponding position until both the apex and base are detected in the newly imputed volume. The major notations used in the paper are summerised in Table 1. The overall workflow is illustrated in Fig. 2.

2.1. Convolutional long short-term memory

Although the success of recurrent Long Short-Term Memory (LSTM) networks, applied to sequence modelling (such as natural language processing) and scene labelling (Sundermeyer et al., 2012; Byeon et al., 2015) tasks has been demonstrated, the inputs to a standard LSTM network is vectorised and encoded through fully connected (FC) layers. This leads to the loss in spatial contextual information and equivariance to scaling and translation, which are essential when dealing with images in visual perception tasks. To address this problem, ConvLSTM (Shi et al., 2015) was proposed to retain relevant spatial information by replacing the FC layers with convolutional layers. The rationale for employing ConvLSTM in this work is to use the convolution and recurrence operations in the input-to-state and state-to-state transitions to leverage in-plane/through-plane spatial correlation information and sequential ordering of slices in SAX cine-MR data, for identification of apical and basal slices.

Following the work of (Shi et al., 2015), the inputs x_z, cell outputs c_z, hidden states h_z, input gate i_z, forget gate f_z, cell gate g_z, and output gate o_z are 3D tensors whose first two dimensions are rows and columns. In our case, the third dimension of the input tensors indicates the slices along the z-dimension (i.e. along the longitudinal axis of the ventricles), instead of temporal acquisitions/data as in (Shi et al., 2015). Let $*$ denote the convolution operator, and let $\circ$ denote the Hadamard product. Then, the key equations of the ConvLSTM cell can be formulated as:

$$
\begin{aligned}
i_z &= \sigma\left(\mathbf{W}_i * x_z + \mathbf{U}_i * h_{z-1} + \mathbf{R}_i \circ c_{z-1} + b_i\right)\\
f_z &= \sigma\left(\mathbf{W}_f * x_z + \mathbf{U}_f * h_{z-1} + \mathbf{R}_f \circ c_{z-1} + b_f\right)\\
g_z &= \tanh\left(\mathbf{W}_g * x_z + \mathbf{U}_g * h_{z-1} + b_g\right)\\
o_z &= \sigma\left(\mathbf{W}_o * x_z + \mathbf{U}_o * h_{z-1} + \mathbf{R}_o \circ c_{z-1} + b_o\right)\\
c_z &= f_z \circ c_{z-1} + i_z \circ g_z\\
h_z &= o_z \circ \tanh\left(c_z\right)
\end{aligned}
\tag{1}
$$

where σ denotes the sigmoid function, **W**, **U** and **R** denote the learnable 2D convolutional kernels, and b denotes the bias term.

The hidden states $h_0, h_1, \cdots h_{z-1}$ and the cell states $c_0, c_1, \cdots c_{z-1}$ are updated based on the input x_z (a stack of three consecutive slices in our case) passing through i_z, f_z and o_z gate activations during each step, as illustrated in Fig. 3. Each ConvLSTM cell encodes both the global sequential through-plane spatial information and local spatial information of the input x_z. All convolutions have a kernel size 3 × 3 with stride 1 × 1. Zero-padding is used to ensure that the output feature maps in each layer have the same spatial dimensions as its inputs. The next subsection presents an effective detection network comprising multiple dense ConvLSTM blocks.

2.2. Detection model with dense ConvLSTM block

We propose a dense ConvLSTM network for incomplete coverage classification. We use two independent but identical models to tackle apical and basal slice detection separately. The model adopts densely connected convolutional networks (DenseNets) (Huang et al., 2017) and ConvLSTM cells to exploit the spatial contextual information in volumetric data. The architecture is illustrated in Fig. 4. The input to this network is a stack of three consecutive, cropped SAX slices (139 × 139 × 3) containing the region of interest (ROI) at either the apex or the base. The trunk architecture consists of three dense blocks with each block containing two ConvLSTM layers, where the original inputs from all preceding layers are concatenated to the output in a feed-forward fashion to strengthen feature propagation and encourage feature reuse. Max-pooling layers are used between two adjacent blocks to change the

Table 1
Major mathematical notations used in the paper.

Notations	Definition
x_z	inputs to the detection model with a slice index z
c_z	cell outputs from a ConvLSTM cell
h_z	hidden states of a ConvLSTM cell
i_z, f_z, g_z, o_z	input gate, forget gate, cell gate and and output gate of a ConvLSTM cell
$\mathbf{W}$, $\mathbf{U}$, $\mathbf{R}$	learnable 2D convolutional kernels
s_i	i-th scalar value in the detection model output
t_i	i-th ground truth label
$n_{b,w,h,c}$	input feature with indices of batch size b, width w, height h, and channel c
$\gamma(\cdot)$, $\beta(\cdot)$	spatial dimension-dependent scale and shift functions
μ_c, σ_c^2	zero mean and unit standard deviation of a given feature batch c
$\mathbf{v}_i$	i-th input three-slice stack
$\mathbf{w}_i$	i-th ground truth slice
$G_i(\mathbf{v}_i)$	i-th generator with i-th input stack
D_k^j	k-th discriminator with j-th feature layer
V^j	VGG net with j-th feature layer
λ_1, λ_2	weighting of the feature matching loss and perceptual loss

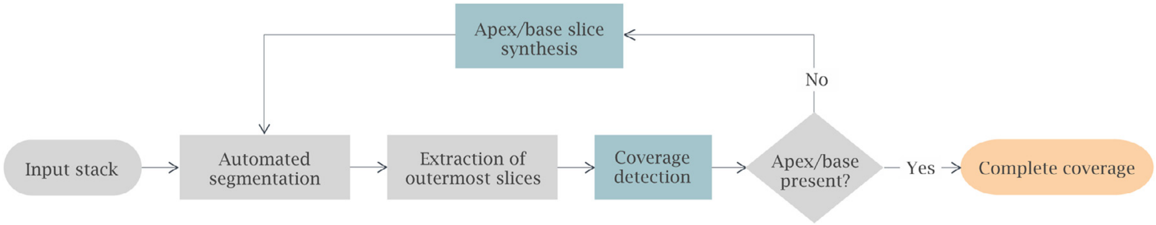

Fig. 2. The flowchart of the proposed two-stage pipeline that can be automatically applied to large-scale CMR data.

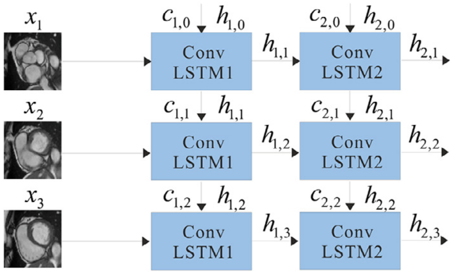

Fig. 3. The architecture of the ConvLSTM network. The hidden states $h_0, h_1, \cdots h_{z-1}$ and the cell states $c_0, c_1, \cdots c_{z-1}$ are updated based on the input (a three-slice stack in our case) passing through i_z, f_z and o_z gate activations during each step. The model utilises the ConvLSTM cells to exploit the spatial contextual information and sequential ordering of slices in SAX cine-MR data.

feature map sizes. We then configure one 3D convolutional layer with (kernel size = 1 × 1 × 1 and stride = 1 × 1 × 1) and four fully-connected layers to extract higher-level features from 2D feature maps learned in a recurrent fashion by the ConvLSTM layers, and learn to predict the final classification result.

As we formulate the apical and basal slice detection separately as two binary classification tasks, we use the binary cross-entropy loss for each detection model. Cross-entropy is the loss function under the inference framework of maximum likelihood and calculates a score that summarises the average difference between the actual and predicted probability distributions for predicting class:

$$BCE = -\frac{1}{N}\sum_{i=1}^{N} t_i \log(s_i) - (1-t_i)\log(1-s_i), \tag{2}$$

where s_i is the i-th scalar value in the model output, t_i is the corresponding ground truth label and N is the output size.

2.3. Image-Conditional GANs

Our data completion model is based on an image-to-image translation architecture (Isola et al., 2017) that learns a mapping from statistically dependent source images y to target images x. Such methods, regarded as a type of image-conditional GAN, can be adapted to address image imputation problems since they can estimate the missing data by modelling the intrinsic manifold of the image data distribution. The generator G and the discrimina-

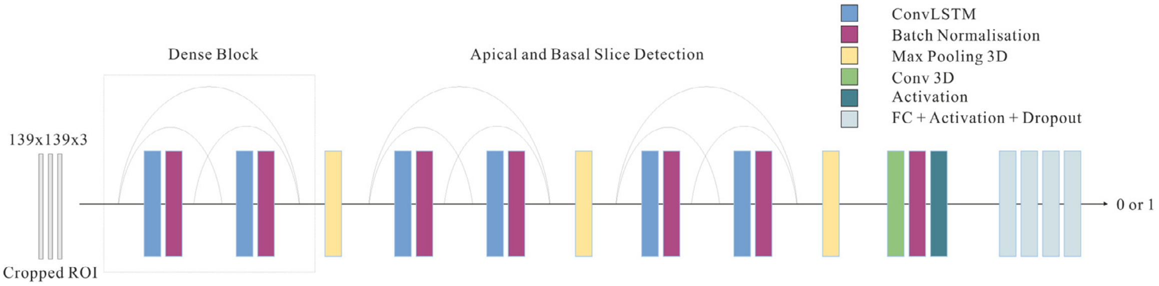

Fig. 4. Structure of the proposed detection network for slice classification. The trunk consists of three dense blocks with each block containing two ConvLSTM layers, where the original inputs from all preceding layers are concatenated to the output in a feed-forward fashion.

tor D are trained simultaneously and adversarially, where G aims to generate images that can fool the discriminator D, while D tries to classify counterfeit images optimally. To improve network stability during training, following the work in Mao et al. (2017) we replaced the negative log-likelihood in the native GAN with a least squared loss function. The optimisation of G and D can be formulated as:

$$\min_G \max_D \mathcal{L}_{cGAN} = \min_G \max_D \left(-\mathbb{E}_{yx}\left[(D(y,x)-1)^2\right] - \mathbb{E}_y\left[D(y,G(y))^2\right]\right). \tag{3}$$

It has been shown an effective strategy to integrate the traditional pixel-wise loss (e.g., L1 or L2 loss measured between the ground truth and generated images) into the GAN objective function and encourage G to create plausible translations of the source image while boosting image generation performance:

$$\mathcal{L}_{L1} = \mathbb{E}_{yx}[\|x - G(y)\|_1]. \tag{4}$$

However, although impressive results were obtained for synthesising natural images, the native image-to-image translation models may be unstable and prone to failure for synthesising images that contain fine structural details and rich quantitative information, which is essential for medical images.

2.4. Encoder-Decoder generator and multi-scale discriminator

Our approach is inspired by recent GAN architectures and refines several key techniques to output cardiac MR volumes that yield anatomically plausible images and accurate clinical parameters of cardiac function, despite the presence of significant proportions of missing data.

The generator is an encoder-decoder model where the input is first passed through a series of layers that down-sample until a bottleneck layer, followed by up-sampling and decoding the bottleneck representation to the size of the output image. The encoder aims to extract the underlying feature information of the input images, while the decoder maps the underlying representation into the target images with the same size as the input. The encoder comprises five 2D convolutional layers (kernel size = 3 × 3, padding = 1 × 1 and stride = 2 × 2), batch normalisation and activation ReLu layers. The decoder follows a full pre-activation residual network architecture (i.e. BatchNorm-ReLu-convolution) (He et al., 2016) that consists of a series of the residual blocks, followed by nearest neighbour up-sampling layers. Each residual block contains two convolutional layers (kernel size = 3 × 3, padding = 1 × 1), where a learned residue of input is added to the output to ensure the characteristics of original features are retained.

For the discriminator, instead of using a deeper network that would increase the network capacity and cause overfitting, a multi-scale discriminator (Durugkar et al., 2017; Nguyen et al., 2017; Wang et al., 2018) is adopted to operate on different receptive fields, which simultaneously encourages the generator to synthesise globally coherent images and capture fine structural details. Each discriminator uses modules of the form convolution-BatchNorm-LeakyReLu.

2.5. 3D Conditional batch normalisation

To ensure visual properties and relevant texture details are effectively propagated through the decoder, we employ conditional batch normalisation (CBN), which has been adopted in several previous studies (De Vries et al., 2017; Miyato and Koyama, 2018; Zhang et al., 2018a; Chen et al., 2019; Park et al., 2019; Xia et al., 2020). CBN employs a new conditioning strategy to incorporate external conditioning information (such as labels, embedding, masks or input noise vectors) into the image generation pathway through batch normalisation. It is implemented as a learning-based affine transformation over modulated features with parameters inferred from auxiliary data. In image synthesis, CBN enables an image to be translated from one domain into another while consistently respecting the constraints specified by conditioning data. In this work, we used three adjacent slices (i.e., the network's own input data) as inputs/conditioning information to the CBN module, to capture spatial features and fine structural details that are most relevant to the missing slices, which contributes significantly to our missing slice imputation task (as demonstrated later on in an ablation study in Section 4.5). Specifically, in each CBN layer, we first normalise the extracted features to zero mean and unit standard deviation. Then, the normalised features are modulated/denormalised using the affine transformation whose scale and shift parameters are learned from neighbouring slices using a CNN network. Mathematically, in the batch normalisation setting, input feature batch $n_{b,w,h,c} \in \mathbb{R}^{B\times W\times H\times C}$ ($b \in B, w \in W, h \in H,$ and $c \in C$ denote the batch size, width, height, and channel of the feature map, respectively) is normalised in a channel-wise manner:

$$n'_{b,w,h,c,} = \gamma_{w,h,c}(\mathbf{v}) \times \frac{n_{b,w,h,c} - \mu_c}{\sigma_c + \epsilon} + \beta_{w,h,c}(\mathbf{v}), \tag{5}$$

with

$$\mu_c = \frac{1}{N}\sum_{b,w,h} n_{b,w,h,c}, \quad \sigma_c^2 = \frac{1}{N}\sum_{b,w,h}\left(n_{b,w,h,c} - \mu_c\right)^2, \tag{6}$$

where $N = B \times W \times H$, ϵ is a small number to avoid division by zero, $\mathbf{v}$ denotes input 3-slice stack, $\gamma(\cdot)$ and $\beta(\cdot)$ are spatial dimension-dependent functions that can be formulated as CNN layers. The modulation parameters of all CBN layers within the generator were learned simultaneously through the GAN training. To leverage contextual information contained across slices, we propose to use 3D convolutional kernels (kernel size = 3 × 3 × 3, padding = 1 × 1 × 1 and stride = 1 × 1 × 1) to learn γ and β, and seek to the superior performance. The structure of the proposed CBN is illustrated in Fig. 5. The input to the CBN is a three-slice stack of dimensions 256 × 256 × 3, while the feature maps to be modulated have different dimensions resulting from the up-sampling layers of the decoder. Hence, we downsampled the input image stack to different levels such that the feature-wise scale shift parameters can be directly applied to the latter.

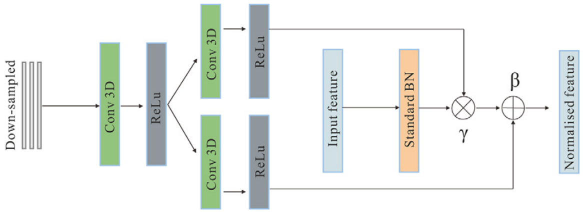

Fig. 5. Structure of the proposed 3D CBN module: the extracted features are first normalised to zero mean and unit standard deviation. Then, the normalised features are modulated using the affine transformation whose scale and shift parameters (i.e., γ and β) are learned using 3D convolutional layers.

2.6. Cascaded generative model for multi-slice imputation

Rather than simultaneously synthesising two or more slices that is unreliable and increases the estimation error, we address the multi-slice imputation problem using a cascaded architecture stacking on multiple networks, to generate each new slice in a recursive fashion. As subsequent networks learn to tolerate and handle synthesised images as inputs, the network can compensate for the error accumulation in the generation of multiple missing slices. Fig. 6 shows the structure of the proposed architec-

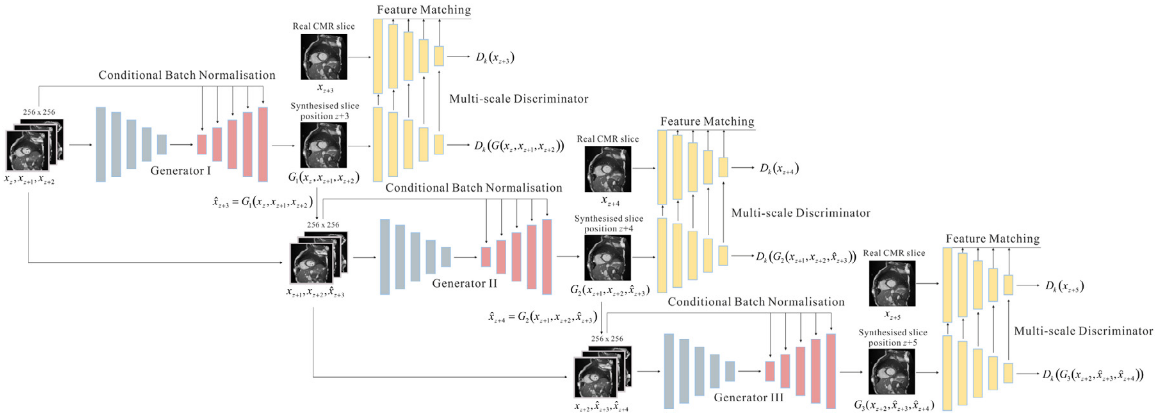
Feature Matching
Real CMR slice
x_{z+3}
$D_k(x_{z+3})$
Conditional Batch Normalisation
Multi-scale Discriminator
256 x 256
Synthesised slice position z+3
$D_k(G(x_z, x_{z+1}, x_{z+2}))$
Generator I
x_z, x_{z+1}, x_{z+2}
$G_1(x_z, x_{z+1}, x_{z+2})$
$\hat{x}_{z+3} = G_1(x_z, x_{z+1}, x_{z+2})$
Feature Matching
Real CMR slice
x_{z+4}
$D_k(x_{z+4})$
Conditional Batch Normalisation
Multi-scale Discriminator
256 x 256
Synthesised slice position z+4
$D_k(G_2(x_{z+1}, x_{z+2}, \hat{x}_{z+3}))$
Generator II
$x_{z+1}, x_{z+2}, \hat{x}_{z+3}$
$G_2(x_{z+1}, x_{z+2}, \hat{x}_{z+3})$
$\hat{x}_{z+4} = G_2(x_{z+1}, x_{z+2}, \hat{x}_{z+3})$
Feature Matching
Real CMR slice
x_{z+5}
$D_k(x_{z+5})$
Conditional Batch Normalisation
Multi-scale Discriminator
256 x 256
Synthesised slice position z+5
$D_k(G_3(x_{z+2}, \hat{x}_{z+3}, \hat{x}_{z+4}))$
Generator III
$x_{z+2}, \hat{x}_{z+3}, \hat{x}_{z+4}$
$G_3(x_{z+2}, \hat{x}_{z+3}, \hat{x}_{z+4})$

Fig. 6. Schematic view of the proposed cascaded architecture. The first network takes a 3-slice stack $[x_z, x_{z+1}, x_{z+2}]$ as conditioning input to synthesise a CMR slice $G(x_z, x_{z+1}, x_{z+2})$ at the adjacent position $z+3$. Then, the initial input is re-stacked by progressively discarding the innermost slice x_{z-1+i} and appending the newly synthesised slice $\hat{x}_{z+2+i}$ and construct the new input $\mathbf{v}_i$ to the subsequent networks. The model aims to generate the ith extended slice using the ith network in the stacked architecture, i.e., $\hat{x}_{z+2+i} = G_i(\mathbf{v}_i)$, until both the apex and base are detected by the classification network.

ture. Each network in the cascaded model consists of a generator and a multi-scale discriminator (operates at two scale levels). The model aims to generate the ith extended slice using the ith network G_i in the stacked architecture, i.e., $\hat{x}_{z+2+i} = G_i(\mathbf{v}_i)$, until both the apex and base are detected in the imputed volume by the classification network. The first generator takes three outermost slices $\mathbf{v}_1 = [x_z, x_{z+1}, x_{z+2}]$ in the original incomplete volumes as input and synthesises a missing slice $\hat{x}_{z+3} = G_1(\mathbf{v}_1)$ at the position $z+3$. The second and third generators are trained on the updated three-slice stacks with newly synthesised slices inserted. As the cascaded network is much bigger than a single network, the training of such a network is unstable and can encounter over-fitting. To counteract this issue, our strategies in practice are: 1) the stacked networks are trained one after the other in a sequential manner; 2) the subsequent network is initialised with the weights of the previously trained network in the series and then fine-tuned with the new augmented three-slice stacks. By doing so, the training can avoid overfitting while retaining the benefit of slice imputation refinement.

2.7. Optimisation

We adapt the image-to-image translation framework described in Eq. (3) and remove the pixel-wise L1 loss that struggles with capturing high-frequency details. Instead, we employ a feature matching loss (Salimans et al., 2016) and a perceptual loss (Johnson et al., 2016) to optimise the proposed GAN. While the feature matching loss is based on the discriminator D and matches the statistics of feature representations in multiple intermediate layers of D, the perceptual loss measures perceptual differences in content and style between real and generated images on high-level features extracted from a pre-trained VGG-16 network V. Thus, our final joint objective for each GAN network (a generator and a multi-scale discriminator) in the cascaded model combines the adversarial loss, the feature matching loss, and the perceptual loss as:

$$\mathcal{L}_{Final} = \min_G \left(\left(\max_{D_1, D_2} \sum_{k=1,2} -\left(\mathbb{E}_{\mathbf{v}w}\left[(D_k(\mathbf{v}_i, w_i) - 1)^2 \right] + \mathbb{E}_v\left[D_k(\mathbf{v}_i, G(\mathbf{v}_i))^2 \right] \right) \right) + \lambda_1 \sum_{k=1,2} \mathbb{E}_{\mathbf{v}w} \sum_{j=1}^{F} \frac{1}{N_j} \left[\left\| D_k^j(\mathbf{v}_i, w_i) - D_k^j(\mathbf{v}_i, G(\mathbf{v}_i)) \right\|_1 \right] + \lambda_2 \mathbb{E}_{\mathbf{v}w} \sum_{j=1}^{P} \frac{1}{M_j} \left[\left\| V^j(w_i) - V^j(G(\mathbf{v}_i)) \right\|_1 \right] \right) \tag{7}$$

where i indicates the ith network, j indicates the jth layer features, N_j and F are the number of features in each layer and the total number of layers in D, respectively. M_j and P represent the number of features in the layer j and the total number of layers in V, respectively. k denotes the index of the discriminator operating at two different scales. $\lambda_1$and λ_2 are used to control the relative weighting of the feature matching loss and perceptual loss, to the adversarial loss. The conditioned G and D networks can be optimised by $\mathcal{L}_{Final}$ to infer the missing slices $w_1 = x_{z+3}$, $w_2 = x_{z+4}$ and $w_3 = x_{z+5}$ from the augmented input 3-slice stacks $\mathbf{v}_1 = [x_z, x_{z+1}, x_{z+2}]$, $\mathbf{v}_2 = \left[x_{z+1}, x_{z+2}, \hat{x}_{z+3}\right]$ and $\mathbf{v}_3 = \left[x_{z+2}, \hat{x}_{z+3}, \hat{x}_{z+4}\right]$, respectively.

3. Experimental setup

3.1. Datasets

Cardiac MR images from the UKBB were used to train and validate the proposed method. Images were acquired using a clinical wide bore 1.5T MR system (MAGNETOM Aera, Syngo Platform VD13A, Siemens Healthcare, Erlangen, Germany) equipped with an 18-channel anterior body surface coil (45 mT/m and 200 T/m/s gradient system). 2D cine b-SSFP SAX image stacks were acquired with the following acquisition protocol: in-plane spatial resolution 1.8 × 1.8 mm, slice thickness 8 mm, slice gap 2 mm, image size 198 × 208. The number of slices in the SAX stack typically ranges between 10 and 12. Each slice was acquired at 50 cardiac phases. Further acquisition details can be found in (Petersen et al., 2015).

For training and testing the proposed coverage detection and slice imputation networks, we used visually quality-controlled UKBB CMR data from a previous study (Carapella et al., 2016; Petersen et al., 2017). Quality assessment was carried out by cardiologists through visual inspection using a three-grade quality score system (1 = optimal, 2 = sub-optimal, 3= unreliable) for each aspect of image quality such as image plane orientation, coverage, data consistency and artefacts[1]. We constructed the ground-truth data using 4102 SAX image sequences with complete heart coverage labelled from a spreadsheet by experienced cardiologists, i.e., both apex and base slices are present.

[1] Quality assessment can be accessed from the UK Biobank Resource under the returned data #2541 by application. http://www.ukbiobank.ac.uk/registerapply/

Also, we evaluate the generalisation of the proposed method on the publicly accessible, Kaggle Second Annual Data Science Bowl (ADSB) dataset[2]. The ADSB testing dataset contains 200 subjects and was compiled by the National Institutes of Health and Children's National Medical Center. The slice thickness ranges from 6 mm to 8 mm, and the in-plane spatial resolution varies from 0.61 to 1.75 mm. Each subject contains 30 cardiac phases over the heart cycle (scanned from the end-diastolic (ED) phase) and the number of slices in the SAX stack typically ranges between 6 and 14.

3.2. Network training

We addressed the detection of the apical and basal slices separately. To create a training dataset for the detection models, we extracted the three outermost slices as positive samples for basal or apical slice detection. For negative samples, i.e. not containing the base/apex, we selected four adjacent three-slice stacks near the positive samples. We constructed the training set from images at two cardiac phases, i.e. ED and end-systolic (ES), with optimal image quality. With masks obtained from a 2D CNN-based segmentation method (Bai et al., 2018), we cropped all the input image stacks to triplets with the size of 139 × 139 × 3 to extract the ROI. In total, we have 12,301 slice triplets for the apical slice detection (5,401 positive samples and 6900 negative samples) and 13,446 triplets for basal slice detection (6,379 positive samples and 7067 negative samples). All samples were split into three subsets with a ratio of 6:2:2 for the training set, the validation set, and the test set, respectively.

The training of the detection networks was optimised using the Adam optimiser with the following hyperparameters: a learning rate of 1×10^{-4}, the first and the second momentum of the gradient estimate 0.9 and 0.999, dropout rate 0.5 after FC layers. Trainable weights were randomly initialised from a truncated normal distribution centred on 0. The models converged after 48 epochs and 35 epochs for the apex detection and base detection, respectively.

Similarly, we also treated the apical and basal slice imputation as two independent problems. We randomly chose 3602 subjects for training and validation, and then tested our models on the 500 subjects. For each of training subjects, three sets of 4 adjacent slices (with one slice shift for each set) from the top and bottom of the ED and ES volumes were extracted to form pairs of input slice stacks and ground-truth images, resulting in 43,224 slice quadruplets (21,612 for each of the apex and base regions). Also considering there exist natural variations in heart size over different patients, each of the three cascaded generators is not bound to a specific spatial location, as we trained each generator with "a wide range" of input samples extracted from different spatial positions from the same subject.

The input image stacks were constructed in the direction from the middle slice towards the base for the basal slice imputation network, and from the middle slice towards the apex for the apical slice imputation network. In the training of the second and third generators, we re-stacked the input slices by appending the inferred slice from the previous generator and dropping the innermost slice.

The cascaded networks were trained one after the other in a sequential manner. We trained the first network for 50 epochs. The subsequent networks were initialised with the weights of the previously trained network and fine-tuned for another 20 epochs. All stacked networks have a common architecture and were optimised using the same objective function shown in Eq. (7). The training adopts the Adam optimiser with an initial learning rate of 2×10^{-4}, for both the generator and discriminator. The decay rates of the first and the second momentum of the gradient estimate were set to 0.5 and 0.999, respectively. All the stacks were resized to 256 × 256. The relative weighting factors of the feature matching loss and the perceptual loss to the GAN loss were empirically set as $\lambda_1 = 10$ and $\lambda_2 = 10$.

3.3. Evaluation design

We conducted several experiments to assess the accuracy and robustness of the proposed methods. The first experiment was performed by evaluating the accuracy of the coverage detection models. To demonstrate the advantages of the proposed method, we compared it to 2D and 3D CNN-based classification models. The 2D model is a state-of-the-art Inception-v3 network (Szegedy et al., 2016) that uses label smoothing and factorising convolutions to improve network efficiency. The 2D model takes a single slice as input and predicts a probability that the slice corresponds to negative or positive apical/basal slice. The 3D CNN used employed a trunk structure similar to the proposed detection network but with all ConvLSTM layers replaced by 3D convolutional layers (kernel size = 3 × 3 × 3, padding = 1 × 1 × 1 and stride = 1 × 1 × 1). We also compared a state-of-the-art missing slice detection method proposed in Zhang et al. (2018b). Following the network design in Zhang et al. (2018b), we incorporated a fisher-discriminative (FD) fully connected layer in the competing 3D CNN network, resulting in a refined 3D CNN+FD network. We used the same training and testing approaches for these models. To quantitatively assess classification performance, we used the established metrics, i.e. specificity, sensitivity, accuracy and the area under the curve (AUC) for the receiver operating characteristic (ROC).

In the second experiment group, we validated the proposed image imputation networks on 500 subjects from quality-controlled UKBB CMR data. Three truncation levels were considered by systematically removing the 1–3 topmost/bottommost slice(s) from each complete image stack (i.e. with full coverage). The correlation coefficient (CC) and the peak signal-to-noise ratio (PSNR) were used to measure the image quality of the synthesised CMR slices, relative to the original slices. Furthermore, we also quantified the statistical differences of standard clinical cardiac measurements, e.g. the LV end-diastolic volume (LVEDV), LV stroke volume (LVSV) and LV ejection fraction (LVEF) etc., between the full and imputed cardiac image volumes. Measurements of these cardiac indices were derived from segmentation results using a state-of-the art CNN method (Bai et al., 2018).

To demonstrate the feasibility and impact of the proposed pipeline on analysing large-scale datasets, we retrospectively applied the proposed pipeline to 37,396 UKBB subjects. We assessed differences between cardiac volume and functional indices extracted from actually acquired and imputed CMR volumes. Bland-Altman analyses were used to evaluate the correlation and agreement between these cardiac clinical measurements.

4. Results

4.1. Detection model evaluation

To assess the performance of the classification models, we performed ROC analysis on the 2460 testing apical images and 2689 basal images. The results are shown in Fig. 7. The proposed detection network yields better area under the ROC curve (AUC) for both the apical and the basal slice detection (94.84% and 95.88%), compared to the 2D CNN model (92.66% and 92.96%), 3D CNN model (93.44% and 95.13%) and 3D CNN+FD model (93.59% and 95.75%), which means it has a better measure of separability.

[2] https://www.kaggle.com/c/second-annual-data-science-bowl/

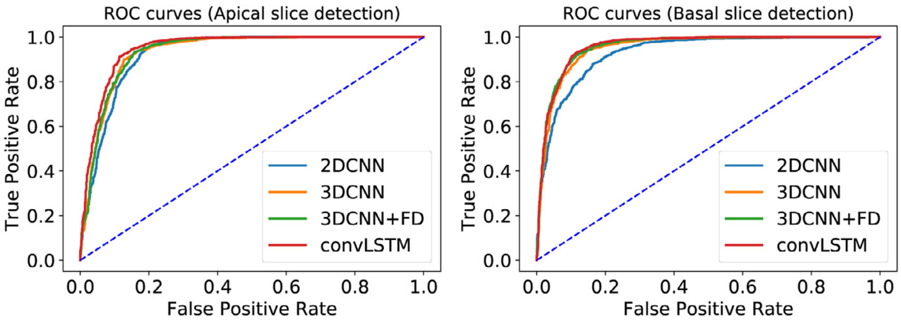

Fig. 7. ROC curves between different learning models for apical and basal slice detection.

Table 2
Comparison of classification results between the proposed detection model (convLSTM), the 2D CNN model, 3D CNN and 3D CNN+FD model. The bold-face font highlights the largest/best value in each measure.

Method	Sensitivity (%)		Specificity (%)		Accuracy (%)		AUC (%)	
	Apical	Basal	Apical	Basal	Apical	Basal	Apical	Basal
2D CNN	83.34	87.89	86.68	83.53	85.28	85.60	92.66	92.96
3D CNN	90.11	82.85	87.17	**92.09**	88.45	87.76	93.44	95.13
3D CNN+FD	**93.65**	91.90	83.57	88.38	88.66	90.03	93.59	95.75
convLSTM	91.32	**92.77**	**88.47**	88.66	**89.71**	**90.59**	**94.84**	**95.88**

Table 3
Summary of the statistical significance using the McNemar' s test computed between the proposed detection model and the competing methods.

p-value	Apex			Base		
	2D CNN	3D CNN	3D CNN+FD	2D CNN	3D CNN	3D CNN+FD
conLSTM	0.001	0.002	0.038	0.001	0.001	0.112

The results of the binary classification test with the optimal thresholds returned by the G-mean analysis are reported in Table 2. For each threshold we calculate the corresponding G-mean from the sensitivity and the specificity. Once computed, we locate the index for the largest G-mean score and use this index to determine the optimal threshold. For the 2D CNN model (Inception-v3 (Szegedy et al., 2016)), the accuracy rates for the apical and basal detection are 85.28% and 85.60%, respectively. The detection performance improves to an accuracy of 88.45% and 87.76% with 3D convolutional kernels, by incorporating inter-slice contextual information into the model. The best accuracy rates are yielded by the proposed ConvLSTM model, with 89.71% and 90.59% for the apical slice/basal slice detection. We found that with a Fisher discriminant regulariser on the CNN features, the 3D CNN+FD model showed superior results over the baseline 3D CNN model by improving the discriminative power of learned features, reflected by the better area under the ROC curve (AUC) i.e., 93.59% vs. 93.44% and 95.75% vs. 95.13% for both the apical and the basal slice detection. The 3D CNN+FD model yields the best sensitivity (93.65%) amongst all investigated detection models. However, its AUC values are worse than the proposed convLSTM model. This demonstrates the effectiveness of using the convolution and recurrence operations to leverage in-plane and through-plane spatial correlation information by considering the sequential ordering of slices as a rich descriptor of cardiac anatomy. While the ConvLSTM model reaches the best specificity rate of 88.47% for apical slice detection, its rate for base classification is 88.66%, worse than that of 92.09% from the 3D CNN model. Statistical significance is presented in Table 3 by performing the McNemar' s test computed between the detection model and the competing methods. We can see that the proposed approach achieves statistically significant improvements over the 2D CNN and 3D CNN models, considering a significance level of p<0.05. Although the proposed approach also yields statistically significant over 3D CNN+FD model for apex detection (p = 0.038), the p-value is larger than the chosen significance level for basal slice detection.

4.2. Imputation model evaluation

A visual comparison of the synthesised slices by the proposed imputation network and the ground-truth images for 7 representative subjects of the UKBB dataset at the ED phase is depicted in Fig. 8 (similar results obtained from the ES phase are shown in the appendix). From top to bottom, results of three degrees of incomplete coverage were shown. The qualitative examples are randomly selected on a pool of samples with performances around the mean CC generated. Subjects 1–4 represent the cases where one slice from either the cardiac apex or base was removed to generate sub-optimal coverage volumes, before using our method to impute the missing slice. The proposed method generates anatomically plausible results in terms of preserving fine structural details and realistic textures by learning about these structures from the neighbouring slice features. The inferred slices are visually comparable to the reference CMR ones; see LV blood pool, RV, right atrium (RA), and aorta in the synthesised slices. Additionally, the method can maintain high image quality when inferring two consecutive missing slices, as shown in Subjects 5 and 6 in Fig. 8. We observed some differences in the background tissues between the second imputed slice (denoted as Ex2) and the corresponding ground-truth image. This is expected as the second imputed slices are further away (20 mm spatial distance) from the seen slices due to relatively large through-plane spacing in CMR data. Synthesising three topmost or bottommost slices in the image volumes is more challenging and results in fewer similarities in the background, for the third imputed slice (denoted as Ex3) compared with the reference. The visual inspection of detailed structures also revealed minor degradation in image quality and the presence of texture artefacts.

We also visually assess the quality of the imputed slices in the through-plane direction, i.e. the long-axis (LAX) view for 6 subjects in Fig. 9. Each SAX stack is overlaid on top of the LAX 2-chamber view images (LAX 2CH) provided. As shown in Fig. 9, SAX stacks missing 1–3 slice(s) (in either the basal or apical direction) result

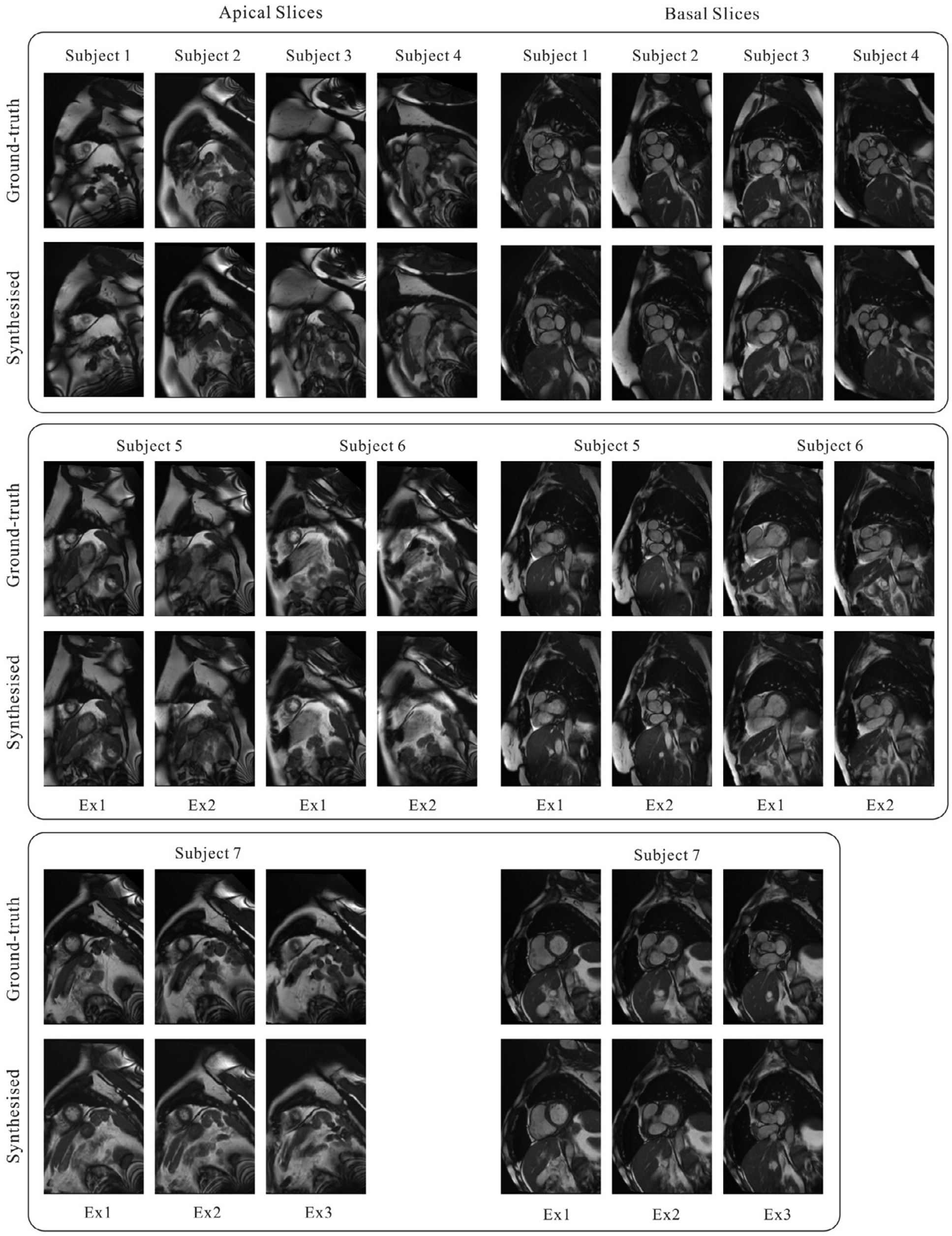

Fig. 8. Qualitative comparison of the ground truth and the synthesised slices for 7 subjects of the UKBB dataset at the ED phase (similar results obtained at the ES phase are shown in the appendix). From top to bottom, results of three degrees of incomplete coverage are shown. Ex1, Ex2 and Ex3 represent the first, the second and the third extended slice, respectively. The CMR images were reproduced with the permission of UK Biobank.

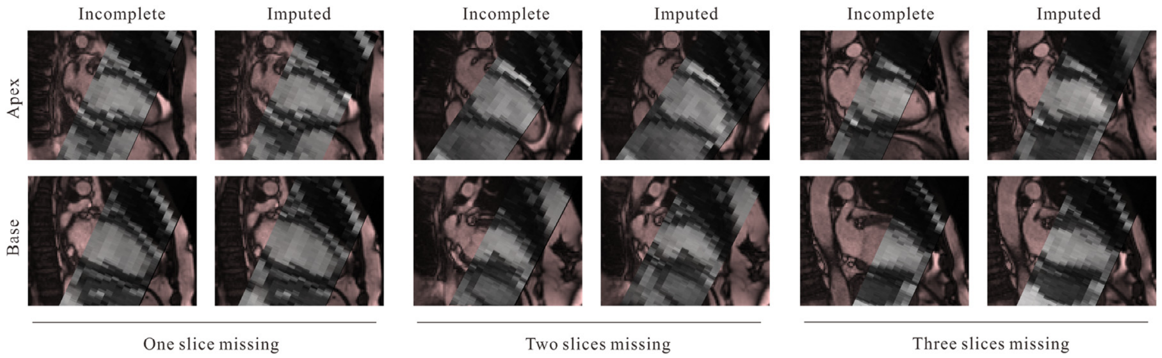

Fig. 9. Visual comparison of incomplete coverage stacks and imputed stacks on the LAX view for 6 representative subjects. Each SAX stack is superimposed on top of the LAX 2CH view for better visualisation. From left to right, results of three degrees of incomplete coverage are shown. The CMR images were reproduced with the permission of UK Biobank.

Table 4
Summary of the CC and PSNR measurements between the ground truth and synthesised images from the proposed imputation approach on 500 subjects at both the ED and ES phase. Ex1, Ex2 and Ex3 represent the first, the second and the third extended slice, respectively.

CC		One Slice Missing	Two Slices Missing		Three Slices Missing		
		Ex1	Ex1	Ex2	Ex1	Ex2	Ex3
Apex	ED	0.893 ± 0.031	0.894 ± 0.028	0.805 ± 0.052	0.895 ± 0.031	0.809 ± 0.045	0.740 ± 0.063
	ES	0.890 ± 0.032	0.891 ± 0.028	0.803 ± 0.051	0.892 ± 0.032	0.806 ± 0.043	0.737 ± 0.059
Base	ED	0.895 ± 0.035	0.896 ± 0.045	0.819 ± 0.058	0.906 ± 0.032	0.821 ± 0.061	0.761 ± 0.073
	ES	0.901 ± 0.033	0.895 ± 0.041	0.827 ± 0.057	0.897 ± 0.034	0.821 ± 0.058	0.771 ± 0.071
PSNR		One Slice Missing	Two Slices Missing		Three Slices Missing		
		Ex1	Ex1	Ex2	Ex1	Ex2	Ex3
Apex	ED	23.83 ± 1.74	23.86 ± 1.76	21.37 ± 1.83	23.99 ± 1.84	21.36 ± 1.63	20.15 ± 1.83
	ES	23.81 ± 1.88	23.79 ± 1.91	21.38 ± 1.95	23.87 ± 1.85	21.34 ± 1.71	20.17 ± 1.97
Base	ED	25.32 ± 1.63	25.09 ± 1.50	22.99 ± 1.52	25.15 ± 1.22	22.75 ± 1.41	21.79 ± 1.57
	ES	25.41 ± 1.59	25.04 ± 1.52	23.02 ± 1.44	24.97 ± 1.31	22.74 ± 1.43	21.81 ± 1.50

Table 5
Quantitative assessment of the segmentation accuracy by comparing the automated segmentation results between the reference slices and synthesised slices on 500 test sets. Ex1, Ex2 and Ex3 represent the first, the second and the third extended slice, respectively.

		Dice Score			Hausdorff Dist. [mm]		
		Ex1	Ex2	Ex3	Ex1	Ex2	Ex3
Apex	ED	0.904 ± 0.071	0.823 ± 0.096	0.697 ± 0.111	4.248 ± 2.329	5.365 ± 2.191	5.787 ± 2.529
	ES	0.847 ± 0.089	0.741 ± 0.105	0.668 ± 0.103	4.435 ± 2.059	5.655 ± 2.296	6.818 ± 2.557
Base	ED	0.923 ± 0.050	0.895 ± 0.062	0.838 ± 0.092	4.708 ± 1.268	6.428 ±2.725	9.036 ± 4.767
	ES	0.862 ± 0.073	0.816 ± 0.071	0.801 ± 0.077	5.807 ± 2.551	7.818 ± 2.895	8.391 ± 3.248

in incomplete LV coverage, whereas the proposed method compensates for this by accurately synthesising the missing apical or basal slices, in a manner consistent with the original data acquired.

Mean and standard deviation of image quality measurements on 500 subjects at both the ED and ES phase are presented in Table 4. Focusing on results of the ED phase first, the slice imputation method yields the best results when one slice is missing from the stack, which is reflected by the computed CC of 0.893 ± 0.031 and 0.895 ± 0.035 for the apex and base region, respectively. These metrics values decrease to 0.740 ± 0.063 and 0.761 ± 0.073 when generating the third missing slices. While we observed that high image quality is retained, the lower CC and PSNR values are mainly caused by the dissimilarity in the background between the imputed and acquired images. Consistent with the visual inspection, we found that the model can synthesise better images in the basal direction than those in the apical direction, due to large variation in the appearance of the apex slices. This is convenient for cardiac quantification as the absence of the basal slices has a larger impact on volume calculation. Overall, there are no significant differences in the quality of synthesised slices between the ED and ES phase.

To assess the anatomical plausibility of synthesised images, we also computed the segmentation results based on the synthesised images and compared them with those derived from the ground truth images. Fig. 10 shows the automated segmentation of the LV on the synthesised slices, illustrating retained accuracy of segmentation results for those synthesised slices with the proposed method. Note that, the automatic segmentation model was trained on the real cardiac cine MR images but still can yield good segmentation results from the synthesised slices, demonstrating the synthesised slices are anatomically plausible and realistic. Table 5 summarises the quantitative assessment of the LV segmentation accuracy by comparing the automated segmented results between the reference and synthesised slices on 500 test sets. The Dice score and Hausdorff distance were used as the metrics. All values are shown as mean ± standard deviation. It is observed that in general the Dice scores of the segmentation results for the apex are inferior to those of the base. This is because considerable variability in the shape of the LV near the apex in anisotropic 2D cine images, which leads to the difficulty in ensuring highly accurate and precise automatic segmentation of the LV in this region. We also found the degradation in segmentation accuracy when synthesising more slices, as expected. Next, we computed the clinical indices to evaluate the impact of these segmentation errors.

To assess the impact of the proposed imputation model, we compared the clinical cardiac volumetric and functional parameters derived from full, incomplete, and imputed image stacks. In total, there were 9 cardiac indices investigated, including the LV end-diastolic volume (LVEDV), end-systolic volume (LVESV), LV stroke volume (LVSV), LV ejection fraction (LVEF), LV myocardial mass (LVM), RV end-diastolic volume (RVEDV) and end-systolic volume (RVESV), RV stroke volume (RVSV) and RV ejection fraction (RVEF) computed from 500 subjects.

The mean and standard deviation of those measurements, categorised by the three sub-optimal coverage levels, are presented in Table 6, 7 and 8, along with an analysis of the mean absolute differences between the incomplete, imputed CMR images, and reference, as shown in Fig. 11. As expected, the largest impact on heart coverage is caused by the absence of basal slices. For instance, missing one slice, two slices and three slices reduces the LVEDV by an average of 11.8 mL, 32.7 mL and 54.7 mL, respectively, indicating the progressive under-estimation of the clinical parameter. As a result, these differences cause a dramatic decrease in the computed functional parameters such as LVSV and LVEV, accordingly. The proposed method can compensate for the insufficient coverage and yields a good agreement on all clinical indices with the reference, e.g. with the mean differences of only -0.7 mL, 0.1 mL and 2.6 mL in LVEDV for one to three missing slice scenarios, respectively. Statistical significance of the results was verified by performing the Student's *t*-test between the derived cardiac indices from different methods. In Table 8, the p-values for the RVEF and RVSV for the base are lower than 0.05, when comparing the proposed method to the reference value, as missing three basal

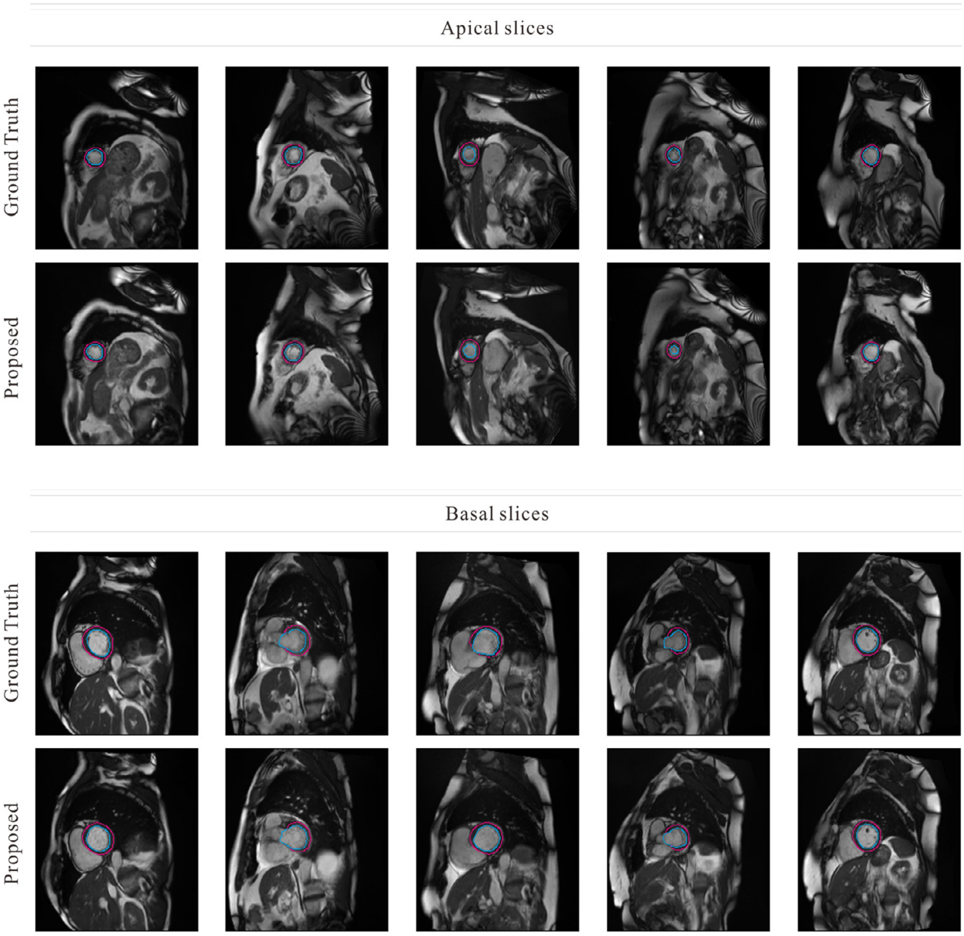

Fig. 10. Examples of the segmentation results of the LV to illustrate the accuracy of the automated segmentation on the synthesised slices with the proposed method. The CMR images were reproduced with the permission of UK Biobank.

Table 6
Mean and standard deviation of the clinical cardiac indices computed from 500 UKBB subjects, between the complete, incomplete, and compensated images. Here, one slice in either the apical or basal region is missing.

	Complete	Apex				Base			
		Incomplete		Compensated		Incomplete		Compensated	
Parameters	Mean ± Std.	Mean ± Std.	p-value	Mean ± Std.	p-value	Mean ± Std.	p-value	Mean ± Std.	p-value
LVEDV (mL)	143.6 ± 34.5	140.4 ± 34.2	0.29	143.6 ± 34.6	**0.99**	131.8 ± 34.6	<0.001	144.3 ± 35.0	**0.72**
LVESV (mL)	59.9 ± 21.2	59.1 ± 20.8	0.66	59.9 ± 21.3	**0.97**	59.2 ± 21.5	0.51	60.2 ± 21.3	**0.87**
LVSV (mL)	83.6 ± 18.9	81.2 ± 18.8	0.14	83.5 ± 18.9	**0.96**	72.7 ± 17.8	<0.001	84.1 ± 19.2	**0.61**
LVEF (%)	58.8 ± 6.69	58.4 ± 6.71	0.47	58.7 ± 6.71	**0.95**	55.6 ± 7.09	<0.001	58.9 ± 6.79	**0.80**
LVM (g)	83.6 ± 20.8	80.6 ± 20.3	0.11	83.6 ± 20.8	**0.96**	78.0 ± 20.3	0.002	84.0 ± 20.9	**0.77**
RVEDV (mL)	152.4 ± 36.8	149.6 ± 36.1	0.39	152.2 ± 36.8	**0.95**	144.3 ± 37.5	0.01	153.4 ± 36.9	**0.76**
RVESV (mL)	66.7 ± 22.3	66.1 ± 21.9	0.73	66.7 ± 22.4	**0.99**	66.3 ± 22.4	0.85	66.8 ± 22.3	**0.98**
RVSV (mL)	85.7 ± 19.2	83.5 ± 18.9	0.21	85.5 ± 19.2	**0.93**	78.1 ± 20.2	<0.001	86.6 ± 19.6	**0.58**
RVEF (%)	56.8 ± 6.53	56.4 ± 6.61	0.50	56.7 ± 6.58	**0.95**	54.4 ± 7.07	<0.001	56.9 ± 6.67	**0.73**

slices is a challenging case. Also, the third imputed slices are further away (30 mm spatial distance) from the seen slices due to the relatively large through-plane spacing in CMR data, but still have a significant impact on volume calculation (occupying a relatively large area). Although the p-values for the RVEF and RVSV indices are lower than 0.05, the proposed method still effectively reduces the relative error from 58.5% to 5.6% for RVEF and 30.2% to % 2.99% for RVSV. The analysis of the mean absolute differences between different measurements (cf. Fig. 11) also highlights the robustness of the cascaded model and demonstrates the reliability and accuracy of imputed images even in the presence of multiple missing slices.

We also evaluated the regional myocardial wall thickness, which is often used as a biomarker for quantifying regional dysfunction. We analysed the mean differences of myocardial wall thickness between the imputed and complete image stacks at both the cardiac ED and ES phase. Fig. 12 shows the analysis results in the bulls-eye plot based on the AHA 17-segment model. It can be seen that the imputation method yields relatively small differences, with respect to the wall thickness derived from the complete data, for both the apical (segments 13–17) and basal region (segments 1–6). The biggest difference is observed for the apical slices at the ES phase when three slices are missing. This is because, due to considerable variability in the shape of the my-

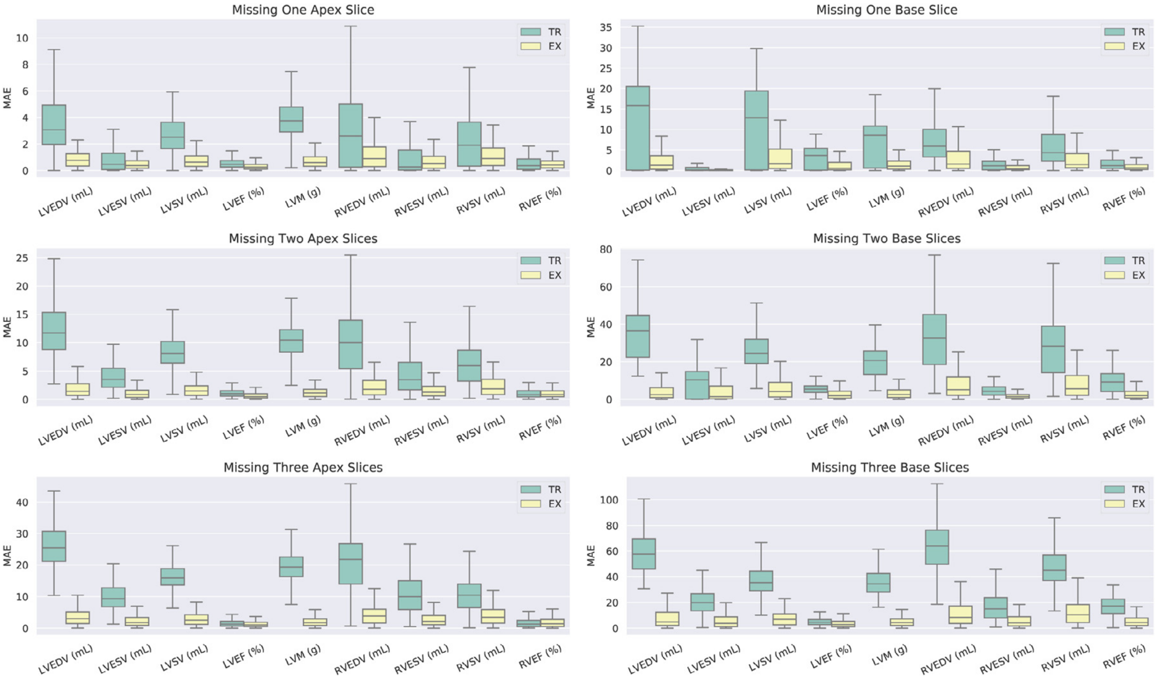

Fig. 11. Box plots of the absolute differences between the incomplete images (denoted as TR), compensated images (denoted as EX) and the complete images, across different incomplete coverage levels.

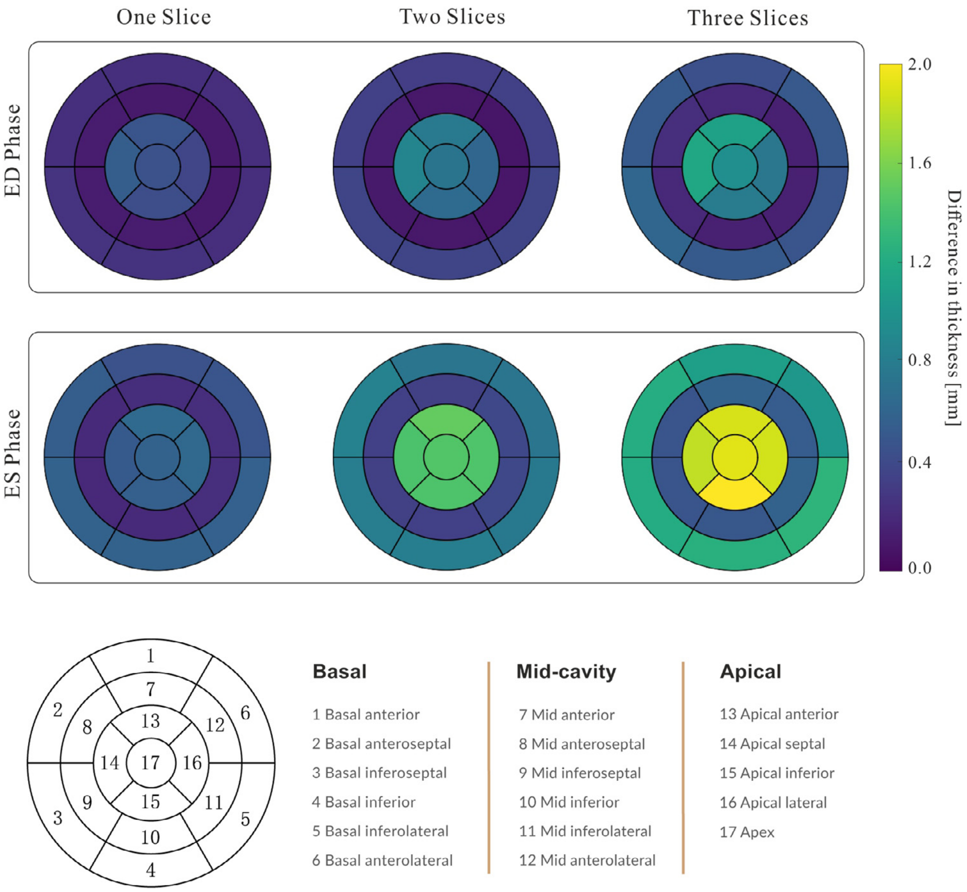

Fig. 12. Bulls-eye plots of the differences in regional wall thickness analysis (AHA 17-segment model) between the synthesised images and the reference. Results are shown at both the ED and ES phase across different incomplete coverage levels.

Table 7
Mean and standard deviation of the clinical cardiac indices computed from 500 UKBB subjects, between the complete, incomplete, and compensated images. Here, two consecutive slices in either the apical or basal region are missing.

	Complete	Apex				Base			
		Incomplete		Compensated		Incomplete		Compensated	
Parameters	Mean ± Std.	Mean ± Std.	p-value	Mean ± Std.	p-value	Mean ± Std.	p-value	Mean ± Std.	p-value
LVEDV (mL)	143.6 ± 34.5	132.1 ± 33.5	0.001	143.1 ± 34.7	**0.87**	110.9 ± 34.4	<0.001	143.5 ± 34.2	**0.99**
LVESV (mL)	59.9 ± 21.2	55.9 ± 19.8	0.03	59.5 ± 21.0	**0.81**	54.1 ± 21.2	0.002	61.2 ± 21.2	**0.48**
LVSV (mL)	83.6 ± 18.9	76.2 ± 18.7	<0.001	83.6 ± 19.3	**0.98**	56.8 ± 14.3	<0.001	82.3 ± 19.2	**0.45**
LVEF (%)	58.8 ± 6.69	58.1 ± 6.80	0.28	58.9 ± 6.63	**0.83**	52.1 ± 6.88	<0.001	57.8 ± 6.85	**0.10**
LVM (g)	83.6 ± 20.8	74.7 ± 19.6	<0.001	83.4 ± 20.8	**0.95**	66.4 ± 19.1	<0.001	83.4 ± 20.9	**0.94**
RVEDV (mL)	152.4 ± 36.8	142.6 ± 34.8	0.002	151.5 ± 36.9	**0.78**	121.1 ± 37.7	<0.001	152.3 ± 37.1	**0.98**
RVESV (mL)	66.7 ± 22.3	62.9 ± 21.0	0.05	66.5 ± 22.4	**0.90**	63.7 ± 22.6	0.13	67.2 ± 22.3	**0.80**
RVSV (mL)	85.7 ± 19.2	79.6 ± 18.4	<0.001	85.0 ± 19.5	**0.70**	57.4 ± 20.8	<0.001	85.1 ± 21.0	**0.75**
RVEF (%)	56.8 ± 6.53	56.3 ± 6.67	0.45	56.7 ± 6.71	**0.84**	47.2 ± 9.18	<0.001	56.2 ± 7.45	**0.38**

Table 8
Mean and standard deviation of the clinical cardiac indices computed from 500 UKBB subjects, between the complete, incomplete, and compensated images. Here, three consecutive slices in either the apical or basal region are missing.

	Complete	Apex				Base			
		Incomplete		Compensated		Incomplete		Compensated	
Parameters	Mean ± Std.	Mean ± Std.	p-value	Mean ± Std.	p-value	Mean ± Std.	p-value	Mean ± Std.	p-value
LVEDV (mL)	143.6 ± 34.5	118.9 ± 32.3	<0.001	141.7 ± 34.9	**0.54**	88.9 ± 28.6	<0.001	141.0 ± 34.4	**0.41**
LVESV (mL)	59.9 ± 21.2	49.6 ± 18.5	<0.001	58.3 ± 20.6	**0.37**	43.5 ± 18.8	<0.001	59.6 ± 21.8	**0.88**
LVSV (mL)	83.6 ± 18.9	69.2 ± 18.3	<0.001	83.4 ± 19.8	**0.90**	45.4 ± 12.7	<0.001	81.4 ± 18.6	**0.17**
LVEF (%)	58.8 ± 6.69	58.2 ± 7.32	0.32	59.3 ± 6.68	**0.37**	52.2 ± 7.20	<0.001	58.3 ± 7.21	**0.45**
LVM (g)	83.6 ± 20.8	66.7 ± 18.6	<0.001	82.5 ± 20.4	**0.57**	53.1 ± 16.9	<0.001	81.3 ± 20.2	**0.16**
RVEDV (mL)	152.4 ± 36.8	131.6 ± 33.5	<0.001	149.8 ± 36.3	**0.42**	90.2 ± 33.7	<0.001	147.5 ± 36.7	**0.14**
RVESV (mL)	66.7 ± 22.3	56.8 ± 19.9	<0.001	66.4 ± 22.1	**0.89**	54.8 ± 22.8	<0.001	66.5 ± 22.7	**0.93**
RVSV (mL)	85.7 ± 19.2	74.7 ± 17.9	<0.001	83.4 ± 19.2	**0.17**	35.5 ± 15.4	<0.001	80.9 ± 22.3	**0.01**
RVEF (%)	56.8 ± 6.53	57.4 ± 7.02	0.02	56.1 ± 6.82	**0.27**	39.6 ± 9.78	<0.001	55.1 ± 8.91	**0.02**

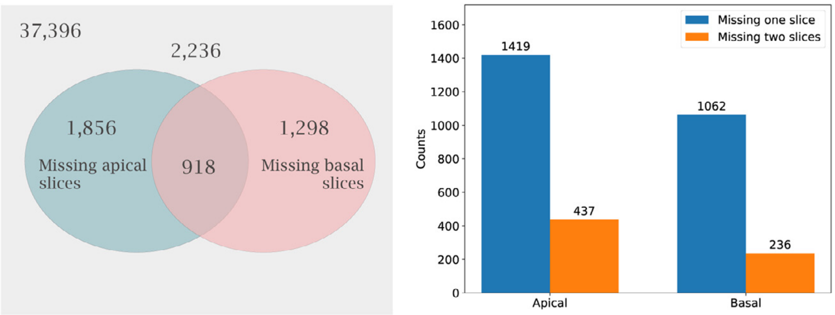

Fig. 13. Results obtained by analysing 37,396 UKBB datasets with the proposed automatic pipeline. There were 2236 subjects classified as a sub-optimal coverage in terms of missing at least one slice in either the basal or apical direction.

ocardium near the apex, automatic wall thickness assessment in the apical region can be error prone due to the difficulty in ensuring highly accurate and precise segmentation of the myocardium in this region, using anisotropic cine-CMR image stacks.

4.3. Large-Scale dataset analysis

By retrospectively applying the proposed pipeline to 37,396 UKBB subjects, we detected in total 2236 cases as a sub-optimal coverage in terms of missing at least one slice in either the basal or apical direction (i.e. ~5.9% of incomplete coverage rate), and compensated for those insufficient coverage volumes. The results are shown in Fig. 13, from which we can see that 1298 subject cases are missing basal slices (1,062 cases for one slice and 236 for two slices), and 1856 are missing apical slices (1,419 for one and 437 for two slices), while 918 cases are missing both in the same volume. The mean and standard deviation of clinical parameters are presented in Table 9 and the Bland-Altman analysis is shown in Fig. 14. Statistically significant differences were found between the original images and compensated ones regarding LVEDV, LVSV, LVEF, LVM, RVEDV, RVSV and RVEF (p<0.001). Note that the mean values of the clinical parameters (from these 2236 subjects) are overall larger than the reference mean values in Table 6, even

Table 9
Mean and standard deviation of the cardiac clinical indices computed from 2236 UKBB subjects, between the incomplete and compensated volumes. Note that the mean values of these clinical parameters are relatively larger than the reference mean values, as the ventricles with larger size tend to be insufficiently covered during the acquisition in real-world scenario.

	Incomplete	Compensated		
Parameters	Mean ± Std.	Mean ± Std.	MAE	p-value
LVEDV (mL)	152.5 ± 37.9	158.7 ± 37.6	6.15	<0.001
LVSV (mL)	87.8 ± 21.9	93.5 ± 21.5	5.65	<0.001
LVEF (%)	58.0 ± 7.17	59.2 ± 7.04	1.40	<0.001
LVM (g)	90.2 ± 24.6	94.0 ± 24.6	3.83	<0.001
RVEDV (mL)	161.2 ± 40.5	166.9 ± 40.2	5.73	<0.001
RVSV (mL)	89.0 ± 22.3	93.7 ± 22.1	4.79	<0.001
RVEF (%)	55.9 ± 7.29	56.7 ± 6.97	1.07	<0.001

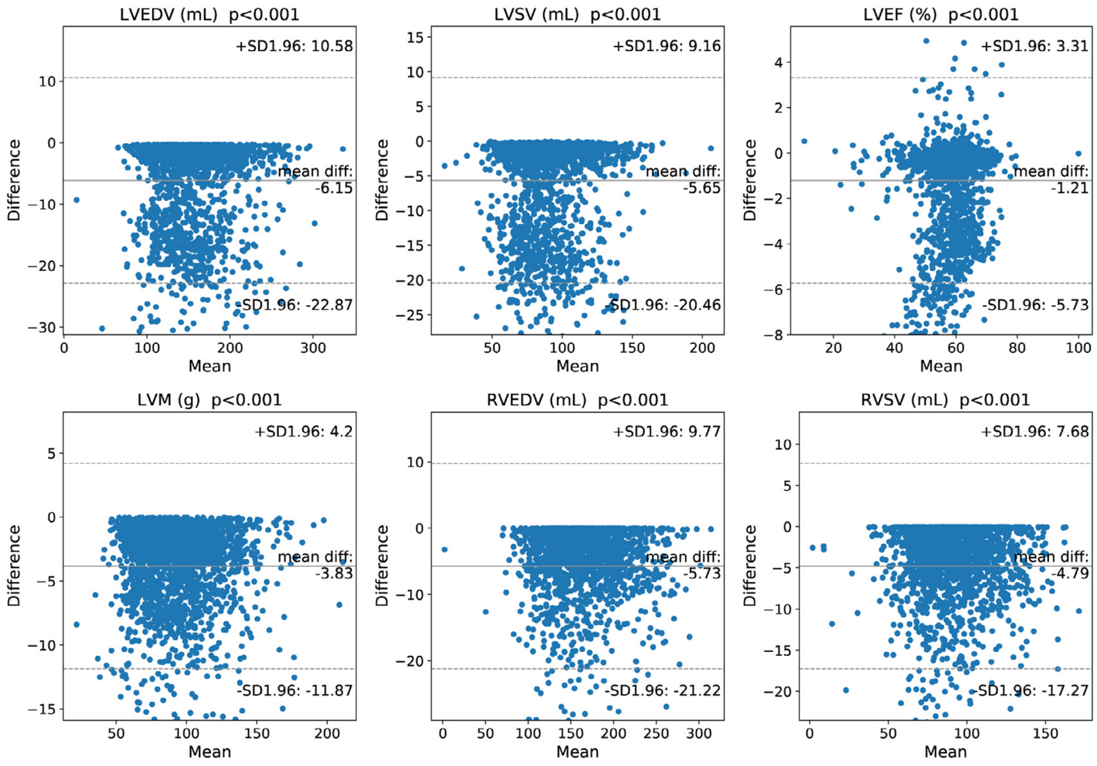

Fig. 14. Bland-Altman analyses for the LVEDV, LVSV, LVEF, LVM, RVEDV and RVSV measurements using the acquired cine CMR images (incomplete) and the imputed cine CMR images (complete). Statistically significant differences were found regarding those clinical parameters (p<0.001).

for the incomplete coverage data, as the ventricles with larger size tend to be insufficiently covered during the acquisition in real-world scenario. The Bland-Altman analysis revealed that the mean difference with 95% confidence interval (CI) between the original acquired and compensated images are -6.15 mL (95% CI, -22.87 mL to 10.58 mL) for LVEDV; -3.83 g (95% CI, -11.87 g to 4.2 g) for LVM; and -5.73 mL (95% CI, -21.22 mL to 9.77 mL) for RVEDV (cf. Fig. 14). Note that the distributions lie towards the negative side of the axis as we systematically synthesised more slices to the image stacks to compensate for the under-estimation, thereby, increasing the volume of the ventricles. These statistical analyses emphasise the statistical differences between the cardiac indices calculated from the incomplete CMR images and the compensated ones in UKBB and highlight the potential of the data completion pipeline for large-scale CMR population studies.

4.4. Cross-Database evaluation (ADSB dataset)

We also evaluated the proposed generative model on the publicly accessible ADSB dataset. After excluding several cases with stacks of only four or less slices, 191 subjects of original volume images were used as the reference and we manually removed two slices from the top and bottom to simulate incomplete data. Qualitative results of two subjects at the ED phase are shown in Fig. 15, where all the results were obtained by the generative models pretrained on the UKBB data. From these results, we can observe that the performance of the proposed method drops slightly relative to that of the UKBB data, in terms of anatomical plausibility and texture artefacts. This is expected as the ADSB dataset differs in appearance compared with the UKBB data. Fig. 16 shows the analysis of the absolute errors between the incomplete, imputed CMR images and reference in the ADSB dataset for the computed cardiac indices, when two slices either at the apical or basal region are missing. We can see that errors are dramatically reduced in the volume calculation by employing the proposed method, despite the high inter-subject variability of ADSB images. This demonstrates the generalisation ability and robustness of the proposed approach.

4.5. Ablation study

This section presents the ablation study results to support the architectural design. First, we systematically analysed the effectiveness and the contribution of each component in the proposed method. The comparison was conducted on three variants that correspond to replacing the CBN with standard batch normalisation, using a single discriminator and removing feature matching loss in turn, namely, w/o CBN, w/o CBN + MD and w/o CBN + MD + FM. For fair comparison, we retrained these variant networks using the same hyperparameters as the proposed method. Results from each of these network configurations are then compared with the proposed model.

Fig. 17 illustrates a visual comparison of the images generated in this ablation study, when the first slice is missing. The proposed method with all the components integrated produces results not only most visually comparable to that of the reference, but also perceptually appealing and anatomically plausible. Quantitative results in Table 10 confirm this observation and indicate that the CBN and multi-scale discriminator yield better results than the conventional batch normalisation and a single discriminator. Removing or replacing them results in a significant drop in performance for both the apical and basal slice synthesis (e.g., a CC of 0.892 ± 0.031 to 0.855 ± 0.038 for the apex and 0.898 ± 0.034 to 0.864 ± 0.04 for the base). We can see that the feature matching loss term further boosts the quality of the images synthesised, which is reflected by an average improvement of 0.836 to 0.855 and 21.13 dB to 22.52 dB for the CC and PSNR for the apex, and 0.848 to 0.864 and 22.22 dB to 23.13 dB for the CC and PSNR for the base, respectively. The progressively increasing Fréchet inception distance (FID) values also demonstrate the degradation of image quality. Statistical significance test with the Wilcoxon signed

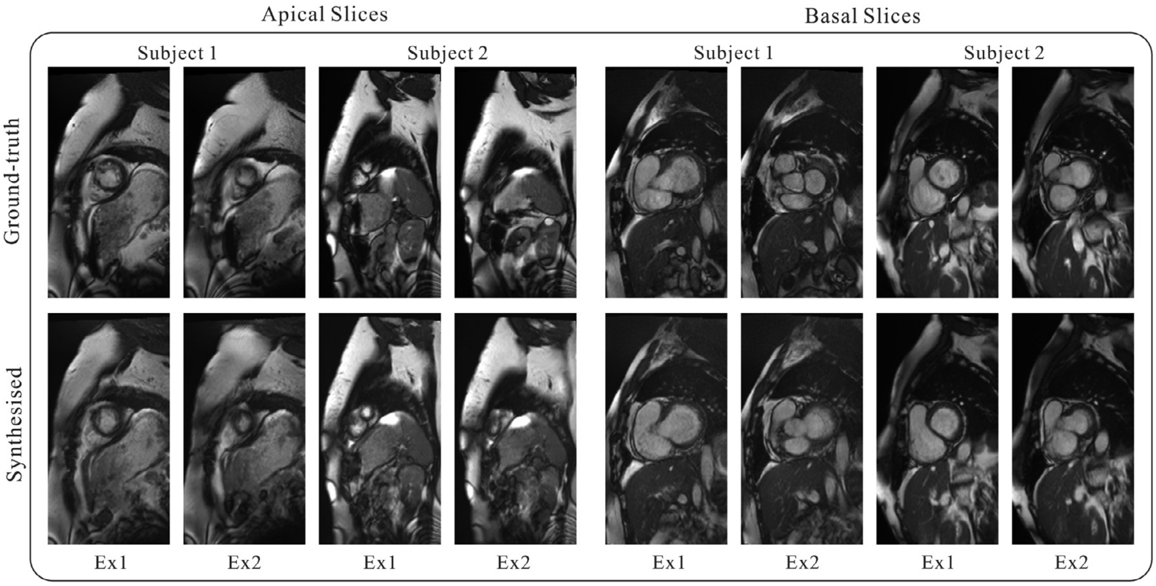

Fig. 15. Qualitative comparison of the ground truth and the synthesised slices for 2 subjects of the ADSB dataset using the pre-trained models on the UKBB dataset. Ex1 and Ex2 represent the first, the second extended slice, respectively.

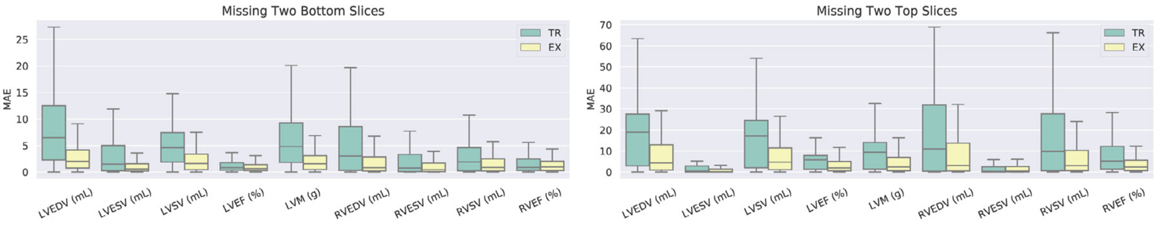

Fig. 16. Box plots of the absolute differences between the incomplete images (denoted as TR), compensated images (denoted as EX) and the complete images for the ADSB dataset.

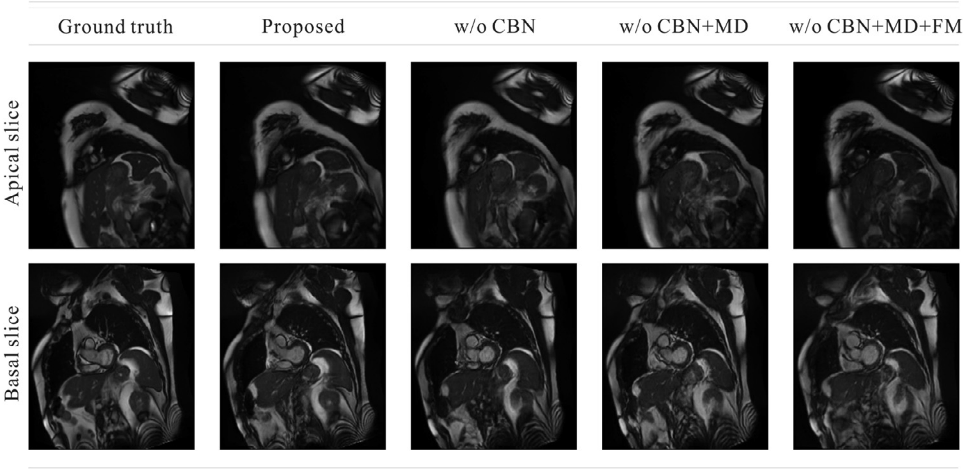

Fig. 17. An overview on results of the ablation study. Comparison of three variants that correspond to replacing the CBN with standard batch normalisation, using a single discriminator and removing feature matching loss in turn, namely, w/o CBN, w/o CBN + MD and w/o CBN + MD + FM, respectively. The CMR images were reproduced with the permission of UK Biobank.

rank test shows the computed p-value < 0.05 in terms of CC and PSNR metrics, suggesting that the proposed method achieves statistically significant improvements over its counterparts.

Also, to demonstrate the contribution of the cascaded structure, we compared it with another two imputation strategies: 1) Using a single generator that learns to recursively generate the next slice based on three consecutive slices at end of the image stack (denoted as “single (3 to 1)”) and 2) a single generator that learns to produce three consecutive slices simultaneously (denoted as “single (3 to 3)”). Qualitative and quantitative results are presented in Fig. 18 and Table 11, respectively. In Fig. 18, we show an example of the third imputed slice from different methods compared with the ground truth slice. We can see that simultaneously synthesising three consecutive slices is more challenging, leading to severe degradation in image quality. Using one generator that learns to infer the next slice recursively yields improvements but suffers from

Table 10
Summary of the ablation study results evaluated on the generated images when the first slice is missing. Comparison of three variants that correspond to replacing the CBN with standard batch normalisation, using a single discriminator and removing feature matching loss in turn, namely, w/o CBN, w/o CBN + MD and w/o CBN + MD + FM, respectively. The symbol “*” indicates the statistical significance (p-value < 0.05) compared with the proposed method.

	Apex			Base		
Metrics	CC	PSNR	FID	CC	PSNR	FID
Proposed	**0.892 ± 0.031**	**23.82 ± 1.81**	**23.55**	**0.898 ± 0.034**	**25.37 ± 1.61**	**20.59**
w/o CBN*	0.863 ± 0.04	22.84 ± 1.96	29.11	0.868 ± 0.041	23.24 ± 1.53	28.19
w/o CBN+MD*	0.855 ± 0.038	22.52 ± 1.55	34.19	0.864 ± 0.04	23.13 ± 1.53	29.71
w/o CBN+MD+FM*	0.836 ± 0.039	21.13 ± 1.38	43.42	0.848 ± 0.041	22.22 ± 1.41	31.34

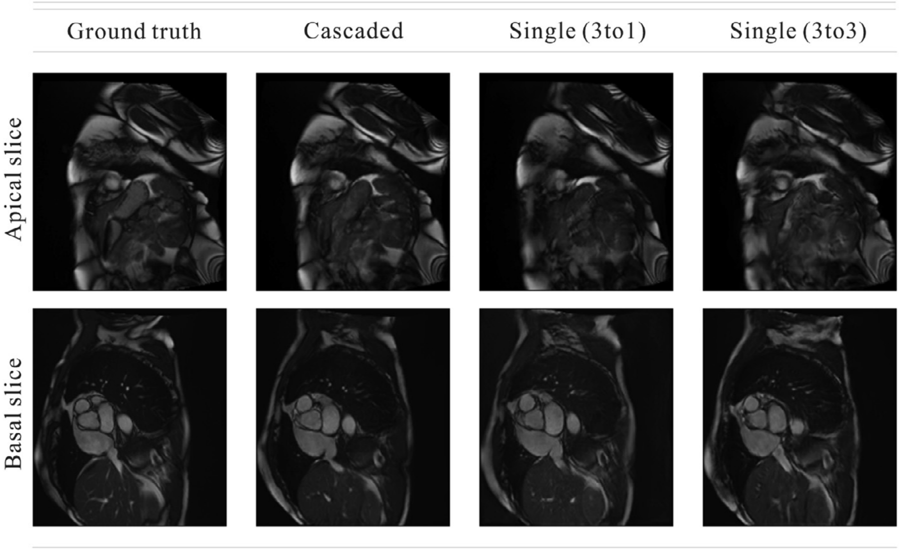

Fig. 18. Qualitative comparison of the proposed cascaded model with two native imputation strategies on the third imputed slice. The CMR images were reproduced with the permission of UK Biobank.

Table 11
Summary of the CC, PSNR and FID measurements between the ground truth and synthesised images from the three imputation approaches on 500 subjects. The symbol “*” indicates the statistical significance (p-value < 0.05) compared with the cascaded method.

	Apex			Base		
Metrics	CC	PSNR	FID	CC	PSNR	FID
Cascaded (3 to 1)	**0.738 ± 0.061**	**20.16 ± 1.91**	**38.27**	**0.767 ± 0.071**	**21.80 ± 1.53**	**31.58**
Single (3 to 1)*	0.707 ± 0.058	19.33 ± 1.46	49.67	0.715 ± 0.070	20.56 ± 1.41	40.73
Single (3 to 3)*	0.681 ± 0.055	18.71 ± 1.52	98.87	0.703 ± 0.073	20.13 ± 1.49	79.53

the errors accumulated during imputation of the first and second slices. In contrast, the proposed cascaded model learns to accommodate for such errors in the synthesised images, enabling their effective use as input stacks to the subsequent generators. This is because, in the recursive single generator scenario, the model is trained with image stacks comprising only “real slices”, to learn to synthesise the missing slice. Consequently, when multiple slices are missing, the trained generator is used to recursively synthesise the missing slices (during inference). However, as in this scenario, the generator is never trained with synthesised slices as part of the input image stack, the recursive synthesis is less effective. The cascaded model proposed in this study retrains the generator with image stacks that include one or more synthesised slices and thus is designed exactly for this purpose. The proposed method yields the most comparable results to the ground truth image in terms of visual quality, similarity and plausible textures. Besides visual inspection, the CC, PSNR and FID values computed over 500 test sets in Table 11 also demonstrated the improvement of the cascaded method over the other imputation strategies. The computed p-value < 0.05 shows that the proposed cascaded structure yields statistically significant improvements in performance over its counterparts.

5. Discussions

The development of automatic and generic approaches to compensate for incomplete cardiac coverage in CMR images, can have a transformative impact on high-throughput image analysis of population studies by avoiding the inclusion of thousands of sub-optimal CMR images in quantitative analyses and preventing the exclusion of identified incomplete data from such analyses. Violation of either may reduce statistical power and introduce bias aggregated over a given cohort. In this work, we proposed an efficient and robust two-stage pipeline to address this problem, i.e. automatic detection of apical and basal slices, followed by slice synthesis for the missing position. The detection network adopted ConvLSTM networks to leverage the inter-slice spatial contextual information in volumetric data. Experimental results demonstrated the effectiveness and efficiency of the proposed model and its superior performance over the 2D and 3D convolutional networks. As opposed to the methods reported in (Tarroni et al., 2018), the proposed model uses only the SAX stack and thus does not depend on the availability of CMR data from LAX views. This is useful in a practical setting as it does not preclude the use of the workflow in scenarios where the LAX views are absent.

In practice, we prefer high sensitivity than specificity during slice detection since the extra synthesised slices above basal slice or below apical slice, if any, will have a negligible effect on cardiac volume calculation as they are returned to zeros by manual/automatic segmentation methods. During image acquisition, sufficient margin is also left above and below the LV cavity in SAX image stacks, according to the established guidelines for cardiac MR image acquisition (Schulz-Menger et al., 2013).

For apical and basal slice synthesis, we developed a conditional GAN that encourages visual properties and fine details are effectively propagated through the image generation pathway. We also designed a cascaded network to address multi-slice imputation. Experimental results showed that the proposed approach not only yields visually comparable slices to the acquired data with a full coverage, but also retains the accuracy of anatomical and functional cardiac parameters of clinical interest. For instance, three missing basal slices has the most severe impact on volume calculation in the current investigation. The mean differences to the reference of LVEDV, LVM, RVEDV are -2.6 mL, -2.3 g and -4.9 mL, respectively, which are thus considerably smaller than those obtained from the incomplete image stacks (-54.7 mL, -30.5 g and -62.2 mL), indicative of the anatomical validity of the imputed image volumes.

Although we only focused on scenarios where up to three consecutive slices are missing in SAX stacks in this study, the model can be extended to synthesise more absent slices by choosing additional three-slice blocks towards the mid-cavity region in the training samples. The sensitivity and robustness of the model to larger portions of missing data are yet to be assessed and could be the focus of future work. We also observed that the quality of synthesised basal slices is better than that achieved for the apical slices, due to large variations in appearance of the apical slices. This is preferred for cardiac quantification as the absence of the basal slices has a significant impact on volume calculation.

As an attempt, we also investigated a variant of the proposed imputation method by incorporating three LAX images (i.e., 2CH, 3CH and 4CH views) as an additional input to the network to impute the missing SAX slices. We found no obvious differences between the two approaches in terms of the quantitative metrics (CC, PSNR and FID) evaluated.

Future work would involve the following aspects. First, in this work we proposed to synthesise a SAX slice by learning relevant features from three neighbouring slices. The number of slices used was selected through a pilot study. We found that fewer than 3 slices will limit the ability to accurately model the intrinsic manifold of the image data, whereas more than 3 slices will also be less effective and increase the errors due to 1) overfitting caused by increased complexity and nonlinearity of the model and 2) raised influence by the presence of slice misalignment resulting from patient motion. Further incorporating motion compensation algorithms or physical constraints and shape priors into the generative model may improve the results and will be the subject of future work. Second, as CMR cine images typically cover the full cardiac cycle, it may be beneficial to exploit the temporal aspect of the dynamic CMR data and leverage redundant information from different cardiac phases to boost the slice detection and imputation performance further. Third, the UKBB dataset used in the experiments involves general population subjects, i.e., it contains both healthy and diseased patients with a prevalence approximating that of the general population. Hence, future studies should assess in more detail the generalisability and performance of the proposed generative network on CMR images across pathologies more specifically. Hence, future studies should assess in more detail the generalisability and performance of the proposed generative network on CMR images across pathologies more specifically. Last, after the proposed heart completion an automatic quality control step can be applied to assess whether the synthesised CMR slices are realistic, in order to facilitate clinical translation of the proposed approach. Automatic quality control may be approach in several ways, for example – (i) slice-wise quantification of cardiac morphological indices (such as myocardial thickness, blood pool area etc.) could be used to identify whether there is a smooth transition in the synthesised slice from the adjacent slices. Large changes in these values in the synthesised slices might be indicative of incorrect synthesis (when considering healthy populations); (ii) an independent classification network could be trained using the synthesised slices for training. Training such a network, however, would require grading of the synthesised slices a priori based on quality, by experienced cardiologists/cardiac imaging experts.

6. Conclusion

In this work, we proposed an effective two-stage pipeline for detecting and synthesising missing slices in cardiac apex and base, to address incomplete heart coverage, which hinders accurate measurement of cardiac volume and functional assessment. The detection model employed several dense blocks consisting of ConvLSTM layers, to leverage 3D contextual feature and exploit the sequential ordering of SAX slices, and achieve reliable classification outcomes. The imputation network incorporated visual properties and fine details into the image generation and thus can infer slices that are anatomically plausible and comparable to the acquired complete data. Extensive experimental results demonstrated that the proposed approach is robust and reliable. Notably, the accuracy of subsequent quantification can be improved for CMR datasets with sub-optimal coverage, without the need for re-scanning the patient or completely discarding such a data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Yan Xia: Conceptualization, Methodology, Software, Visualization, Validation, Formal analysis, Writing – original draft. **Nishant Ravikumar:** Methodology, Writing – review & editing. **Alejandro F. Frangi:** Conceptualization, Writing – review & editing, Supervision.

Acknowledgements

This research has been conducted using the UK Biobank Resource under Application 11350. The authors are grateful to all UK Biobank participants and staff. The work of AFF is funded by the Engineering and Physical Sciences Research Council (EPSRC) through TUSCA (EP/V04799X/1).

Appendix

A visual comparison of the synthesised slices by the proposed imputation networks and the ground-truth images for 7 subjects of the UKBB dataset at the ES phase is depicted in Fig. 19.

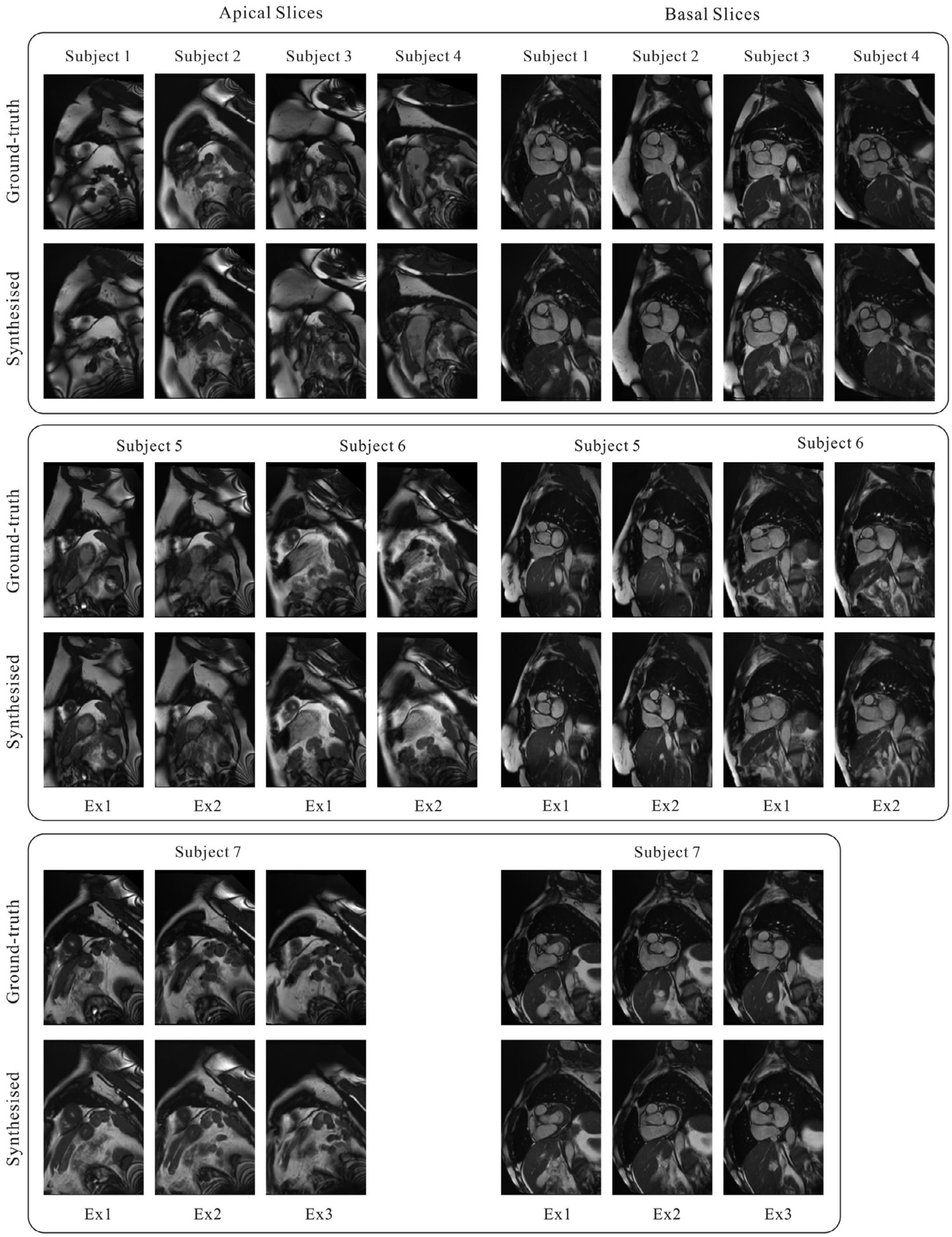

Fig. 19. Qualitative comparison of the ground truth and the synthesised slices for 7 subjects of the UKBB dataset at the ES phase. From top to bottom, results of three degrees of incomplete coverage are shown. Ex1, Ex2 and Ex3 represent the first, the second and the third extended slice, respectively. The CMR images were reproduced with the permission of UK Biobank.

References

Anand, S.S., Tu, J.V., Awadalla, P., Black, S., Boileau, C., Busseuil, D., Desai, D., Després, J.-P., de Souza, R.J., Dummer, T., Jacquemont, S., Knoppers, B., Larose, E., Lear, S.A., Marcotte, F., Moody, A.R., Parker, L., Poirier, P., Robson, P.J., Smith, E.E., Spinelli, J.J., Tardif, J.-C., Teo, K.K., Tusevljak, N., Friedrich, M.G., 2016. Rationale, design, and methods for canadian alliance for healthy hearts and minds cohort study (CAHHM)–a pan canadian cohort study. BMC Public Health 16 (1), 650.

Bai, W., Sinclair, M., Tarroni, G., Oktay, O., Rajchl, M., Vaillant, G., Lee, A.M., Aung, N., Lukaschuk, E., Sanghvi, M.M., et al., 2018. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. Journal of Cardiovascular Magnetic Resonance 20 (1), 65.

Bamberg, F., Kauczor, H.-U., Weckbach, S., Schlett, C.L., Forsting, M., Ladd, S.C., Greiser, K.H., Weber, M.-A., Schulz-Menger, J., Niendorf, T., Pischon, T., Caspers, S., Amunts, K., Berger, K., Bulow, R., Hosten, N., Hegenscheid, K., Kroncke, T., Linseisen, J., Gunther, M., Hirsch, J.G., Kohn, A., Hendel, T., Wichmann, H.-E., Schmidt, B., Jockel, K.-H., Hoffmann, W., Kaaks, R., Reiser, M.F., Volzke, H., 2015. Whole-body MR imaging in the german national cohort: rationale, design, and technical background. Radiology 277 (1), 206–220.

Basty, N., Grau, V., 2018. Super Resolution of Cardiac Cine Mri Sequences Using Deep Learning. In: Image Analysis for Moving Organ, Breast, and Thoracic Images. Springer, pp. 23–31.

Byeon, W., Breuel, T.M., Raue, F., Liwicki, M., 2015. Scene labeling with lstm recurrent neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3547–3555.

Carapella, V., Jiménez-Ruiz, E., Lukaschuk, E., Aung, N., Fung, K., Paiva, J., Sanghvi, M., Neubauer, S., Petersen, S., Horrocks, I., et al., 2016. Towards the Semantic Enrichment of Free-text Annotation of Image Quality Assessment for Uk Biobank Cardiac Cine Mri Scans. In: Deep Learning and Data Labeling for Medical Applications. Springer, pp. 238–248.

Chen, T., Lučić, M., Houlsby, N., Gelly, S., 2019. On self-modulation for generative adversarial networks. In: International Conference on Learning Representations (ICLR).

Dalca, A.V., Bouman, K.L., Freeman, W.T., Rost, N.S., Sabuncu, M.R., Golland, P., 2018. Medical image imputation from image collections. IEEE Trans Med Imaging 38 (2), 504–514.

Dar, S.U., Yurt, M., Karacan, L., Erdem, A., Erdem, E., Çukur, T., 2019. Image synthesis in multi-contrast mri with conditional generative adversarial networks. IEEE Trans Med Imaging 38 (10), 2375–2388.

De Vries, H., Strub, F., Mary, J., Larochelle, H., Pietquin, O., Courville, A.C., 2017. Modulating early visual processing by language. In: Advances in Neural Information Processing Systems, pp. 6594–6604.

Durugkar, I., Gemp, I., Mahadevan, S., 2017. Generative multi-adversarial networks. In: International Conference on Learning Representations (ICLR).

Ferreira, P.F., Gatehouse, P.D., Mohiaddin, R.H., Firmin, D.N., 2013. Cardiovascular magnetic resonance artefacts. Journal of Cardiovascular Magnetic Resonance 15 (1), 41.

Guo, Y., Bi, L., Ahn, E., Feng, D., Wang, Q., Kim, J., 2020. A spatiotemporal volumetric interpolation network for 4d dynamic medical image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4726–4735.

Han, Z., Wei, B., Mercado, A., Leung, S., Li, S., 2018. Spine-GAN: semantic segmentation of multiple spinal structures. Med Image Anal 50, 23–35.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

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

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134.

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. Springer, pp. 694–711.

Klinke, V., Muzzarelli, S., Lauriers, N., Locca, D., Vincenti, G., Monney, P., Lu, C., Nothnagel, D., Pilz, G., Lombardi, M., van Rossum, A.C., Wagner, A., Bruder, O., Mahrholdt, H., Schwitter, J., 2013. Quality assessment of cardiovascular magnetic resonance in the setting of the european CMR registry: description and validation of standardized criteria. Journal of Cardiovascular Magnetic Resonance 15 (1), 55.

Lee, D., Kim, J., Moon, W.-J., Ye, J.C., 2019. Collagan: Collaborative gan for missing image data imputation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2487–2496.

Manjón, J.V., Coupé, P., Buades, A., Fonov, V., Collins, D.L., Robles, M., 2010. Non-local MRI upsampling. Med Image Anal 14 (6), 784–792.

Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S., 2017. Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2794–2802.

Miyato, T., Koyama, M., 2018. cGANs with projection discriminator. In: International Conference on Learning Representations (ICLR).

Nguyen, T., Le, T., Vu, H., Phung, D., 2017. Dual discriminator generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2670–2680.

Park, T., Liu, M.-Y., Wang, T.-C., Zhu, J.-Y., 2019. Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2337–2346.

Pennell, D.J., 2003. Cardiovascular magnetic resonance: twenty-first century solutions in cardiology. Clinical medicine 3 (3), 273.

Petersen, S.E., Matthews, P.M., Bamberg, F., Bluemke, D.A., Francis, J.M., Friedrich, M.G., Leeson, P., Nagel, E., Plein, S., Rademakers, F.E., Young, A.A., Garratt, S., Peakman, T., Sellors, J., Collins, R., Neubauer, S., 2013. Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of UK biobank-rationale, challenges and approaches. Journal of Cardiovascular Magnetic Resonance 15 (1), 46.

Petersen, S.E., Matthews, P.M., Francis, J.M., Robson, M.D., Zemrak, F., Boubertakh, R., Young, A.A., Hudson, S., Weale, P., Garratt, S., Collins, R., Piechnik, S., Neubauer, S., 2015. UK Biobank'S cardiovascular magnetic resonance protocol. Journal of cardiovascular magnetic resonance 18 (1), 8.

Petersen, S.E., Sanghvi, M.M., Aung, N., Cooper, J.A., Paiva, J.M., Zemrak, F., Fung, K., Lukaschuk, E., Lee, A.M., Carapella, V., et al., 2017. The impact of cardiovascular risk factors on cardiac structure and function: insights from the uk biobank imaging enhancement study. PLoS ONE 12 (10), e0185114.

Plenge, E., Poot, D.H., Niessen, W.J., Meijering, E., 2013. Super-resolution reconstruction using cross-scale self-similarity in multi-slice MRI. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 123–130.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training GANs. In: Advances in neural information processing systems, pp. 2234–2242.

Sánchez, I., Vilaplana, V., 2018. Brain mri super-resolution using generative adversarial networks. In: International conference on Medical Imaging with Deep Learning: Amsterdam, 4-6th July 2018, pp. 1–8.

Schulz-Menger, J., Bluemke, D.A., Bremerich, J., Flamm, S.D., Fogel, M.A., Friedrich, M.G., Kim, R.J., von Knobelsdorff-Brenkenhoff, F., Kramer, C.M., Pennell, D.J., et al., 2013. Standardized image interpretation and post processing in cardiovascular magnetic resonance: society for cardiovascular magnetic resonance (scmr) board of trustees task force on standardized post processing. Journal of Cardiovascular Magnetic Resonance 15 (1), 1–19.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-c., 2015. Convolutional lstm network: a machine learning approach for precipitation nowcasting. Adv Neural Inf Process Syst 28, 802–810.

Sundermeyer, M., Schlüter, R., Ney, H., 2012. Lstm neural networks for language modeling. Thirteenth annual conference of the international speech communication association.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826.

Tarroni, G., Oktay, O., Bai, W., Schuh, A., Suzuki, H., Passerat-Palmbach, J., de Marvao, A., O'Regan, D., Cook, S., Glocker, B., Matthews, P., Rueckert, D., 2018. Learning-based quality control for cardiac MR images. IEEE Trans Med Imaging 38 (5), 1127–1138.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., Catanzaro, B., 2018. High-resolution image synthesis and semantic manipulation with conditional GANs. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8798–8807.

Xia, Y., Zhang, L., Ravikumar, N., Attar, R., Piechnik, S.K., Neubauer, S., Petersen, S.E., Frangi, A.F., 2020. Recovering from missing data in population imaging–Cardiac MR image imputation via conditional generative adversarial nets. Med Image Anal 67, 101812.

Yang, H., Sun, J., Carass, A., Zhao, C., Lee, J., Xu, Z., Prince, J.L., 2018. Unpaired Brain MR-to-CT Synthesis Using a Structure-constrained CycleGAN. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 174–182.

Yurt, M., Dar, S.U.H., Erdem, A., Erdem, E., Çukur, T., 2019. Mustgan: multi--stream generative adversarial networks for MR image synthesis. arXiv preprint arXiv:1909.11504.

Zhang, H., Goodfellow, I., Metaxas, D., Odena, A., 2018. Self-Attention generative adversarial networks. Stat 1050, 21.

Zhang, L., Gooya, A., Frangi, A.F., 2017. Semi-supervised assessment of incomplete LV coverage in cardiac MRI using generative adversarial nets. In: International Workshop on Simulation and Synthesis in Medical Imaging. Springer, pp. 61–68.

Zhang, L., Gooya, A., Pereanez, M., Dong, B., Piechnik, S.K., Neubauer, S., Petersen, S.E., Frangi, A.F., 2018. Automatic assessment of full left ventricular coverage in cardiac cine magnetic resonance imaging with fisher-discriminative 3-D CNN. IEEE Trans. Biomed. Eng. 66 (7), 1975–1986.

Zhuang, X., Shen, J., 2016. Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. Med Image Anal 31, 77–87.