UNIVERSITY OF LEEDS

This is a repository copy of *Recovering from missing data in population imaging – Cardiac MR image imputation via conditional generative adversarial nets*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/176404/

Version: Accepted Version

Article:
Xia, Y, Zhang, L, Ravikumar, N et al. (5 more authors) (2021) Recovering from missing data in population imaging – Cardiac MR image imputation via conditional generative adversarial nets. Medical Image Analysis, 67. 101812. ISSN 1361-8415

https://doi.org/10.1016/j.media.2020.101812

© 2020, Elsevier. This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/.

Reuse
This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Medical Image Analysis (2021)

Recovering from Missing Data in Population Imaging – Cardiac MR Image Imputation via Conditional Generative Adversarial Nets

Yan Xia[a,b,*], Le Zhang[c,d], Nishant Ravikumar[a,b], Rahman Attar[a,b], Stefan K. Piechnik[e], Stefan Neubauer[e], Steffen E. Petersen[f], Alejandro F. Frangi[a,b,g,*]

[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] *Queen Square Institute of Neurology, University College London, London, UK*
[d] *Centre for Medical Image Computing, Department of Computer Science, University College London, London, UK*
[e] *Oxford Center for Clinical Magnetic Resonance Research (OCMR), Division of Cardiovascular Medicine, John Radcliffe Hospital, University of Oxford, Oxford, UK*
[f] *William Harvey Research Institute, Barts Heart Centre, Barts Health NHS Trust, Queen Mary University of London, London, UK*
[g] *Medical Imaging Research Center (MIRC), University Hospital Gasthuisberg, and Cardiovascular Science and Electronic Engineering Departments, KU Leuven, Leuven, Belgium*

ARTICLE INFO

Article history:
Received –
Received in final form –
Accepted –
Available online –

Communicated by –

2000 MSC:

Keywords: Deep learning, Data imputation, Conditional generative adversarial net, Conditional batch normalisation, Multi-scale discriminator, Cardiac MRI

ABSTRACT

Accurate ventricular volume measurements are the primary indicators of normal/abnormal cardiac function and are dependent on the Cardiac Magnetic Resonance (CMR) volumes being complete. However, missing or unusable slices owing to the presence of image artefacts such as respiratory or motion ghosting, aliasing, ringing and signal loss in CMR sequences, significantly hinder accuracy of anatomical and functional cardiac quantification, and recovering from those is insufficiently addressed in population imaging. In this work, we propose a new robust approach, coined Image Imputation Generative Adversarial Network (I2-GAN), to learn key features of cardiac short axis (SAX) slices near missing information, and use them as conditional variables to infer missing slices in the query volumes. In I2-GAN, the slices are first mapped to latent vectors with position features through a regression net. The latent vector corresponding to the desired position is then projected onto the slice manifold, conditioned on intensity features through a generator net. The generator comprises residual blocks with normalisation layers that are modulated with auxiliary slice information, enabling propagation of fine details through the network. In addition, a multi-scale discriminator was implemented, along with a discriminator-based feature matching loss, to further enhance performance and encourage the synthesis of visually realistic slices. Experimental results show that our method achieves significant improvements over the state-of-the-art, in missing slice imputation for CMR, with an average SSIM of 0.872. Linear regression analysis yields good agreement between reference and imputed CMR images for all cardiac measurements, with correlation coefficients of 0.991 for left ventricular volume, 0.977 for left ventricular mass and 0.961 for right ventricular volume.

1. Introduction

Cardiac Magnetic Resonance (CMR) imaging is generally accepted as the reference standard for several aspects in cardiovascular medicine, such as the accurate assessment of left and right ventricular (LV/RV) mass, volume and function, and the quantification of myocardial fibrosis, providing insights into cardiac diseases in a single diagnostic session (Pennell, 2003). Due to its excellent reproducibility of quantitative measurements compared with other modalities, CMR is a robust and attractive non-invasive technique for population imaging and is chosen in several population-based studies (Marwick et al., 2013), such as the UK Biobank (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).

However, CMR image quality can be adversely affected by

*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: y.xia@leeds.ac.uk (Yan Xia), a.frangi@leeds.ac.uk (Alejandro F. Frangi)

numerous imaging artefacts, which could ultimately lead to reduced diagnostic accuracy. For instance, accurate and reproducible ventricular volume assessments, and associated Ejection Fraction (EF) and Stroke Volume (SV), are important in the management of various cardiac diseases because they are strong predictors of clinical outcomes (Knauth et al., 2008). The measurements directly relate to the volume of these ventricular chambers, whose extents are defined by the basal and apical slices (Frangi et al., 2001). Most published studies addressing computation of EF and SV assume a complete data set, i.e. all imaging planes being available for all cardiac time frames. In population studies, clinical trials, or clinical routine, this assumption may not hold in CMR imaging as certain datasets may have missing or corrupted information due to imaging artefacts and acquisition/storage errors (Ferreira et al., 2013; Van der Graaf et al., 2014). For instance, short axis (SAX) cine stacks are reconstructed from images acquired across multiple breath-holds (typically 1-3 slices/breath-hold). Consequently, cardiac and respiratory motion, together with fast flowing blood in the vicinity, may cause inter-slice motion artefacts. In addition, adjacent tissues with different characteristics or implants may cause local loss of signal in certain slices (Van der Graaf et al., 2014). Due to insufficient radiographer experience in scan acquisition planning, natural cardiac muscle contraction, breathing motion, and imperfect triggering, CMRs may display incomplete LV coverage, posing further challenges to quantitative LV characterisation and accurate diagnosis (Zhang et al., 2018b).

Although several automated, learning-based image quality assessment (QA) techniques have been proposed for CMR images (Tarroni et al., 2018; Zhang et al., 2018b), a common strategy to account for missing/corrupted slices and unavailable features is to disregard incomplete samples in the cohort (Klinke et al., 2013). However, excluding data not only reduces statistical power and causes bias, but is also of ethical and financial concern as partially acquired subject data remains unused, and limits the application of such methods to similarly complete datasets. Thus, image imputation for these corrupted/missing slices is an important step after QA in order to enable quantitative assessment of cardiac function. However, large inter-slice spacing in CMR imaging and variation in the intensity distributions of anatomy across slices, pose a significant challenge to traditional data imputation and interpolation approaches, as illustrated in Fig. 1.

Generally, data imputation based methods have been proposed to deal with this problem, such as using the mean of the data or model-based missing data estimation (García-Laencina et al., 2010). Stochastic regression imputation methods can use the information provided by the data to solve the collinearity problem caused by the high correlation of predicted variables (Schlomer et al., 2010). If the missing mechanism is random, the missing variable can be imputed by the marginal distribution of the observed data using maximum likelihood estimation (MLE) (Dong and Peng, 2013). This has been achieved previously by fitting parametric mixture models (such as Gaussian mixtures, for example) to the observed data using the EM algorithm, and sampling from the model to impute the missing data (Richardson and Weiss, 2018).

Traditional interpolation methods can be adapted to this scenario, by using intensity/object-based interpolation. Intensity-based interpolation methods are essentially weighted average schemes of the input image, and thus yield blurring effects and unrealistic results. In object-based methods, on the other hand, the extracted information from available slices is used for guiding the interpolation process in a more accurate manner. An important object interpolation category is based on image registration, where corresponding points between consecutive slices are found and then the interpolation is applied to find the in-between slices. Such methods include the modified version of control grid interpolation (CGI) (Frakes et al., 2008), multi-resolution registration-based slice interpolation (Leng et al., 2013), and higher order splines-based interpolation (Horváth et al., 2017). However, registration-based slice interpolation methods have several limitations: first, the consecutive slices must have similar anatomical features. Second, the registration method must be able to estimate the correct transformation to match these similar features. Violation of either of these aspects yields false correspondence maps, which leads to incorrect interpolation results. Additionally, such methods do not leverage statistical information concerning the anatomy and appearance that can be extracted from populations, which implies that they are sensitive to the input data, and susceptible to producing physically implausible solutions due to the presence of local minima. A comprehensive summary of common methods for slice interpolation was described in (Grevera and Udupa, 1998; Grevera et al., 1999).

Data-driven image synthesis using generative adversarial networks (GANs) (Goodfellow et al., 2014) has significantly improved various medical image synthesis tasks, such as data augmentation, super-resolution, denoising and cross-modality domain adaption (Zhuang and Shen, 2016; Han et al., 2018; Yang et al., 2018; Sánchez and Vilaplana Besler, 2018), to name a few. 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 (Lee et al., 2019). GANs comprise two sub-networks that are trained simultaneously, namely, the generator and the discriminator. They are trained by optimising a suitable adversarial loss and automatically adapting to the differences between the synthesised and real images in the target domain. In addition, GANs can be extended to a conditional model (Mirza and Osindero, 2014) if both the generator and discriminator are conditioned on some extra auxiliary information, such as class labels or data from other modalities. Recent studies on image-to-image translation have demonstrated that such conditional GANs are highly effective in learning the mapping between statistically dependent source and target domain images (Isola et al., 2017).

In this paper, we adapt the conditional GAN to generate missing CMR slices, following manual, semi- or fully-automatic quality control (QC). We propose a novel, robust method for image imputation based on a generative adversarial network (I2-GAN) to infer missing slices, conditioned on information in adjacent image slices. The main contributions of our approach are:

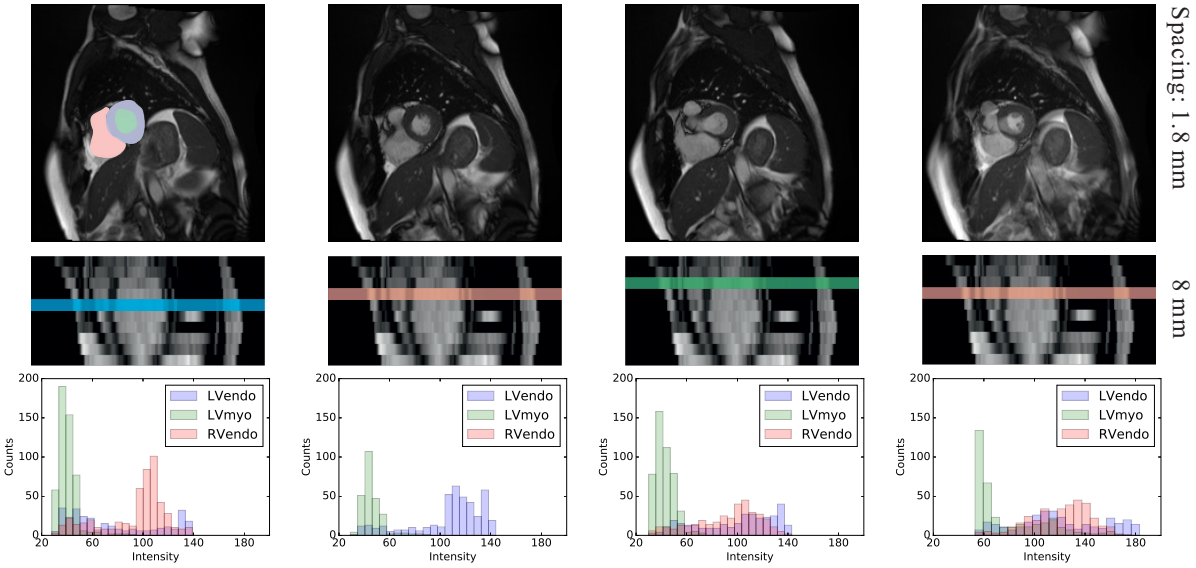

Fig. 1. Typical CMR images from UK Biobank with in-plane spatial resolution 1.8×1.8 mm, slice thickness 8 mm. The three rows display short axis (SAX) slices, long axis (LAX) slices, and histogram of corresponding slice over three selected cardiac ROIs (marked as different colours in the first image), respectively. Large inter-slice spacing and variation in the intensity distributions of the same anatomical regions across consecutive slice positions can be observed. The first three images in the top row show three consecutive slices and mean imputation of the middle slice is shown in the forth image. The colour bars in LAX slice show the corresponding slice position for SAX slices. LVendo, LVmyo and RVendo represent LV endocardium, LV myocardium and RV endocardium, respectively.

(1) A novel deep conditional GAN architecture is proposed for generating missing SAX slices for CMR images across different positions. First, a 3D regression network learns features relevant to identifying the position of the missing CMR slice and predicts its corresponding position. Conditioned on these pre-trained, extracted feature representations, a generator and discriminator are subsequently trained to synthesise a realistic image in place of the missing slice.

(2) To synthesise visually appealing CMR images that retain the accuracy of standardised quantification analysis, we design a dedicated generator and discriminator in this work. The generator contains residual blocks, where all normalisation layers are conditioned and modulated with auxiliary slice information to ensure that fine details are effectively propagated through the network. A multi-scale discriminator is employed to ensure recovery of both global and local spatial features. Moreover, the typical pixel-wise loss function is replaced with a feature matching loss that operates in feature space to prevent unsatisfying perceptual quality.

(3) This paper extends the preliminary version of our method presented at the 2019 International Conference on Medical Image Computing and Computer-Assisted Intervention (Zhang et al., 2019), by extensively refining each component of the architecture to obtain more realistic imputation results. In addition, we comprehensively analyse the model by evaluating its performance across a larger cohort of subjects at, a single cardiac phase (end-diastole), and across multiple time points in the cardiac cycle. Additionally, we evaluate the clinical viability of the proposed framework by extracting quantitative cardiac functional indices from the imputed volumes and comparing them against values estimated for the original (complete) volumes.

The paper is organised as follows. Section II introduces the I2-GAN model and learning algorithm. Section III describes details of the experiments conducted to validate the model, and Section IV presents the results of statistical analyses conducted to evaluate model performance. Subsequently, we discuss important characteristics of our model in contrast with the state-of-the-art and its relevance in real clinical scenarios in Section V, before providing concluding remarks for our work in Section VI.

2. Method

In this section, we construct a generative model for missing slice imputation (MSI), present the resulting learning algorithm, and describe our image synthesis pipeline. The overview of the method is: first, a 3D regression net is used to predict the missing slice position y and extract features f relevant to regions in the vicinity of missing slices, for a given image volume. Then, a generative net takes the pre-computed intrinsic slice features f as conditioning input to synthesise a cardiac cine MRI slice at the corresponding position. Finally, a multi-scale discriminator distinguishes the generated samples from the real images, and simultaneously matches the features of the inferred slices with those of the real slices (i.e. using a feature matching loss). The network architecture is illustrated in Fig. 2.

2.1. Generative Adversarial Network (GAN)

A GAN consists of two networks: the generator G and the discriminator D (Goodfellow et al., 2014). In an image synthesis task, G builds a mapping function that maps a random noise vector z to an output image x from a target distribution $p(x)$, and D outputs a single scalar that represents the probability that

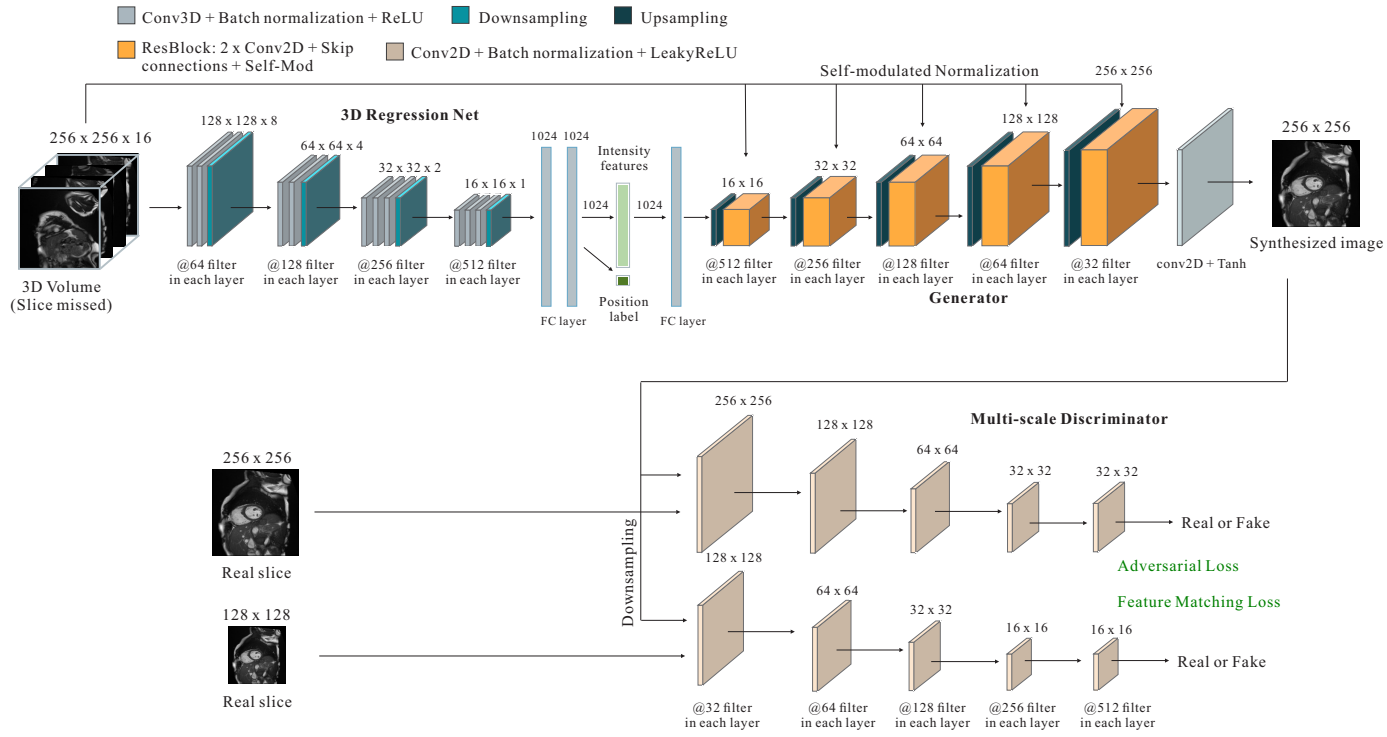

Fig. 2. Structure of the proposed I2-GAN network for cardiac MSI. The 3D regression network maps the input volume to a vector containing intensity features and position label. The former is the condition to feed to the generator. The generator contains several residual blocks, where all normalisation layers are self-modulated with auxiliary slice information to ensure that fine details are propagated throughout the network. The multi-scale discriminative net helps capture both global and local spatial features.

a sample s is drawn from $p(x)$. While G learns to synthesise realistic-looking images indistinguishable from real images, D tries to distinguish real images from the generated ones. The minimax objective for GANs can be formulated as follows[1]:

$$\min_G \max_D \mathcal{L}_{GAN} = \min_G \max_D \mathbb{E}_x \left[\log D(x)\right] + \mathbb{E}_z \left[\log\left(1 - D(G(z))\right)\right]. \quad (1)$$

To improve network stability during training, the negative log-likelihood in $\mathcal{L}_{GAN}$ can be replaced by a squared loss function (Mao et al., 2017):

$$\mathcal{L}_{LSGAN} = -\mathbb{E}_x \left[(D(x) - 1)^2\right] - \mathbb{E}_z \left[D(G(z))^2\right]. \quad (2)$$

2.2. Image Imputation Conditional GAN (I2-GAN) for CMR

We adapt the conditional generative model (Mirza and Osindero, 2014) and aim to transform the features from CMR volumes with full ventricular coverage into the query CMR volumes, which miss slices in certain positions. The conditional generator is defined as $G : \mathbb{R}^Z \times \mathbb{R}^F \rightarrow \mathbb{R}^S$, where F is the dimension of the intrinsic intensity, and S is the dimension of cardiac slice. The discriminator is denoted as $D : \mathbb{R}^S \rightarrow \mathbb{R}$. The optimisation of the G and D can be reformulated as:

$$\mathcal{L}_D = \mathbb{E}_x \left[(D(x) - 1)^2\right] + \mathbb{E}_f \left[D(G(f))^2\right] \quad (3)$$

$$\mathcal{L}_G = \mathbb{E}_f \left[D(G(f))^2\right]. \quad (4)$$

The architecture of the regression net, generator and discriminator will be discussed, in subsequent sections.

2.3. 3D Regression Net

As visual perception tasks in medical image analysis benefit from leveraging inter-slice context (Kamnitsas et al., 2017), a deep 3D convolutional neural network (CNN) is used to learn CMR image intensity features f relevant to predicting the position of missing slices. Note that the input cardiac volumetric image $\mathbf{X}$ for this network is an incomplete image stack with the missing slice. As shown in Fig. 2, the trunk architecture of the regression net consists of four 3D convolutional layers with (kernel size = 3^3, padding = 2^3 and stride = 2^3) and 2 fully-connected layers for feature extraction. The ReLU activation is used after each convolutional layer. We then configure two layers for extracting the 512-dimensional feature vector f and regressing a single scalar y, representing the missing slice position (normalised to the range [0,1]), separately. While training the regression net, the loss function converges quickly. Thus, we can easily learn missed slice position index y, and compute the intensity feature vector f. To evaluate the accuracy of the regression network, we conducted a 5-fold cross validation study over 4848 subjects from the UKBB, achieving an average correlation coefficient of 0.99 for correctly identifying the missing slice position. Note that, instead of directly concatenating the position label information y with the extracted intensity feature that may weaken the position signal, we incorporate it into the

[1]Here, we denote $\mathbb{E}_{x\sim p_{\text{data}}(x)}$ as $\mathbb{E}_x$ and $\mathbb{E}_{z\sim p_z(z)}$ as $\mathbb{E}_z$ for simplicity.

generator implicitly by encoding the two adjacent slices around the missing position through conditional batch normalistation layers introduced in the next section, to feed relevant spatial features as well as strengthen the position information.

2.4. Generator with Self-modulated Normalisation

Our generator network follows a full pre-activation ResNet architecture (He et al., 2016) implemented in recent popular conditional GAN and pix2pix models (Karras et al., 2017; Zhu et al., 2017; Isola et al., 2017; Miyato et al., 2018; Zhang et al., 2018a). The model consists of the residual blocks, followed by nearest neighbour upsampling layers. As shown in Fig. 3, each residual block contains two convolutional layers where a learned residue of input is added to the output to ensure the characteristics of original images are retained.

For feature normalisation, we are motivated by the concept of conditional batch normalisation (CBN) that has been adopted in multiple previous studies (De Vries et al., 2017; Miyato and Koyama, 2018; Zhang et al., 2018a; Chen et al., 2019; Park et al., 2019) and suggests a new conditioning mechanism to incorporate external conditioning information (such as labels, embeddings, masks or generator's own inputs) into image synthesis through batch normalisation. It is typically implemented as a learning-based affine transformation over modulated features with parameters inferred from auxiliary data. More specifically, in CBN layers, the extracted features from the previous layers are first normalised to zero mean and unit deviation. Subsequently, the normalised features are modulated using the affine transformation whose scale and shift parameters are learned from the external conditioning data. In the context of image synthesis, CBN enables an image be translated from one domain into another while consistently respecting the constraints specified by conditioning data. Significant improvements in synthesised image quality have been demonstrated for a wide variety of settings (Ulyanov et al., 2017; Huang and Belongie, 2017; Chen et al., 2019).

Compared with other conditional normalisation that employ external information to modulate the normalised features, our CBN scheme uses the network's own input data and thus is closely related to the self-modulated conditional normalisation approach presented in (Chen et al., 2019). The main difference is that they modulate the features as a function of input latent vector z of the generator (e.g., random noise in unconditional case), whereas we condition the normalisation based on the CMR slices $\mathbf{X}$. More formally, in a standard batch normalisation setting, given input feature batch $h_{b,w,h,c} \in \mathbb{R}^{B\times W\times H\times C}$ ($b \in B, w \in W, h \in H$, and $c \in C$ denote the batch size, width, height, and channel of the feature map, respectively) is normalised in a channel-wise manner as follows (Ioffe and Szegedy, 2015):

$$h'_{b,w,h,c,} = \gamma_c \times \frac{h_{b,w,h,c} - \mu_c}{\sigma_c + \epsilon} + \beta_c, \tag{5}$$

with

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

where ϵ is a small number to avoid division by zero, $N = B \times W \times H$, γ and β are learnable scale and shift parameters during the training phase. Note that $\gamma, \beta \in \mathbb{R}^C$ are spatially independent for each feature channel. In our CBN configuration, both parameters γ and β are replaced by spatial dimension-dependent functions as:

$$h'_{b,w,h,c,} = \gamma_{w,h,c}\left(\mathbf{X}\right) \times \frac{h_{b,w,h,c} - \mu_c}{\sigma_c + \epsilon} + \beta_{w,h,c}\left(\mathbf{X}\right). \tag{7}$$

In this work, we model both functions using a standard CNN structure with empirically adjusted layer depths, depending on the number of adjacent slices used for conditioning. The modulation parameters of all CBN layers within the generator are learned simultaneously through the GAN training.

2.5. Multi-scale Discriminator

The discriminator of I2-GAN takes the generated samples and the real images in the target CMR volume as inputs. In this work, we implement a multi-scale discriminator (Durugkar et al., 2017; Nguyen et al., 2017; Wang et al., 2018) that operates at different image scales, as shown in Fig. 2. The real and synthesised images are downsampled by a factor of 2 and used as inputs to the second discriminator, and the network is trained in a multi-task fashion. The benefits of applying this two-scale image pyramid structure are: the discriminator with the larger receptive field yields a global view of the image and can guide the generator to synthesise globally coherent images, whereas the other discriminator encourages the generator to capture finer details. Batch normalisation and LeakyReLU with slope 0.2 are applied for all the layers in the discriminator.

2.6. Optimisation

Instead of using pixel-wise loss in image-space, which struggles with capturing high-frequency details, a feature matching loss (Salimans et al., 2016; Wang et al., 2018) is employed to optimise the GAN to match the statistics of feature representations in multiple intermediate layers of D. Thus, our final joint objective combines both the adversarial loss and feature matching loss:

$$\mathcal{L}_{I2} = \min_{G}\left(\left(\max_{D_1,D_2}\sum_{k=1,2} -\left(\mathbb{E}_x\left[\left(D_k\left(x\right)-1\right)^2\right] + \mathbb{E}_f\left[D_k\left(G\left(f\right)\right)^2\right]\right)\right) + \lambda\sum_{k=1,2}\mathbb{E}_x\sum_{i=1}^{T}\frac{1}{N_i}\left[\left\|D_k^i\left(x\right) - D_k^i\left(G\left(f\right)\right)\right\|_1\right]\right) \tag{8}$$

where i means the ith layer features in D, N_i is the number of features in each layer, T is total number of layers and λ controls

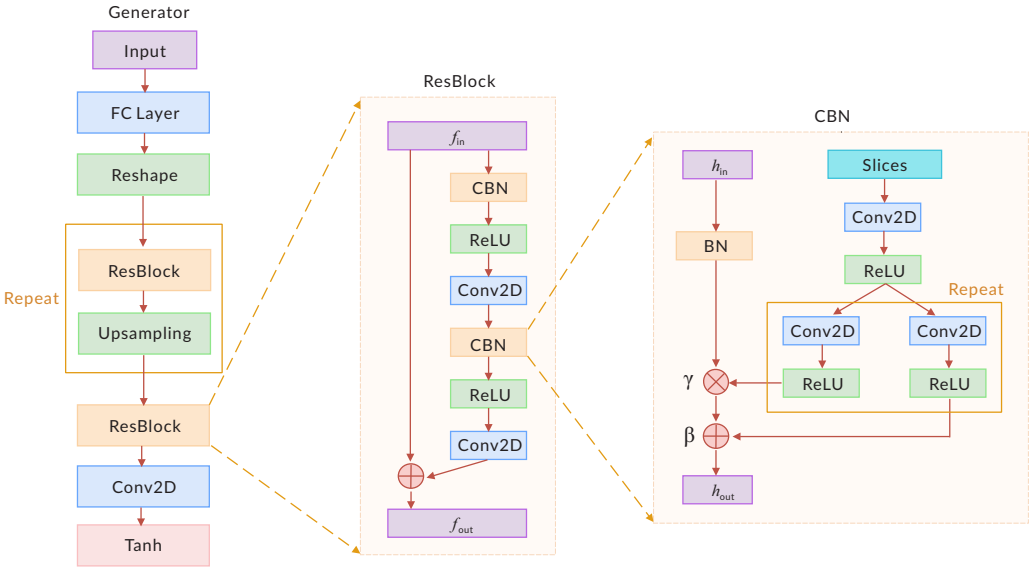

Fig. 3. Structure of the proposed generator. The generator follows a full pre-activation ResNet architecture that consists of residual blocks, followed by nearest neighbour upsampling layers. Conditional batch normalisation (CBN) is implemented in each ResBlock so feature maps are first normalised to zero mean and unit deviation, followed by modulation/de-normalisation using a learned transformation whose parameters are inferred from adjacent CMR slices.

the relative weighting of the feature matching loss to the adversarial loss. The conditioned G and D nets can be optimised by $\mathcal{L}_{I2}$ to infer the missing features from query input volumes.

3. Experimental Setup

3.1. Dataset Description

CMR images from the UK Biobank (UKBB) were used to train and validate the proposed I2-GAN method. CMR images were acquired using a clinical wide bore 1.5T system (MAGNETOM Aera, Syngo Platform VD13A, Siemens Healthcare, Erlangen, Germany) equipped with 18 channels anterior body surface coil (45 mT/m and 200 T/m/s gradient system). 2D cine balanced steady-state free precession (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. Each volumetric sequence contains 50 cardiac phases. Further acquisition details can be found in (Petersen et al., 2015).

To evaluate the generalisation and robustness of our approach, we also investigate the performance of the trained generative model on the Automatic Cardiac Diagnosis Challenge (ACDC) dataset (Bernard et al., 2018) [2]. The publicly available dataset comprises images of 100 subjects acquired with two MRI scanners of different magnetic strengths (1.5 T - Siemens Area, Siemens Medical Solutions, Germany and 3.0 T - Siemens Trio Tim, Siemens Medical Solutions, Germany). Cine MR images were acquired with a conventional SSFP sequence with slice thickness range from 5 mm to 10 mm and sometimes an inter-slice gap of 5 mm. The in-plane spatial resolution varies from 1.34 to 1.68 mm. Each sequence contains 28-40 cardiac phases covering completely or partially one cardiac cycle.

In this work, we focus on the SAX b-SSFP cine CMR datasets for which the ground-truth slices are available and randomly remove one slice to generate incomplete volumes, before using our I2-GAN to synthesise the missed slices. All the 2D slice images were resized to 256×256. The number of slices in the SAX stack typically ranges between 8 and 14. Therefore, we perform a zero-padding on incomplete volumes in the through-plane direction to ensure the input size is consistently $256 \times 256 \times 16$ for the regression net.

3.2. Network Training

The network training procedure for I2-GAN comprises two sequential phases: the individual training of the 3D regression net for extracting the intensity features, and the training of the conditional GAN for image synthesis. In the first phase, the 3D CNN network training was performed for 30 epochs via the minibatch stochastic gradient descent (SGD) and the Adam optimiser, where the learning rate was initially set to 1×10^{-4}. While training the regression network, we set the decay rates of the first and the second momentum of gradient estimates to 0.9 and 0.999, respectively. The mean squared error (MSE) loss was used for the regression problem.

In the second phase, our approach adopted a least-squares GAN (refer to Eq. (2)), and was also trained using 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 estimates were set to 0.5 and 0.999, respectively. In all experiments, the relative weighting of the feature matching loss to GAN loss in Eq. (8) was empirically set to 10. Note that in the generator, instead of performing standard batch normalisation, we first normalised all intermediate feature maps to zero mean and unit deviation, followed by

[2] https://www.creatis.insa-lyon.fr/Challenge/acdc/

Table 1. Summary of 5-fold cross validation results over 4848 CMR datasets from UKBB within a single cardiac phase.

Metrics	Valid Set I (n=970)	Valid Set II (n=970)	Valid Set III (n=970)	Valid Set IV (n=970)	Valid Set V (n=970)
SSIM	0.865 ± 0.028	**0.879 ± 0.025**	0.871 ± 0.025	0.872 ± 0.027	0.868 ± 0.029
PSNR	26.05 ± 1.65	**26.94 ± 1.71**	26.53 ± 1.57	26.88 ± 1.63	26.83 ± 1.40
CC (Regression Net)	**0.994**	0.993	0.991	0.994	0.990

de-normalising/modulating features using a learned transformation whose parameters are inferred from input CMR slices. For the MSI problem, we intuitively used the two adjacent slices as input of Eq. (7) to ensure the normalisation is spatially dependent.

To validate the training stability and reduce bias and variance, we performed 5-fold cross validation over 4848 CMR datasets from UKBB, implying each time, one subset of 970 subjects was used as the validation set, and the other four subsets formed the training set. A summary of the 5-fold cross validation results is presented in Table 1.

3.3. Competing Methods

To demonstrate the advantages of the proposed method, we compared it to two conventional intensity-based and registration-based interpolation methods for slice imputation. The intensity-based method simply computes the linearly weighted average of two adjacent slices and is thus referred to as mean imputation in the rest of the paper. The registration-based slice interpolation approach (Horváth et al., 2017) first uses a symmetric similarity measure to perform structure registration, to calculate displacement fields between neighbouring slices. Then, along every correspondence point trajectory, the displacement fields are utilised to calculate a high order intensity interpolating spline for structural motion. The algorithm was executed using the recommended parameter settings described in their work.

We also compared our method to the standard pix2pix framework (Isola et al., 2017), which comprises an image-conditional GAN architecture that utilises the adversarial and pixel-wise losses, for image-to-image translation tasks. The method takes a set of pairs of corresponding images for training and aims to model the conditional distribution of real images given the input images through minimax optimisation. The pix2pix implementation in this work follows the optimal setup described in their paper that adopts U-Net (Ronneberger et al., 2015) as the generator and a patch-based fully convolutional network as the discriminator. Both the generator and discriminator were trained using the Adam optimiser and a learning rate of 2×10^{-4}, where the first momentum of gradient estimates was selected as 0.5. The network was trained for 100 epochs and the weighting of the L1 pixel-wise loss to adversarial loss was set to 100.

3.4. Evaluation Design

We conducted several experiments to assess the accuracy and robustness of the proposed method. The first experiment (Group 1) was performed by evaluating the quality of the images generated by I2-GAN on a pair of training and testing datasets comprising 4848 subjects within a single cardiac phase. To quantitatively assess synthesis performance, the structural similarity index measurement (SSIM) (Wang et al., 2004) and the peak signal-to-noise ratio (PSNR) were used to measure the image quality of the synthesised CMR images, relative to the original images. The results obtained using our approach were compared with those of the other approaches investigated.

In the second experiment group (Group 2), we evaluated the I2-GAN trained on volumes from one cardiac phase from UKBB and then validated on a different test set comprising 5000 image volumes (264 subjects) extracted from 19 other cardiac time frames, ranging from the end-diastolic (ED) phase to the end-systolic (ES) phase. Since cardiac shape and correspondingly, its morphological indices such as LV/RV volume and LV mass differ remarkably from phase to phase, this experiment was aimed at investigating the generalisation capability of the proposed method, across the full cardiac cycle.

In addition, to assess the impact of MSI in real clinical applications, such as measurements of cardiac function based on blood volumes, we designed an experiment where incomplete data was simulated and volumetric differences between the original (full) and imputed cardiac image volumes were quantified. Measurements of LV volume, RV volume, and LV mass were calculated using the automated LV/RV segmentation and quantification pipeline proposed in (Attar et al., 2019). Datasets from both Group 1 and Group 2 were used for this evaluation, to assess whether there were any statistically significant differences between cardiac functional (volumetric) indices extracted from ground truth and imputed volumes. Pearson's correlation coefficient, linear regression, and Bland-Altman analyses were used to evaluate the correlation and agreement between these cardiac functional measurements.

4. Results

4.1. Intra-Phase Evaluation

First, a qualitative comparison of the synthesised slices is depicted in Fig. 4, between the mean imputation, registration-based interpolation, pix2pix GAN and the proposed I2-GAN approach for 2 subjects of the UKBB dataset in the ED phase. Simple mean imputation yields uncertain and highly blurred edges in the interpolated slices. Results of the registration-based method are marginally better than linear interpolation. However, large dissimilarity between adjacent CMR slices

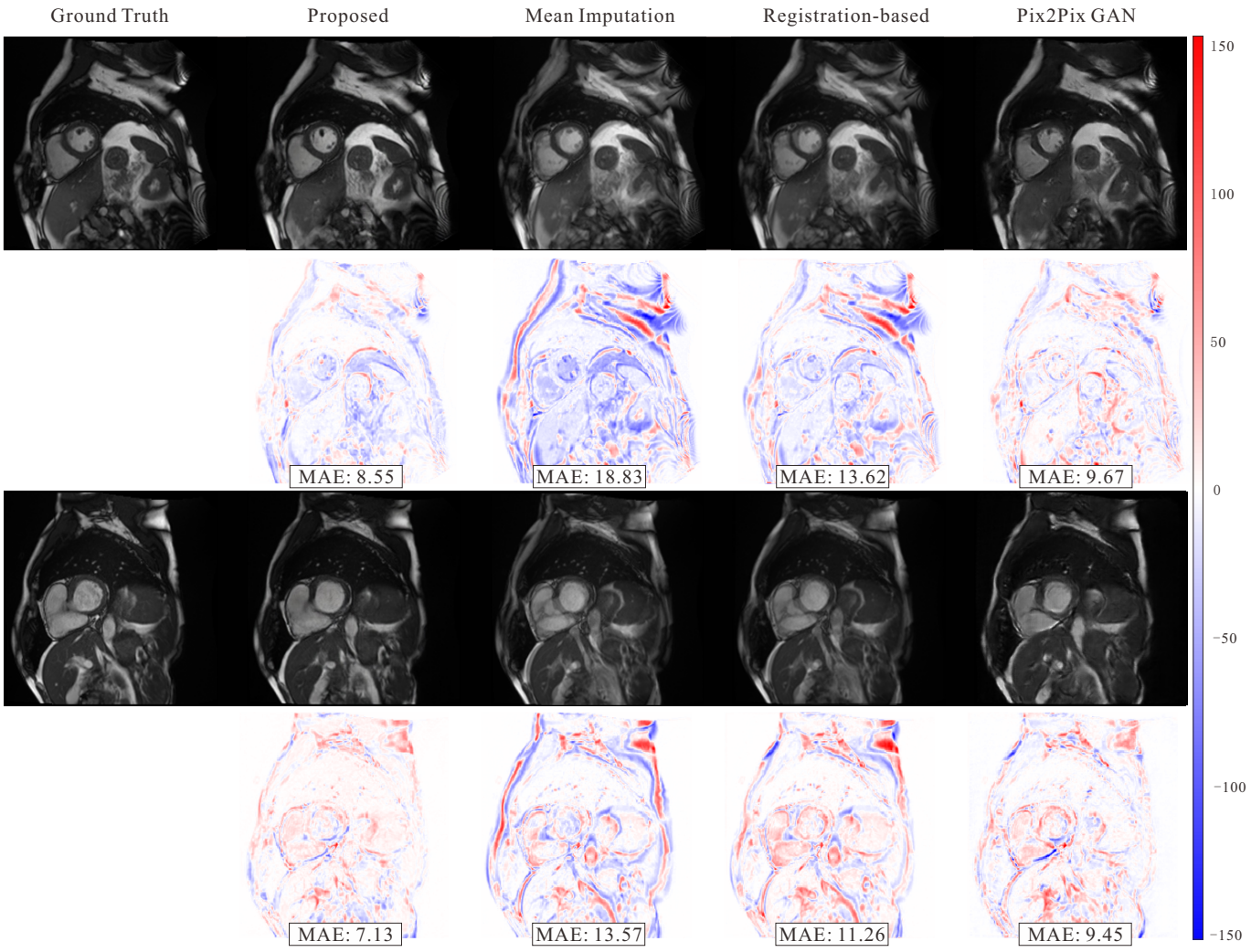

Fig. 4. Qualitative comparison of the ground truth and the synthesised slices between the mean imputation, registration-based interpolation, pix2pix GAN and the proposed I2-GAN approach for 2 subjects from the UKBB dataset in the ED phase (intra-phase). The second and fourth rows show the error image, with the original missed slice at the same position as the ground truth.

yields false correspondence maps during registration, and consequently leads to incorrect interpolation results. By comparison, the slices generated with pix2pix GAN are realistic with sharp edges. As illustrated in Fig. 4, the method reasonably restores anatomical structures that are almost entirely blurred in the interpolation methods, such as LV/RV blood pool and LV myocardium by learning about these structures from the neighbouring slice features. However, the pix2pix method lacks fine anatomical details in areas of soft tissue and myocardium borders and contains some texture artefacts. Our approach generates the most visually comparable result to the reference CMR slices and yields more plausible results in terms of preserving fine structural details and realistic textures in synthesised slices. The observed differences in appearance between the synthesised slices, using each approach, is further highlighted by the corresponding error images in Fig. 4.

As depicted in Fig. 5, similar observations are obtained upon closer investigation of synthesised slices of the cardiac structures. We chose different slice positions from apex to base, in order to assess the capability of the proposed approach to restore the missing slice at different positions within the heart. Again, our method can faithfully approximate the ground truth CMR slices for LV/RV blood pool and LV myocardium and shows the best image quality amongst all methods. Note that the manual segmentation of LV endocardium (LV endo), LV myocardium (LV myo), and RV endocardium (RV endo) of the same UKBB datasets from (Petersen et al., 2017) are also shown in the first column of Fig. 5, to help illustrate the preservation of cardiac tissue boundaries in the synthesised slices.

An overview of the quantitative comparison of different methods on 970 subjects from a single cardiac phase (ED phase) is presented in Fig. 6. The SSIM and PSNR metrics were computed over the entire CMR slice, and over pixels corresponding to three segmented cardiac regions of interest (ROIs) (refer to Fig. 5) using the manual reference masks produced in (Petersen et al., 2017). Consistent with qualitative analysis, the proposed method significantly outperforms the other three methods in terms of the quality of the synthesised missing slices, which is reflected in the average SSIM values for the entire slice: 0.872 ± 0.027 for the proposed method;

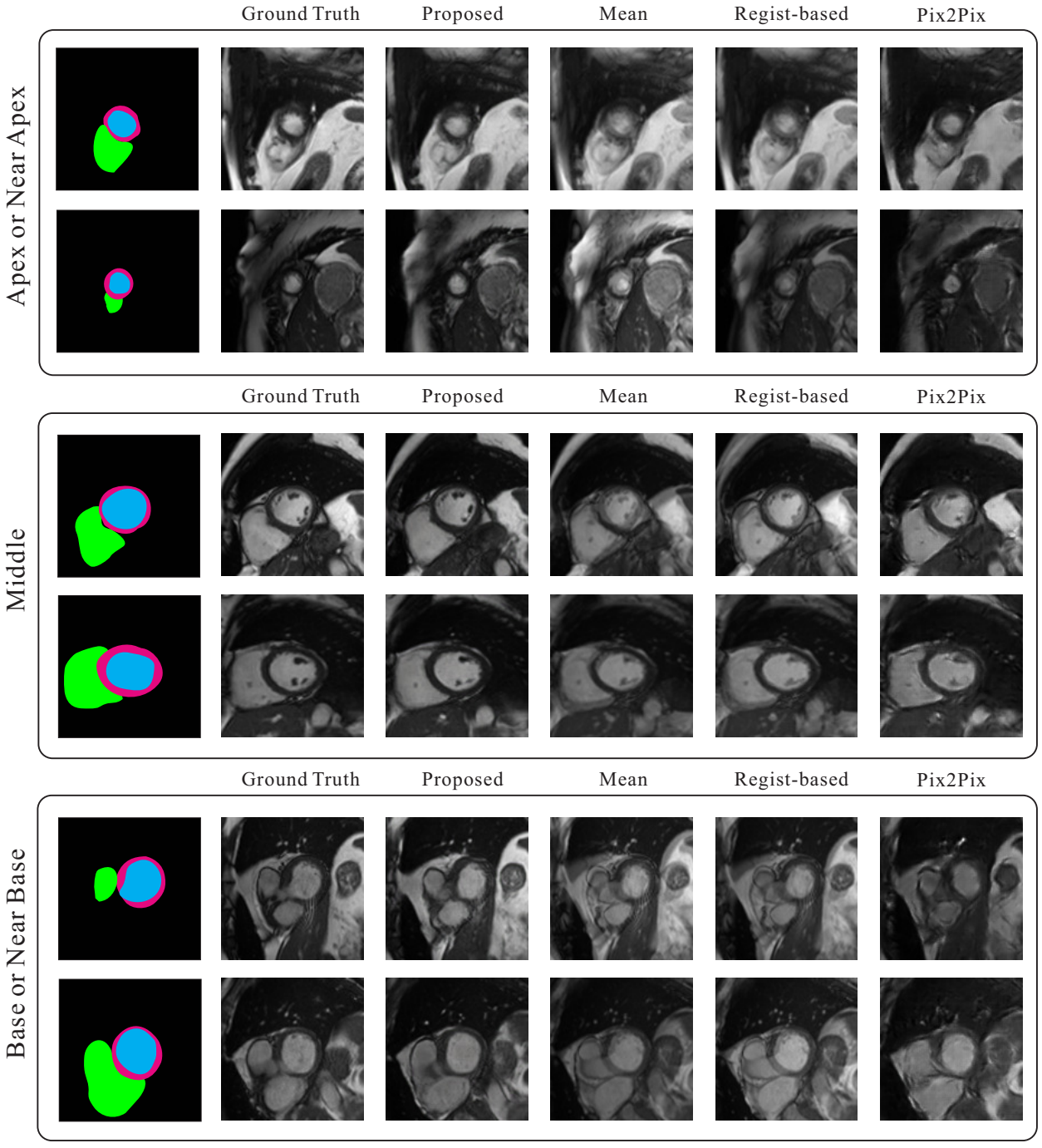

Fig. 5. Close-up for visual comparison of ground truth and synthesised CMR image slices. We compared mean slice imputation, registration-based interpolation, pix2pix GAN and proposed I2-GAN approach, for 6 subjects from the UKBB dataset in the ED phase (intra-phase). The first column depicts the manual segmentation of LV endocardium (blue), LV myocardium (purple), and RV endocardium (green) from (Petersen et al., 2017). Different slice positions from apex to base are chosen and shown.

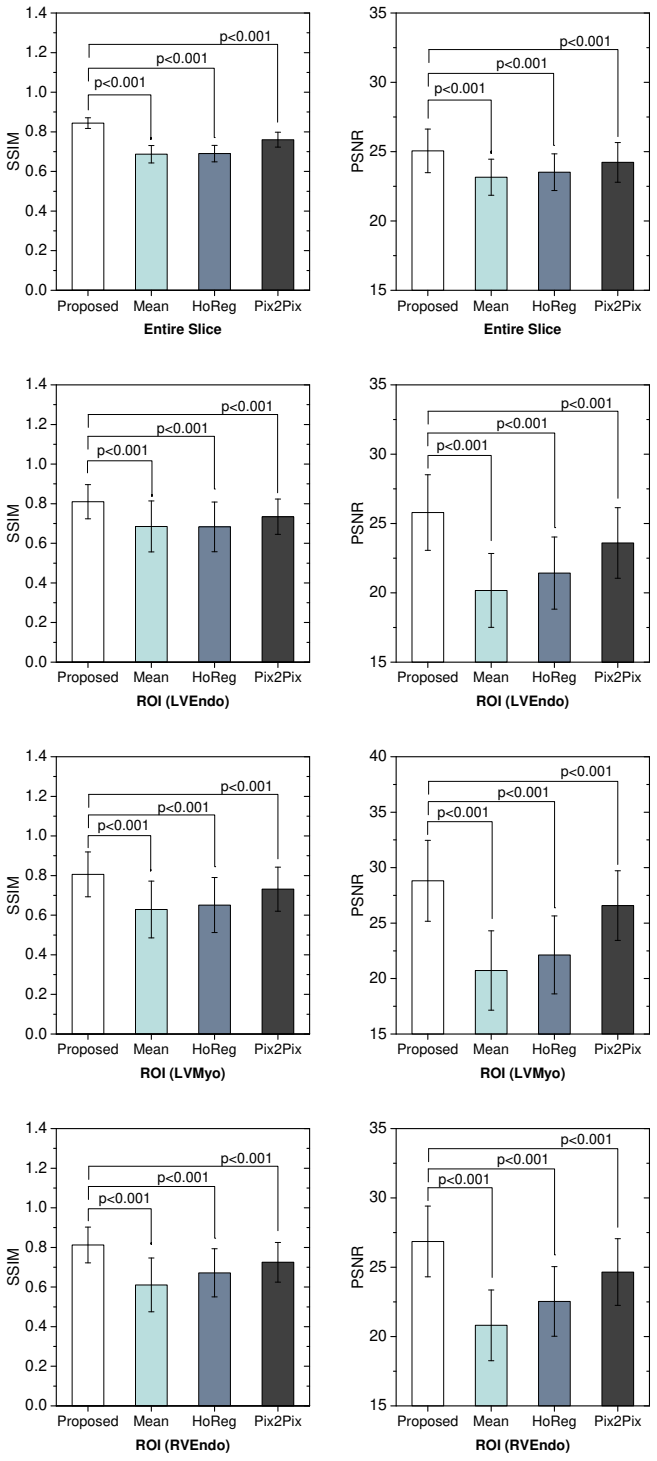

Fig. 6. SSIM and PSNR measurements between the ground truth and synthesised target images from the mean imputation, registration-based interpolation, pix2pix GAN and the proposed I2-GAN approach on 970 subjects from a single cardiac phase (ED phase). The Wilcoxon signed rank test is used for all comparisons (i.e. proposed vs. mean, proposed vs. HoReg and proposed vs. pix2pix) to assess the statistical significance of the results. LVEndo, LVMyo and RVEndo represent LV endocardium, LV myocardium and RV endocardium, respectively.

0.686 ± 0.044 for mean imputation; 0.691 ± 0.041 for registration based interpolation; and 0.790 ± 0.037 for pix2pix GAN. The computed PSNR values are 26.88 ± 1.63 for the proposed method, 23.15 ± 1.30 for mean imputation, 23.51 ± 1.32 for registration-based interpolation and 24.72 ± 1.44 for pix2pix GAN. The nonparametric statistical Wilcoxon signed rank test that does not assume the data to be normally distributed, was used for all comparisons (i.e. proposed vs. mean, proposed vs. registration-based interpolation and proposed vs. pix2pix), indicating that the proposed approach achieved statistically significant improvements over its counterparts, considering a significance level of (p<0.001).

We also investigate the effects of slice position estimation on the results in two scenarios: first, missing slice position is correctly predicted from regression and thus features extracted serve as accurate conditional input to the generative model; second, the estimated slice position is shifted by ±1 relative to the actual missing position, but auxiliary slice information for CBN layers is accurate by knowing the real missing or corrupted slice position beforehand from a previous image quality control step. As discussed previously, the primary objective of this study is to facilitate the imputation of a missing slice, given its position in the SAX stack. Fig. 7 shows the qualitive and quantitative comparison of these two scenarios over 100 subjects on ED phase. We found a mild degradation in the synthesis image quality with inaccurate slice position estimation (a mean SSIM of 0.874 vs. 0.856).

4.2. Inter-Phase Evaluation

Here we present the evaluation results of the proposed method trained on volumes from one cardiac phase (ED) but tested on 5000 volumes selected from 19 other cardiac phases (spanning from ED to ES), for 264 unseen test subjects. We follow the same outline as before, by first visually assessing the synthesised results, followed by quantitative analysis.

Fig. 8 shows different synthesised slices sequentially from apex to base for a subject of the UKBB dataset from a cardiac phase that corresponds to 36.8% of the ED-ES interval, where incomplete data was simulated by systematically removing the corresponding slice in the investigated volume. We observe that the overall synthesised results are still visually comparable to the reference (shown in the first row) and appear consistent between slices, even though the model was trained using a different cardiac phase. The visual inspection of detailed structure also reveals minor degradation in image quality and the presence of artefacts in some of the generated slices, as illustrated by the white arrows in Fig. 8. Fig. 9 depicts the synthesised slices of three subjects for 4 cardiac time frames from ED to ES phase, which in general are of a similarly comparable image quality, to the reference slices. However, we also find that papillary muscles may not always be reconstructed accurately in other time points, as shown in the bottom of the ventricle of the right-most example (indicated by a white arrow).

The obtained metric values computed within the entire slice quantitatively confirm our visual observations. As depicted in Fig. 10, both SSIM and PSNR values show that there are no significant differences in the quality of synthesised slices, between cardiac phases, although in general the proposed method

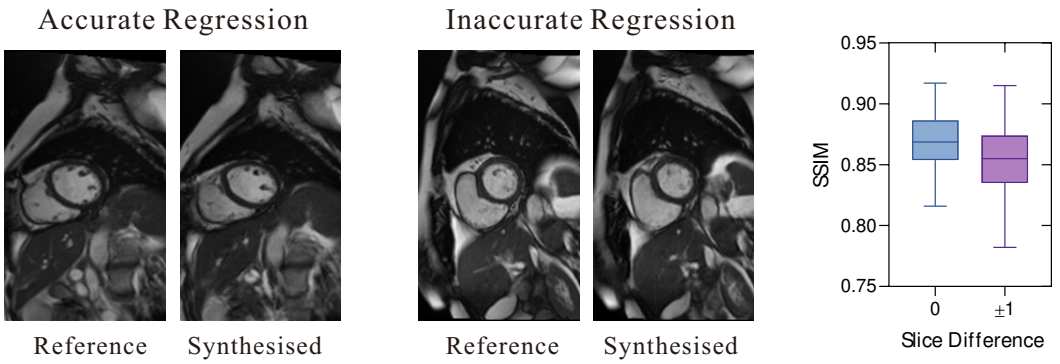

Fig. 7. Visual inspection and quantitative comparison of the impact of accurate regression (correctly predicted missing position) and inaccurate regression (estimated slice position shifted by ±1 compared to the actual one) on the final synthesis quality.

produces marginally inferior results compared to that of intra-phase evaluation (SSIM values: 0.857 ± 0.024 vs. 0.872 ± 0.027 and PSNR values: 25.11 ± 1.48 vs. 26.88 ± 1.63). However, this is to be expected, as we train our network only on ED images.

4.3. ACDC Evaluation Results

We evaluate the proposed generative model trained on the UKBB dataset, on an external publicly available dataset viz. ACDC (Bernard et al., 2018), in order to assess its generalisation capability. Missing slices imputed using the pre-trained I2-GAN for four different subjects from the ACDC dataset, are illustrated in Fig. 11. Here missing slices were imputed for images from cardiac phases, namely, the ED and ES phase, for each subject. White contours in these images represent the delineations of the left ventricular cavity, myocardium, and right ventricle for the reference image based on manual expert segmentations, which are used here to help assess the quality of the proposed approach. From these results, we can observe that although the pre-trained model performs slightly worse compared to that of the UKBB data, it still provides plausible imputation results, as demonstrated by the anatomical details captured within the ventricles and myocardium (such as papillary muscles or myocardial wall).

Fig. 12 shows the accuracy statistics in the ACDC dataset. All metrics were computed over 300 incomplete volumes generated by randomly discarding one slice from the original ED and ES phase data three times with different slice position. Due to high variability of ACDC images found among subject scans, we can see overall SSIM values are worse than those of UKBB data in ED phase (entire slice: 0.821 vs. 0.872, LV: 0.759 vs. 0.797, myocardium: 0.781 vs. 0.790, and RV: 0.758 vs. 0.805, respectively). We did not find substantial differences in accuracy between ED and ES phases. Similarly, our method performs consistently using the PSNR metric, with mean values of 24.36, 22.45, 24.29, 22.12 for entire slice, LV, myocardium, and RV, respectively.

4.4. Cardiac Parameters Calculation

The average values of quantitative cardiac functional (volumetric) indices, estimated using the automated segmentation pipeline proposed in (Attar et al., 2019), are presented in Table 2, along with analysis of the respective differences between reference and I2-GAN synthesised CMR volume. No significant difference was found between the reference and synthesised images concerning all indices, namely, LV volume, LV mass, and RV volume. Note that the intra-phase evaluation yields relatively lower pairwise difference error compared to that of inter-phase evaluation results, with respect to the reference volume. The linear regression analysis showed strong agreement between reference and imputed CMR images for all measurements. As depicted in Fig. 13, the linear regression trendline slopes were higher than 0.99 for both LV volume and RV volume, with correlation coefficients of 0.99 for LV volume and close to 0.97 for RV volume measurements. For intra-phase datasets, Bland-Altman analysis revealed that the mean difference with 95 % confidence interval (CI) between the standard cine CMR and the I2-GAN synthesised CMR were: 0.02 mL (95 % CI, −8.31 mL to 8.35 mL) for LV volume; 0.61 g (95 % CI, −10.78 g to 12.0 g) for LV mass; and 0.86 mL (95 % CI, −26.18 mL to 27.9 mL) for RV volume (also see Fig. 13).

A similar result was also obtained from inter-phase statistical analysis, with the mean differences with 95 % CI being 0.18 mL (95 % CI, −11.26 mL to –11.63 mL) for LV volume, 0.56 g (95 % CI, −17.14 g to 18.26 g) for LV mass, and -0.45 mL (95 % CI, −23.82 to 22.92) for RV volume. These statistical analyses emphasise the high quality of the images synthesised using I2-GAN and highlight its potential for large-scale CMR population imaging applications.

4.5. Ablation Study

This section provides additional ablation study results analysing the contribution of the newly adopted techniques in the proposed method. Three variants, namely I2-GAN w/o CBN, w/o CBN + FM and w/o CBN + FM + MD, corresponding to replacing the CBN layers with conventional batch normalisation, replacing feature matching loss with pixel-wise loss and using a single discriminator for I2-GAN progressively, are investigated and compared with the standard full I2-GAN. For fair comparison, we retrain these variant networks using the same hyperparameters and training epochs as the proposed method and use the same metrics for evaluation.

Fig. 15 and Table 3 show the qualitive and quantatitive results of the ablation study. As depicted in Fig. 15, full version of I2-GAN produces anatomical structures that are visually the

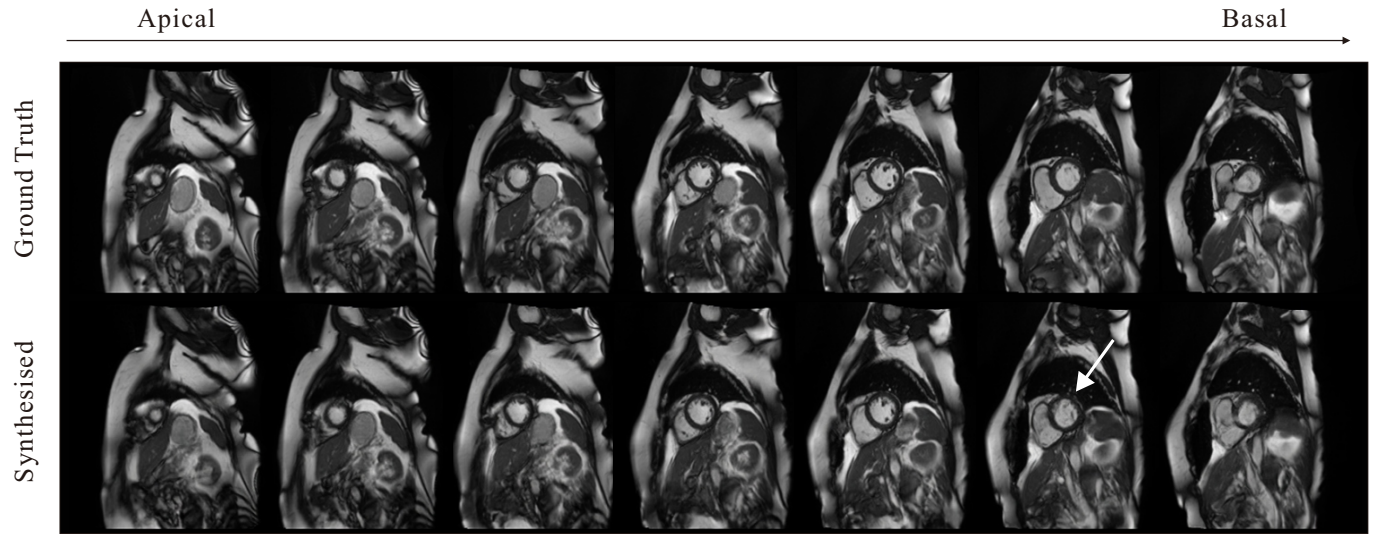

Fig. 8. Qualitative comparison of ground truth and synthesised slices when the I2-GAN is trained on end diastolic (ED) phase and tested on images from a cardiac phase that corresponds to 36.8% of the ED-ES interval. From left to right, different slices from apex to base are shown sequentially. The white arrow indicates the image quality degradation and artefacts in the synthesised slice.

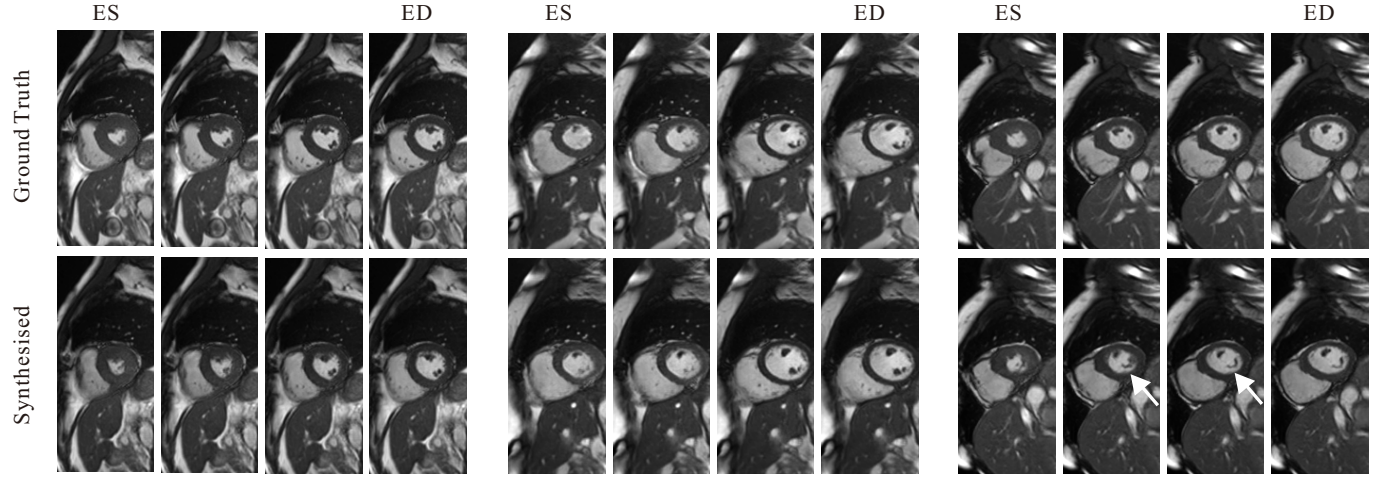

Fig. 9. Qualitative comparison of ground truth and synthesised slices when the I2-GAN is trained on end diastolic (ED) phase and tested on images from different cardiac phases (inter-phase). Three subject examples are shown with slices selected temporally from end-systolic to end-diastolic phase. The white arrow indicates the image quality degradation and artefacts in the synthesised slice.

Table 2. Cardiac volume measurements between the reference volume and I2-GAN imputed volume on the UKBB dataset.

Experiment Group	Parameters	b-SSFP reference images (n=500)	I2-GAN imputed images (n=500)	Pairwise Difference Original vs. Imputed (%)	*p*-value
Intra-Phase	LV volume (mL)	124.76 ± 30.89	124.74 ± 30.82	2.45 ± 3.45	0.992
	LV mass (g)	119.53 ± 26.70	120.14 ± 27.21	3.76 ± 5.01	0.721
	RV volume (mL)	135.62 ± 50.17	134.76 ± 48.44	6.04 ± 8.59	0.783
Inter-Phase	LV volume (mL)	86.53 ± 34.01	86.35 ± 33.92	4.39 ± 6.61	0.932
	LV mass (g)	114.09 ± 28.62	114.65 ± 29.66	4.83 ± 7.35	0.764
	RV volume (mL)	124.08 ± 50.49	124.53 ± 51.71	6.75 ± 9.97	0.891

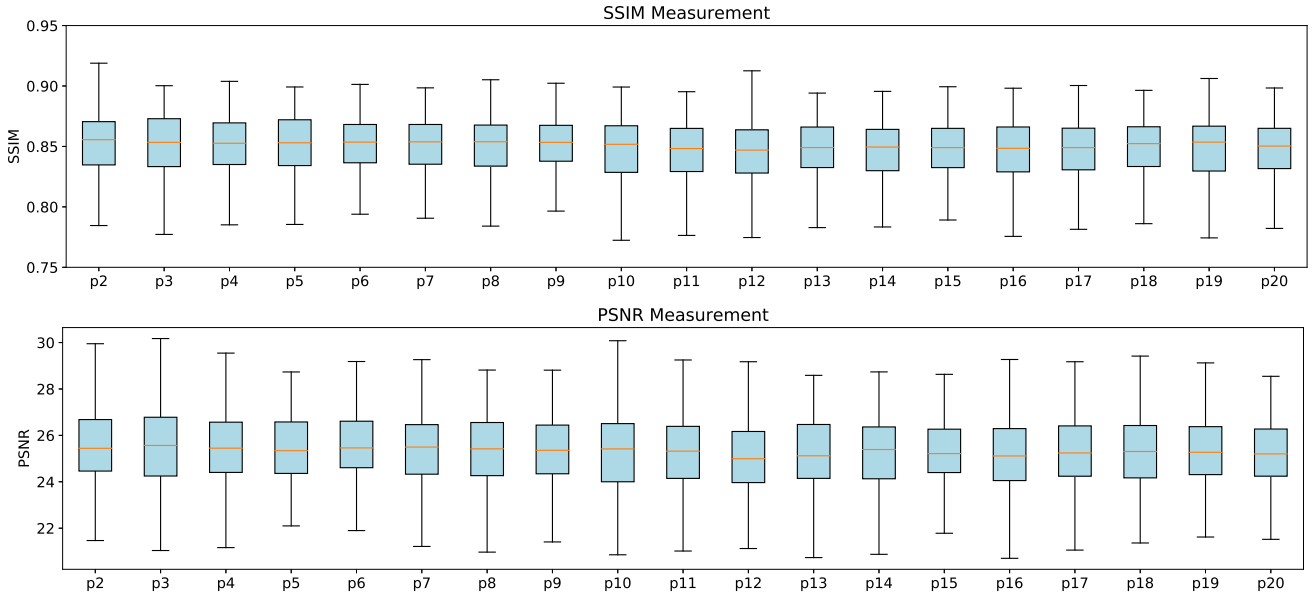

Fig. 10. SSIM and PSNR measurements between the ground truth and synthesised target images from the proposed I2-GAN approach on 5000 data (264 subjects) from UKBB spanning from 19 cardiac time frames (from ED to ES phase).

most similar to ground truth images, such as papillary muscles, myocardial wall and endocardial lumen. The quantitative results indicate that both CBN and discriminator-based feature matching loss yield superior results compared to the conventional batch normalisation and pixel-wise loss term. Removing or replacing them results in a substantial performance decrease (e.g., SSIM of 0.872 ± 0.027 to 0.823 ± 0.028 for whole slice). It also can be seen multi-scale discriminator further contributes to the imputation performance boost, which is reflected by an average increase of ~10% in similarity measurement in ROIs. Furthermore, the Wilcoxon signed rank test was performed to compare the I2-GAN with its variants, indicating that the I2-GAN achieved statistically significant improvements, considering a significance level of (p<0.001).

5. Discussion

Missing slices or presence of various artefacts due to respiratory and patient motion, blurring and signal loss in b-SSFP SAX image stacks, even in a single slice of one cardiac phase, may influence the accuracy of anatomical and functional cardiac measurements and thus often leads to the removal of the entire subject from all subsequent analyses in population imaging applications. In this work, we introduced a new approach that produces realistic and plausible synthesised CMR images with well preserved fine details and texture information. The method is based on a conditional GAN, where the generator is given prior information, for instance the CMR intensity features learned through a dedicated 3D regression network, in our case. Such architectures have been extensively exploited and demonstrated to search in a complex parameter space for a Nash equilibrium through a minimax game, and improve the performance of GANs (Mirza and Osindero, 2014; Odena et al., 2017; Miyato and Koyama, 2018).

One of most popular frameworks for image synthesis tasks is based on the image-to-image translation model (Isola et al., 2017) (often called pix2pix) that uses encoder and decoder stacks with skip connections in between to form the generator, which can also be regarded as an image-conditional GAN. However, many previous studies have shown that adversarial training may be unstable and prone to failure for synthesising images that contain fine details and realistic textures (Johnson et al., 2016; Dosovitskiy and Brox, 2016), which is essential for medical images, as shown in Figs. 4 and 5. Therefore, our main contribution in this work lies in the design of a new, robust GAN architecture to generate CMR images which preserve fine structural details, and enables accurate quantification of anatomical and functional cardiac measurements, compared with those computed by previous methods.

First, our work is motivated by the concept of CBN that incorporates external guidance information (such as labels, embeddings, masks or generator's own inputs) into conditional generative model throughout batch normalisation. CBNs have been adopted in many contexts in computer vision domain such as in image style transfer, visual question answering and visual reasoning (De Vries et al., 2017; Miyato and Koyama, 2018; Zhang et al., 2018a; Chen et al., 2019; Park et al., 2019; Dumoulin et al., 2016) and have been demonstrated to significantly outperform the baseline methods for a wide variety of settings. CBNs can de-normalise the feature maps using spatial constraints (such as auxiliary data or guidance image) and modulate the flow of features during image generation. It utilises spatial conditions for feature-wise manipulation and spatial transformation. The learned modulation parameters encode relevant, correlated spatial features and enable fine structural details and textures to be effectively propagated through the main generation pathway, while respecting certain constraints specific to the input conditioning data. In this work, we intuitively used two

Table 3. Summary of the ablation study results (evaluated in terms of SSIM) over 970 CMR datasets from UKBB within ED cardiac phase.

Methods	Entire Slice n=970	ROI (LVEndo) n=970	ROI (LVMyo) n=970	ROI (RVEndo) n=970
Proposed I2-GAN	**0.872 ± 0.027**	**0.797 ± 0.094**	**0.790 ± 0.124**	**0.805 ± 0.099**
w/o CBN	0.841 ± 0.026	0.733 ± 0.088	0.731 ± 0.118	0.748 ± 0.096
w/o CBN+FM	0.823 ± 0.028	0.713 ± 0.084	0.745 ± 0.103	0.739 ± 0.10
w/o CBN+FM+MD	0.819 ± 0.031	0.711 ± 0.086	0.671 ± 0.120	0.716 ± 0.105

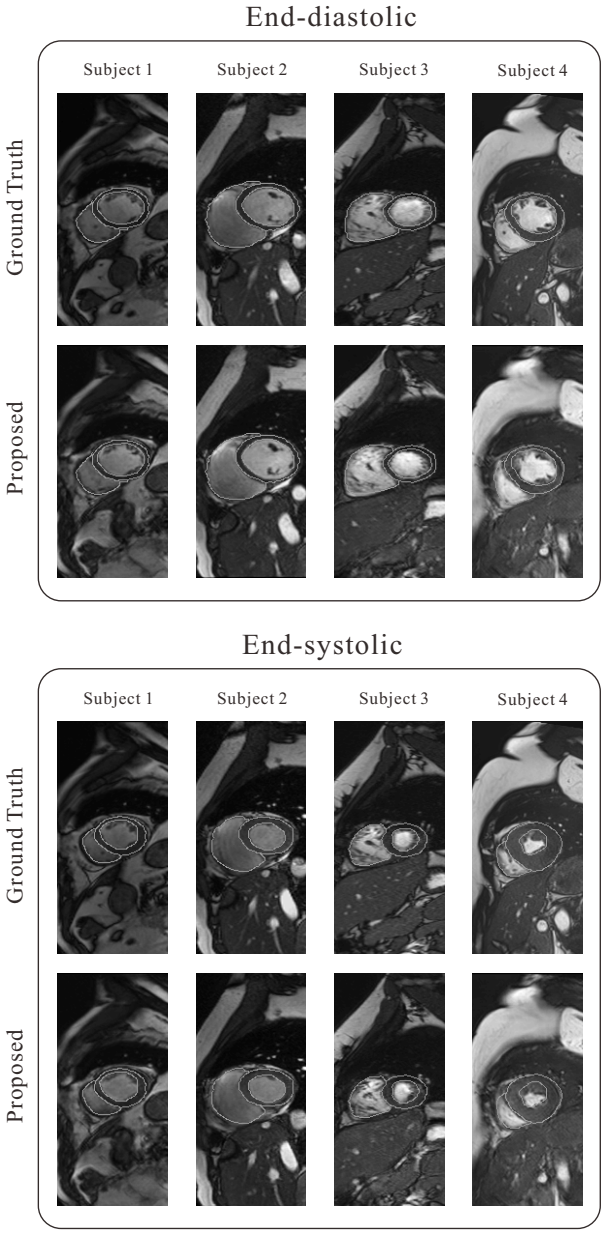

Fig. 11. Visual inspection of the synthesised CMR image slices from 4 subjects of the ACDC dataset with the I2-GAN model trained on the UKBB data. White contours in these displayed images represent the delineations of the left ventricular cavity, myocardium, and right ventricle for the reference image based on manual expert segmentation. Different slice positions are chosen and shown.

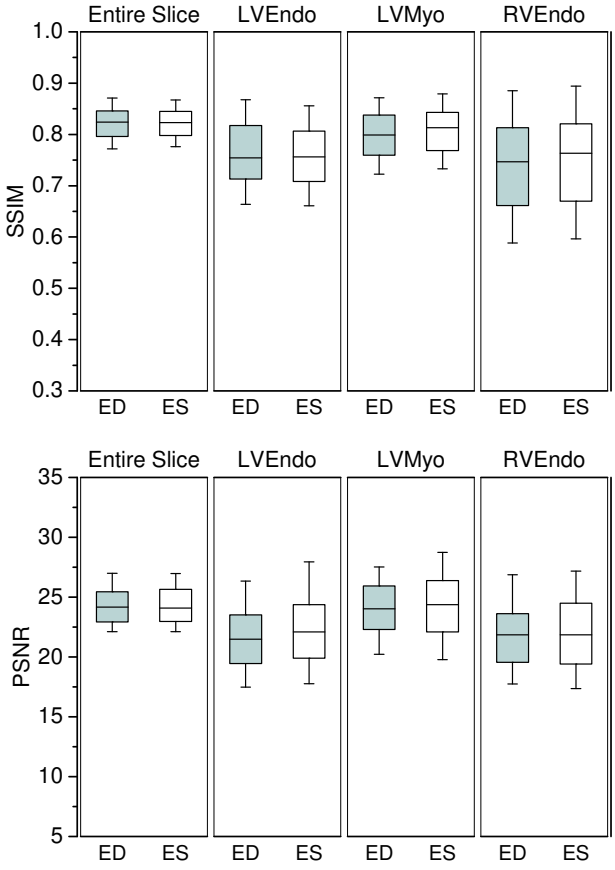

Fig. 12. SSIM and PSNR measurements between the ground truth and synthesised target images from the proposed I2-GAN approach on the ACDC dataset. All metric values were computed over 300 incomplete volumes simulated by randomly discarding one slice from the 100 original ED and ES phase data samples three times. LVEndo, LVMyo and RVEndo represent LV endocardium, LV myocardium and RV endocardium, respectively.

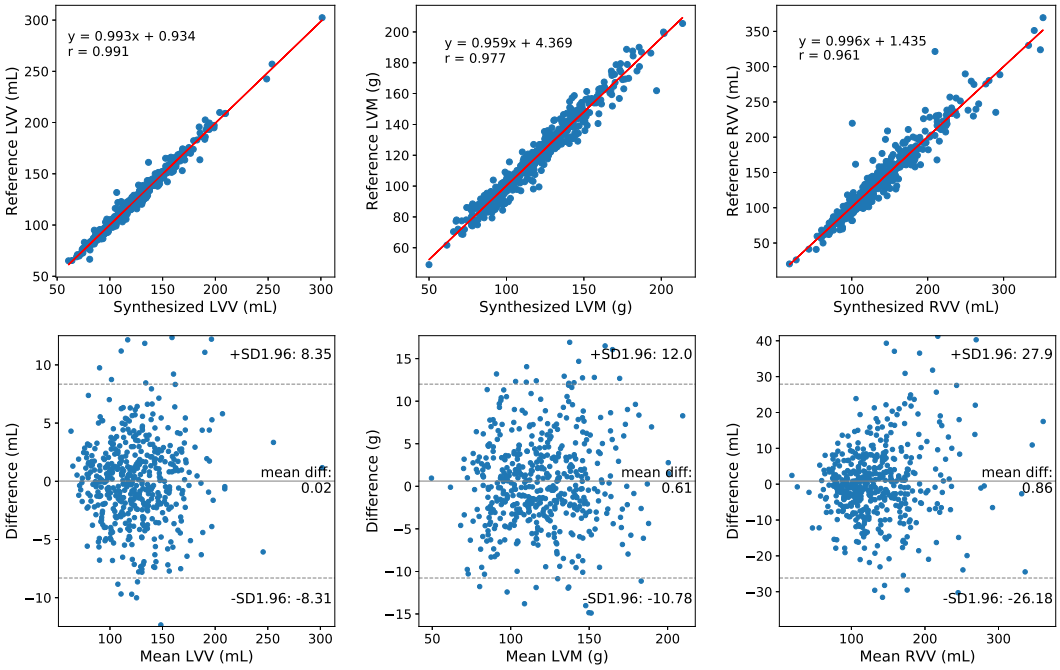

Fig. 13. Linear regression plots and Bland-Altman analysis for LV volume, LV mas, and RV volume measurements using original cine CMR images and I2-GAN imputed cine CMR images for intra-phase evaluation.

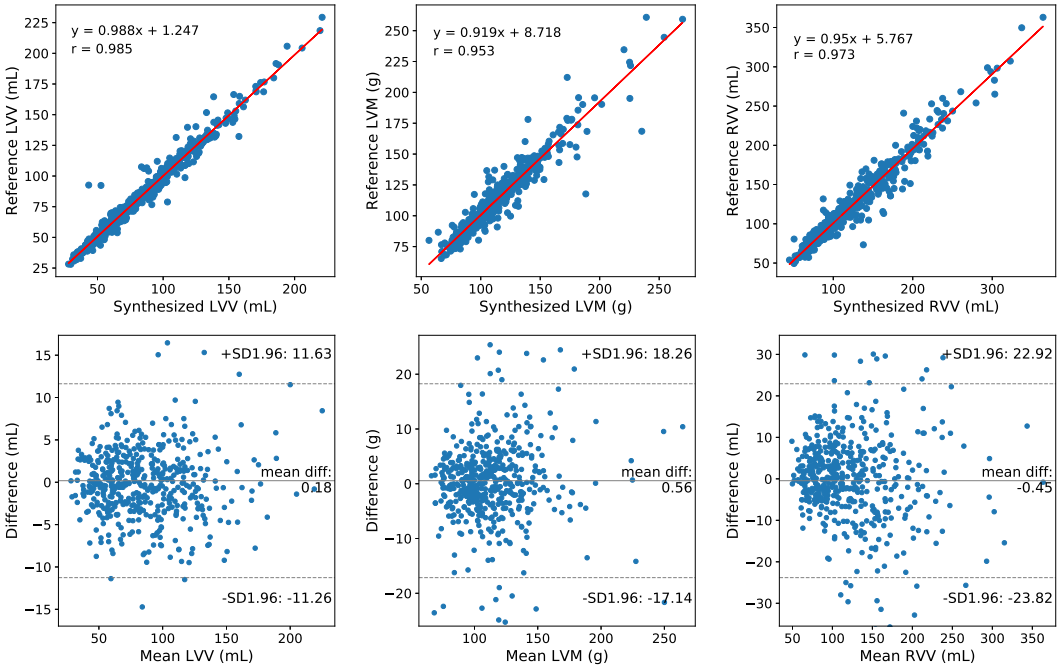

Fig. 14. Linear regression plots and Bland-Altman analysis for LV volume, LV mass, and RV volume measurements using original cine CMR images and I2-GAN imputed cine CMR images for inter-phase evaluation.

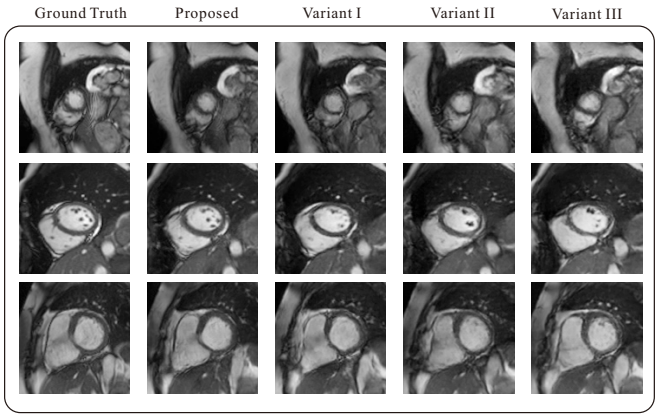

Fig. 15. Visual comparison between the I2-GAN and its three variants in the ablation study for 3 subjects from the UKBB dataset in the ED phase. Variant I, II, and III correspond to I2-GAN w/o CBN, w/o CBN + FM and w/o CBN + FM + MD, respectively. From top to bottom, different slice positions from apex to base are depicted.

adjacent slices as simple but effective complementary guidance, to capture spatial features from relevant regions in the immediate neighbourhood of missing slices, which contributes significantly to the image imputation task, as shown in the ablation study.

We also investigated different types of inputs for the generator, i.e., concatenation of the prior noise with intensity feature and intensity feature vector without random noise, and did not find noticeable differences in the imputation results except minimal texture changes, which indicates that the conditioning features alone are sufficient inputs to the generator.

Second, pixel-wise loss functions struggle to recover high-frequency details such as texture. For instance, minimising mean square error encourages finding pixel-wise averaging of plausible solutions in the original pixel space, which typically results in blurred images and thus yields unsatisfying perceptual quality. To enhance the performance and stabilise the training, we exploit the feature space instead, and employ the discriminator-based feature matching loss combined with adversarial training, as formulated in Eq. (8). The feature matching loss is similar to the perceptual loss proposed in (Johnson et al., 2016; Dosovitskiy and Brox, 2016), and has been applied for image super-resolution (Ledig et al., 2017) and style transfer (Johnson et al., 2016). Feature matching changes the final objective for the generator to minimising the statistical difference between the intermediate feature representations of the real images and the synthesised images. Herein we measured the L1-distance between the means of their feature vectors, and feature matching reformulates the goal from beating the opponent to matching features in real images to obtain fine details and perceptually more convincing results.

In addition, we used a multi-scale discriminator framework that operates at different image scales. The discriminator with the larger receptive field yields a global viewing of the image and guides the generator to synthesise globally consistent images, whereas the discriminator at the small scale encourages the generator to produce finer structural details. This dedicated multi-scale GAN architecture further boosts our image synthesis performance.

We first demonstrated the proposed imputation algorithm on UKBB datasets in two groups of experimental settings and evaluated the results both visually and quantitatively. The set of experiments involves evaluating the quality of the images generated by I2-GAN on the datasets consisting 4848 subjects within a single cardiac phase. We also compared our imputation method with several conventional and state-of-the-art DL methods, namely mean imputation, registration-based interpolation and pix2pix GAN and showed that the proposed method consistently outperformed the other methods by a large margin. The traditional intensity-based and object-based interpolation methods rely strongly on the assumption that the consecutive slices have similar anatomical structure as they restrict their search for finding correspondence points to small neighbourhoods. This assumption is often weak in CMR datasets due to significantly large inter-slice spacing. Besides, the conventional approaches do not leverage statistical information to guide the imputation. These methods thus yielded unsatisfying interpolation results, as expected. The pix2pix method achieved superior results over the interpolation methods, but performed worse than our approach in terms of fine anatomical details and texture. Our model generated more visually realistic and plausible CMR images. The second experiment group was aimed at assessing the robustness of I2-GAN on a different testing set, consisting of cases extracted from 19 other cardiac time frames that range from ED to ES phases, and therefore quantifying its generalisation ability. The proposed imputation method still yielded encouraging results.

We further investigate the impact of slice position estimation in two scenarios in the section 4.1. Our goal is to perform image imputation following manual, semi-automatic, or automated image QA. Scenario 2 thus assumes that the correct positions of missing CMR slices are available *a priori*. This is a limitation of the current approach and mechanisms to mitigate incorrect identification of missing slice positions, and its impact on the overall quality/accuracy of imputed slices will be the subject of future work.

To evaluate the robustness of I2-GAN, we further investigated the preformance of the proposed method on the 100 publicly available ACDC dataset. Even though higher variability of image quality, resolution and structural appearance among these subject scans, the pre-trained model still provides reasonable imputation results.

The conditional normalisation scheme used in this work incorporates two adjacent slices into the image synthesis pathway to modulate the computation flow in the network. All modulation parameters in CBN were learned through a shallow convolutional network embedded in the batch normalisation layers, as shown in Fig. 3. It would be interesting to investigate whether the model benefits from deeper modulation networks or more complex architectures. For instance, we could incorporate the whole image stack into the image imputation network through CBNs with a dedicated attention weighting or a gating mechanism (Oktay et al., 2018; Zhang et al., 2018a), to leverage more useful inter-slice context and additional 3D anatomical information for synthesising globally coherent images. Investigation

in this direction will be the subject of future work.

6. Conclusion

The development of robust and generic techniques for missing data imputation can have a transformative impact on population imaging applications by preventing incomplete data from being completely disregarded, when analysing any given cohort. In this work, we proposed a novel, robust GAN architecture to impute missing slices in CMR images, namely I2-GAN. The approach adopts a slice position regression model and an adversarial training architecture to impute missing slices. Experimental results showed that our model consistently outperformed traditional and state-of-the-art imputation techniques. The statistical analyses highlighted a strong correlation and agreement between key cardiac functional indices measured using the original and imputed image volumes. This indicates that the proposed framework enables accurate cardiac quantification despite missing information in CMR image volumes. Although we only focused on the MSI problem, which assumes the absence of just one slice, the I2-GAN can be easily extended to multi-slice scenarios, particularly for correcting incomplete cardiac coverage involving missing slices at the apical or basal positions. This can be achieved by leveraging additional inter-slice context, information from the full stack of slices or deriving richer, multi-view descriptors from LAX slices. The sensitivity of the model to larger proportions of missing data is yet to be assessed and could be another direction for future work. We also plan to extend the I2-GAN for both intra-phase (spatial) and inter-phase (temporal) super-sampling, in order to facilitate quantitative analysis of cardiac deformation across the whole cardiac cycle. The former in particular would enable the synthesis of isotropic CMR volumes – a common challenge limiting the accuracy of cardiac image registration algorithms. These avenues for future work highlight the versatility of the proposed framework and its relevance as a tool that facilitates quantitative cardiac function analysis in large population studies.

Acknowledgements

This research has been conducted using the UK Biobank Resource under Applications 11350 and 2964. The CMR images presented in Figs. 1, 2, 4, 5, 7-9 and 15 in the manuscript were reproduced with the permission of UK Biobank©. The authors are grateful to all UK Biobank participants and staff. AFF acknowledges support from the Royal Academy of Engineering Chair in Emerging Technologies Scheme (CiET1819/19), EPSRC-funded Grow MedTech CardioX (POC041), and the MedIAN Network (EP/N026993/1) funded by the Engineering and Physical Sciences Research Council (EPSRC). SKP and SN acknowledge the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre based at The Oxford University Hospitals Trust at the University of Oxford, and the British Heart Foundation Centre of Research Excellence. SEP acknowledges support from the NIHR Barts Biomedical Research Centre and from the SmartHeart EPSRC Programme Grant (EP/P0010 09/1).

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., nad 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, 650.

Attar, R., Pereañez, M., Gooya, A., Albà, X., Zhang, L., de Vila, M.H., Lee, A.M., Aung, N., Lukaschuk, E., Sanghvi, M.M., Fung, K., Paiva, J.M., Piechnik, S.K., Neubauer, S., Petersen, S.E., Frangi, A.F., 2019. Quantitative CMR population imaging on 20,000 subjects of the UK Biobank imaging study: LV/RV quantification pipeline and its evaluation. Medical image analysis 56, 26–42.

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, 206–220.

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al., 2018. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging 37, 2514–2525.

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

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.

Dong, Y., Peng, C.Y.J., 2013. Principled missing data methods for researchers. SpringerPlus 2, 222.

Dosovitskiy, A., Brox, T., 2016. Generating images with perceptual similarity metrics based on deep networks, in: Advances in neural information processing systems, pp. 658–666.

Dumoulin, V., Shlens, J., Kudlur, M., 2016. A learned representation for artistic style, in: International Conference on Learning Representations (ICLR).

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, 41.

Frakes, D.H., Dasi, L.P., Pekkan, K., Kitajima, H.D., Sundareswaran, K., Yoganathan, A.P., Smith, M.J., 2008. A new method for registration-based medical image interpolation. IEEE transactions on medical imaging 27, 370–377.

Frangi, A.F., Niessen, W.J., Viergever, M.A., 2001. Three-dimensional modeling for functional analysis of cardiac images, a review. IEEE transactions on medical imaging 20, 2–5.

García-Laencina, P.J., Sancho-Gómez, J.L., Figueiras-Vidal, A.R., 2010. Pattern classification with missing data: a review. Neural Computing and Applications 19, 263–282.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Advances in neural information processing systems, pp. 2672–2680.

Van der Graaf, A., Bhagirath, P., Ghoerbien, S., Götte, M., 2014. Cardiac magnetic resonance imaging: artefacts for clinicians. Netherlands Heart Journal 22, 542–549.

Grevera, G.J., Udupa, J.K., 1998. An objective comparison of 3-D image interpolation methods. IEEE transactions on medical imaging 17, 642–652.

Grevera, G.J., Udupa, J.K., Miki, Y., 1999. A task-specific evaluation of three-dimensional image interpolation techniques. IEEE transactions on medical imaging 18, 137–143.

Han, Z., Wei, B., Mercado, A., Leung, S., Li, S., 2018. Spine-GAN: Semantic segmentation of multiple spinal structures. Medical image analysis 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.

Horváth, A., Pezold, S., Weigel, M., Parmar, K., Cattin, P., 2017. High order slice interpolation for medical images, in: International Workshop on Simulation and Synthesis in Medical Imaging, Springer. pp. 69–78.

Huang, X., Belongie, S., 2017. Arbitrary style transfer in real-time with adaptive instance normalization, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1501–1510.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, pp. 448–456.

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.

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical image analysis 36, 61–78.

Karras, T., Aila, T., Laine, S., Lehtinen, J., 2017. Progressive growing of GANs for improved quality, stability, and variation, in: International Conference on Learning Representations (ICLR).

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, 55.

Knauth, A.L., Gauvreau, K., Powell, A.J., Landzberg, M.J., Walsh, E.P., Lock, J.E., del Nido, P.J., Geva, T., 2008. Ventricular size and function assessed by cardiac MRI predict major adverse clinical outcomes late after tetralogy of Fallot repair. Heart 94, 211–216.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W., 2017. Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4681–4690.

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.

Leng, J., Xu, G., Zhang, Y., 2013. Medical image interpolation based on multi-resolution registration. Computers & Mathematics with Applications 66, 1–18.

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.

Marwick, T.H., Neubauer, S., Petersen, S.E., 2013. Use of cardiac magnetic resonance and echocardiography in population-based studies: why, where, and when? Circulation: Cardiovascular Imaging 6, 590–596.

Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 .

Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y., 2018. Spectral normalization for generative adversarial networks, in: International Conference on Learning Representations (ICLR).

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.

Odena, A., Olah, C., Shlens, J., 2017. Conditional image synthesis with auxiliary classifier GANs, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org. pp. 2642–2651.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention u-net: Learning where to look for the pancreas, in: Conference on Medical Imaging with Deep Learning (MIDL).

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, 273.

Petersen, S.E., Aung, N., Sanghvi, M.M., Zemrak, F., Fung, K., Paiva, J.M., Francis, J.M., Khanji, M.Y., Lukaschuk, E., Lee, A.M., Carapella, V., Kim, Y.J., Leeson, P., Piechnik, S.K., Neubauer, S., 2017. Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort. Journal of Cardiovascular Magnetic Resonance 19, 18.

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, 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, 8.

Richardson, E., Weiss, Y., 2018. On GANs and GMMs, in: Advances in Neural Information Processing Systems, pp. 5847–5858.

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

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 Besler, 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.

Schlomer, G.L., Bauman, S., Card, N.A., 2010. Best practices for missing data management in counseling psychology. Journal of Counseling psychology 57, 1.

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 transactions on medical imaging 38, 1127–1138.

Ulyanov, D., Vedaldi, A., Lempitsky, V., 2017. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6924–6932.

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.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13, 600–612.

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.

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

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

Zhang, L., Pereañez, M., Bowles, C., Piechnik, S., Neubauer, S., Petersen, S., Frangi, A.F., 2019. Missing Slice Imputation in Population CMR Imaging via Conditional Generative Adversarial Nets, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 651–659.

Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE international conference on computer vision, pp. 2223–2232.

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