TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Label-informed cardiac magnetic resonance image synthesis through conditional generative adversarial networks

Citation for published version (APA):
Amirrajab, S., Al Khalil, Y., Lorenz, C., Weese, J., Pluim, J., & Breeuwer, M. (2022). Label-informed cardiac magnetic resonance image synthesis through conditional generative adversarial networks. *Computerized Medical Imaging and Graphics*, *101*, [102123]. https://doi.org/10.1016/j.compmedimag.2022.102123

Document license:
CC BY

DOI:
10.1016/j.compmedimag.2022.102123

Document status and date:
Published: 01/10/2022

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Download date: 12. Oct. 2022

Computerized Medical Imaging and Graphics 101 (2022) 102123

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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

Label-informed cardiac magnetic resonance image synthesis through conditional generative adversarial networks

Sina Amirrajab [a,*,1], Yasmina Al Khalil [a,1], Cristian Lorenz [b], Jürgen Weese [b], Josien Pluim [a], Marcel Breeuwer [a,c]

[a] Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands
[b] Philips Research Laboratories, Hamburg, Germany
[c] Philips Healthcare, MR R&D - Clinical Science, Best, The Netherlands

ARTICLE INFO

Keywords:
Cardiac image synthesis
Image simulation
Semantic image synthesis
Conditional GANs
Image segmentation
Cardiac MRI

ABSTRACT

Synthesis of a large set of high-quality medical images with variability in anatomical representation and image appearance has the potential to provide solutions for tackling the scarcity of properly annotated data in medical image analysis research. In this paper, we propose a novel framework consisting of image segmentation and synthesis based on mask-conditional GANs for generating high-fidelity and diverse Cardiac Magnetic Resonance (CMR) images. The framework consists of two modules: i) a segmentation module trained using a physics-based simulated database of CMR images to provide multi-tissue labels on real CMR images, and ii) a synthesis module trained using pairs of real CMR images and corresponding multi-tissue labels, to translate input segmentation masks to realistic-looking cardiac images. The anatomy of synthesized images is based on labels, whereas the appearance is learned from the training images. We investigate the effects of the number of tissue labels, quantity of training data, and multi-vendor data on the quality of the synthesized images. Furthermore, we evaluate the effectiveness and usability of the synthetic data for a downstream task of training a deep-learning model for cardiac cavity segmentation in the scenarios of data replacement and augmentation. The results of the replacement study indicate that segmentation models trained with only synthetic data can achieve comparable performance to the baseline model trained with real data, indicating that the synthetic data captures the essential characteristics of its real counterpart. Furthermore, we demonstrate that augmenting real with synthetic data during training can significantly improve both the Dice score (maximum increase of 4%) and Hausdorff Distance (maximum reduction of 40%) for cavity segmentation, suggesting a good potential to aid in tackling medical data scarcity.

1. Introduction

Medical image segmentation has seen significant progress in recent years thanks to development of automated deep-learning (DL) methods (Anwar et al., 2018; Litjens et al., 2017). Similarly, with the emergence of deep generative modeling, the medical imaging community has benefited from its solutions for various applications such as image reconstruction, segmentation, registration, and particularly image synthesis (Kazeminia et al., 2020; Yi et al., 2019). Generative adversarial networks (GANs) demonstrate promising results in generating realistic images for tackling medical data scarcity and patient privacy issues. Medical image synthesis using conditional GANs conditioned on the tissue label map, which tries to translate simplified segmentation mask to realistic image, can be understood as the opposite of the image segmentation. This has a number of important advantages: (i) allowing control of the anatomical information (content) of generated images by controlling the segmentation masks, (ii) explicit alignment with the nature of medical images in which anatomy features are separated from the style and appearance produced by imaging modality, (iii) ability for direct usage of generated images for downstream supervised and unsupervised tasks.

In this paper, we propose a cardiac magnetic resonance (CMR) image synthesis framework for translating the anatomical information of a segmentation mask to a realistic image. The framework is able to generate a diverse database of CMR images with ground truth labels that can be used for any downstream task. Preserving the semantic information (content) of the anatomy presented in the segmentation

* Corresponding author.
E-mail addresses: s.amirrajab@tue.nl (S. Amirrajab), y.al.khalil@tue.nl (Y. Al Khalil), cristian.lorenz@philips.com (C. Lorenz), juergen.weese@philips.com (J. Weese), j.pluim@tue.nl (J. Pluim), m.breeuwer@tue.nl (M. Breeuwer).
[1] These authors contributed equally to the work.

https://doi.org/10.1016/j.compmedimag.2022.102123
Received 2 May 2022; Received in revised form 2 September 2022; Accepted 2 September 2022
Available online 11 September 2022
0895-6111/© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

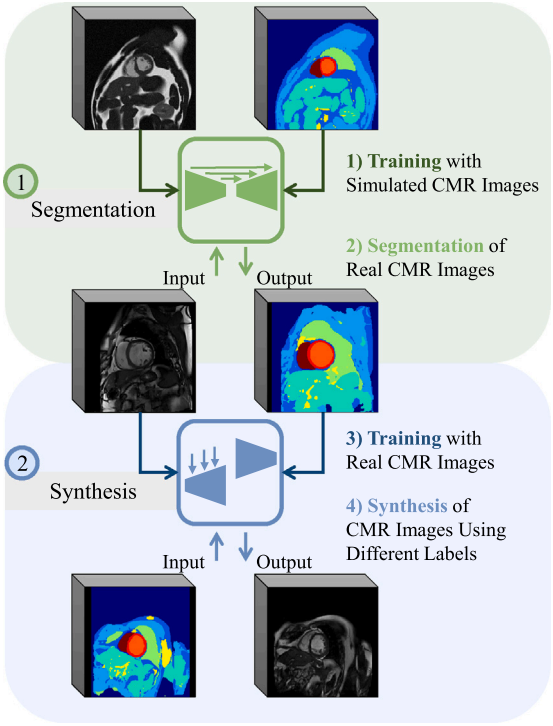

Fig. 1. The proposed framework including **multi-tissue image segmentation** and **synthesis** modules operating offline as follows: **(1) training** the segmentation network using only simulated CMR images with ground truth labels for nine tissue types; **(2) segmenting** real CMR images to provide corresponding multi-tissue labels for each real image; **(3) training** the conditional image generation network using the derived labels alongside with the real images to learn the translation from labels to realistic images; **(4) synthesizing**, in turn, images on different labels.

mask is important when synthesizing data for training a DL medical image segmentation model. Therefore, we propose to incorporate a specific type of conditional GAN that takes segmentation masks as an extra input to the generator architecture. The segmentation mask corresponding to the input image is fed to each layer of the generator using a conditional normalization layer. This is important for learning the textual features (style) of separate classes and preventing information loss when passed through multiple convolutional layers.

The framework consists of two modules: (i) an image segmentation module to provide multi-tissue masks on real CMR images, trained using a physics-based simulated database of CMR images with corresponding labels, and (ii) a semantic image synthesis module to translate input segmentation labels to realistic-looking cardiac images, trained using pairs of real CMR images and corresponding labels. The proposed framework works in a supervised manner, requiring real images with paired multi-tissue segmentation masks. Lack of expert annotations for the cardiac MR images poses another challenge for the adaptation of our method. However, we alleviate this challenge by utilizing simulated cardiac MR images for training a multi-tissue segmentation network that provides labels for organs visible in the field of view, such as lung, abdominal organs, musculoskeletal and skin fat tissue. We perform extensive experiments to analyze the effects of different elements of the framework on the quality of the synthesized images, as well as evaluate the effectiveness and usability of such data for training a supervised DL model for the clinical task of cardiac cavity segmentation.

1.1. Contributions

The contributions of the paper are: (1) Proposing an optimized framework for CMR image synthesis trained with pairs of real images and multi-tissue segmentation masks. The framework is comprised of two modules for image segmentation and synthesis. The first module provides multi-tissue segmentation masks for real images. The second module learns the translation from segmentation masks to realistic images, while preserving the anatomical information captured by tissue labels.

(2) Proposing a simulation-based approach to provide data for training the multi-tissue segmentation network in module (1), while avoiding the need for pairs of real images and ground-truth multi-tissue masks. We leverage a virtual database of simulated cardiac MR images with corresponding ground truth labels derived from the XCAT phantom (a computerized human anatomical model).

(3) Providing detailed assessments of (i) the consequences of using a small or large set of segmentation labels, (ii) the ability to train a segmentation network solely on synthesized data, (iii) the potential of network improvement by augmenting the training set with synthesized data.

In the present paper, we advance our preliminary works (Amirrajab et al., 2020; Abbasi-Sureshjani et al., 2020) substantially: (i) we utilize a more realistic database of simulated CMR images for training and use an optimized multi-tissue segmentation model, which provides an enlarged set of labels essential for high-quality synthesis. Furthermore, we experiment with the number of tissue types within the multi-tissue segmentation mask and choose a different combination of tissue classes compared to the previous work; (ii) we employ a multi-vendor database of real CMR images to demonstrate the ability of the generator to learn the vendor-specific appearance of the images, without using the style encoder network used in the previous works; (iii) we provide a detailed explanation of the network architectures, experiments and evaluation metrics to assess the quality; (iv) We provide a detailed assessment of both the image-quality of the synthesized images and the effectiveness and usability of the generated data for two scenarios: (1) training a heart cavity segmentation network completely using synthetic data to evaluate whether we can achieve the same performance when we utilize the same amount of real data; (2) augmenting real data with addition of synthetic data to boost the performance; (v) A data reduction experiment is carried out where the number of images for training the image synthesis network is reduced to analyze its effects on the quality of the synthesized images. Additionally, a human visual study is performed for qualitative assessment of image realism.

2. Related work

In this section we briefly highlight some of the recent works in the area of conditional and unconditional image synthesis with emphasis on medical applications. Comprehensive overviews of the generative adversarial networks in medical imaging can be found in Yi et al. (2019), Kazeminia et al. (2020) and Singh and Raza (2021).

2.1. Generative adversarial networks

Generative adversarial networks (GANs) aim to learn the underlying process of data generation by forcing the generator to produce fake samples that are indistinguishable from real images (Goodfellow et al., 2014). Image synthesis using GANs has gained major attention thanks to the adversarial strategy for training the generator to synthesize images from a random noise vector. Deep convolutional GAN (Radford et al., 2015) is a variant for unsupervised learning of image representations that uses convolutional neural networks in the architecture of the generator and discriminator. Such a model has been adopted for unconditional medical image synthesis for various applications such as synthesizing CT lung nodules (Chuquicusma et al., 2018), CT liver and brain lesions (Frid-Adar et al., 2018; Zhang et al., 2021), MRI brain (Bermudez et al., 2018), and skin lesion (Baur et al., 2018). Another methodology called progressive growing of GANs (Karras et al., 2017) is used for synthesizing high-resolution retina fundus and brain MRI data (Beers et al., 2018). All of the mentioned work is categorized under the unconditional image synthesis domain, in the sense that a random noise vector is fed to the generator, with no other information for controlling the generation procedure.

2.2. Conditional image synthesis

Conditional image synthesis, on the other hand, relies on using different types of auxiliary information for guiding the generator to produce results, given an input condition (Mirza and Osindero, 2014). Among the first works to investigate conditional GANs for the problem of image-to-image translation is pix2pix by Isola et al. (2017), which demonstrates a general-purpose approach for synthesizing realistic photos from label maps, edge maps, or black and white images. The framework is supervised in the sense that it needs pairs of input–output images from two domains for training. The input image from the first domain is fed to the generator that adopts a U-Net based architecture for translation to the target domain via adversarial loss. In contrast to L1 loss that leads to blurry images, adversarial loss proves to be more effective in generating high-quality images with fine-grain details. The key success lies behind the ability of the discriminator to learn a trainable loss function to capture the subtle differences between real and fake images. The photo realism and resolution of the pix2pix results have further improved through utilizing a course-to-fine generator with residual blocks and a multi-scale discriminator in the pix2pixHD framework proposed by Wang et al. (2018b). Other approaches for unsupervised, unpaired image-to-image translation are used in cases when paired training data from two domains is not available (Zhu et al., 2017; Liu et al., 2017; Huang et al., 2018). Both paired and unpaired methods are successfully employed in the medical domain for the application of cross modality image synthesis — translating an image from one modality (CT) to another one (MR). MR images are translated to CT images (and vice versa) of the brain for radiotherapy treatment planing (Wolterink et al., 2017; Liu et al., 2021), as well as the heart for data augmentation (Chartsias et al., 2017; Zhang et al., 2018).

In the present work, we are interested in translating multi-label segmentation masks to realistic medical images, for the application of providing training data for DL-based CMR image analysis. We aim to provide more control over the underlying anatomy of the synthesized images while producing realistic data with high-quality image appearance at the same time. Upon successful training of this label-informed network, the segmentation mask can be manipulated during the test time to generate a new image with altered anatomical characteristics. Such a synthesized image with corresponding (altered) segmentation mask can be used to relieve data scarcity by providing more labeled images for training an analysis model. We analyze both the quality of the synthesized images and the usability of the synthesized data for a clinical task of cardiac cavity segmentation. Experiments are conducted to evaluate the effects of multi-tissue versus cavity-tissue labels, multi-vendor data, and the number of training data on the realism of the generated images. We further perform a small human visual study for scoring the image quality, in addition to calculating the well-known quality metrics.

3. Method

An overview of the proposed framework is presented in Fig. 1. Two consecutive modules, operated for segmenting real cardiac MR images and for synthesizing images, are the main building blocks of our framework. The segmentation module is designed to provide the corresponding multi-tissue labels for the input images, while the synthesis module utilizes such labels with paired images to learn the translation from segmentation to realistic image appearance. Consequently, the synthesis network can generate cardiac images based on the anatomy characterized by the labels.

3.1. Image segmentation network

We adopt a U-Net architecture (Ronneberger et al., 2015), completely trained using simulated CMR images, for the task of multi-tissue image segmentation. Several modifications to the original architecture are made to optimize the network, including utilization of Leaky ReLU non-linearity and Batch normalization (BN) after each convolutional layer for stabilizing the training, as well as dropout regularization with a rate of 0.5 to avoid overfitting and improve the generalization. The network consists of five down-sampling and up-sampling blocks, trained with a batch size of 32 2D simulated CMR images, for generating pixel-wise predictions for nine tissue classes, including the background. We use a Focal Tversky loss from Abraham and Khan (2019) for training, optimized using Adam for stochastic gradient descent, with an initial learning rate of 10^{-4}. We apply early stopping when the learning rate drops below 10^{-6} and train the network for a total of 350 epochs using 200 simulated volumes for training.

To alleviate the burden of obtaining real data with corresponding labels for supervised training of the U-Net, we utilize a virtual population of simulated cardiac MR images with accurate ground truth labels for training, as discussed in Section 4.1. The predicted multi-tissue segmentation masks are used for two purposes: (i) providing paired data for training the image synthesis network, and (ii) providing input labels for synthesizing new images using the trained generators.

3.2. Image synthesis network

Preserving the content information of the segmentation labels, which represents the anatomy in the context of medical images, is crucial. Traditionally, this is achieved by feeding the segmentation mask as the input of an encoder–decoder architecture with skip connections, forming a U-Net generator (Isola et al., 2017; Wang et al., 2018b). This approach is sub-optimal for preserving the information of the binary segmentation masks, and common (unconditional) normalization layers such as batch normalization (Ioffe and Szegedy, 2015) and instance normalization (Ulyanov et al., 2016) tend to wash away the semantic information of the segmentation mask, as pointed out by Park et al. (2019). On the contrary, a conditional normalization layer, applied throughout the generator architecture, takes the segmentation map as the condition to compute modulation parameters for de-normalizing the activation in a semantic-preserving manner. A SPatially Adaptive DE-normalization (SPADE) layer is therefore proposed by Park et al. (2019) to replace all the unconditional normalization layers to inject the information of the segmentation map into each feature layer of the network. Conditioning the network using SPADE layers was found to significantly improve the performance of image synthesis compared to directly inputting the segmentation mask to the first layer of the generator. Moreover, the generator can have a much lighter architecture which makes it easier and more stable to train, requiring a fewer number of training data.

3.2.1. Generator architecture

The generator architecture, inspired by Park et al. (2019), is comprised of several SPADE residual blocks (He et al., 2016), followed by up-sampling layers to increase the spatial dimension of the input activation. The semantic information of the segmentation mask is injected into each residual block using a SPADE conditional normalization layer, as shown in Fig. 2.

3.2.2. Normalization layers

Normalization layers play a crucial role in conditioning the generator on the input label. The Batch normalization (BN) (Ioffe and Szegedy, 2015) layer has been recognized as an effective component for stabilizing the training of modern deep neural networks by normalizing the statistics of feature activation maps after each convolution followed by a non-linearity layer. The BN first normalizes features in the input

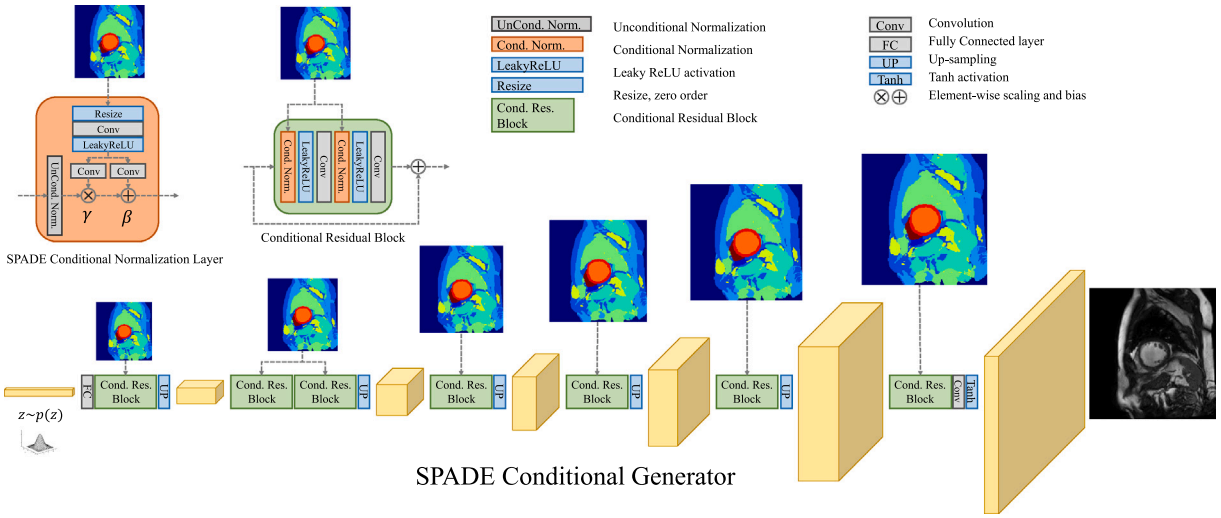

Fig. 2. Label-conditional generator. Adopted from the architecture proposed by Park et al. (2019), it consists of several conditional residual blocks that take the segmentation map as the condition and inject its semantic information via SPatially Adaptive (DE)normalization (SPADE) layers.

mini-batch to its mean and standard deviation, and then uses the scale and bias values for applying the affine transformation. Given the input activation map with the height of H^i and width of W^i for a mini-batch size of N and C^i number of channels ($x^i \in \mathbb{R}^{N\times C^i\times H^i\times W^i}$) at the ith layer of the network, each individual feature channel is normalized as follows:

$$BN_{\gamma,\beta}(x^i) = \gamma\left(\frac{x^i - \mu_c(x^i)}{\sigma_c(x^i) + \epsilon}\right) + \beta, \tag{1}$$

where $\gamma, \beta \in \mathbb{R}^C$ are unconditional scale and bias parameters learned during the training. ϵ is small constant added to the denominator to avoid dividing by zero, $\mu_c(x^i), \sigma_c(x^i) \in \mathbb{R}^C$ are the mean and standard deviation calculated across mini-batch and spatial dimensions ($n \in N, h \in H^i, w \in W^i$) for each feature channel ($c \in C^i$) as follows:

$$\mu_c(x^i) = \frac{1}{NH^iW^i}\sum_{n=1}^{N}\sum_{h=1}^{H^i}\sum_{w=1}^{W^i} x^i_{n,c,h,w} \tag{2}$$

$$\sigma_c(x^i) = \sqrt{\frac{1}{NH^iW^i}\sum_{n=1}^{N}\sum_{n=1}^{H^i}\sum_{w=1}^{W^i}\left(x^i_{n,c,h,w} - \mu_c(x^i)\right)^2} \tag{3}$$

Different from BN, Park et al. (2019) propose to replace the modulation parameters with spatially varying $\gamma^i_{c,h,w}(s^i)$ and $\beta^i_{c,h,w}(s^i)$ dependent on the input segmentation mask (s^i). Following BN layer, the normalized activation is modulated with spatially-dependent scaling and bias factors calculated using a shallow, two-layer modulation convolutional network. Calculated $\gamma^i_{c,h,w}(s^i)$ and $\beta^i_{c,h,w}(s^i)$ are multiplied and added to the normalized activation function element-wise as follows:

$$SPADE_{c,\gamma_c,\beta_c}(x^i, s^i) = \gamma^i_{c,h,w}(s^i)\left(\frac{x^i - \mu_c(x^i)}{\sigma_c(x^i) + \epsilon}\right) + \beta^i_{c,h,w}(s^i), \tag{4}$$

3.3. Discriminator architecture

During the conditional GAN training, the discriminator should also receive the information about the conditional input. In our case, this is the anatomy represented in the segmentation mask. The corresponding ground truth label is concatenated with both the real image and synthesized image (in a channel-wise manner) to construct two pairs of real image-label and synthesized image-label inputs. These two pairs of inputs are then passed to the discriminator in the same batch (as suggested in Park et al. (2019)) to provide two separate predictions for the real image and synthesized image conditioned on the label. Motivated by successful GAN frameworks, we adopt a multi-scale discriminator for high-resolution image synthesis, to differentiate subtle differences between finer details presented in real and synthesized images (Wang et al., 2018b). Two identical PatchGAN discriminators (Isola et al., 2017) with four convolutional layers operate at two image scales that differ by a factor of 2 (see Fig. 3).

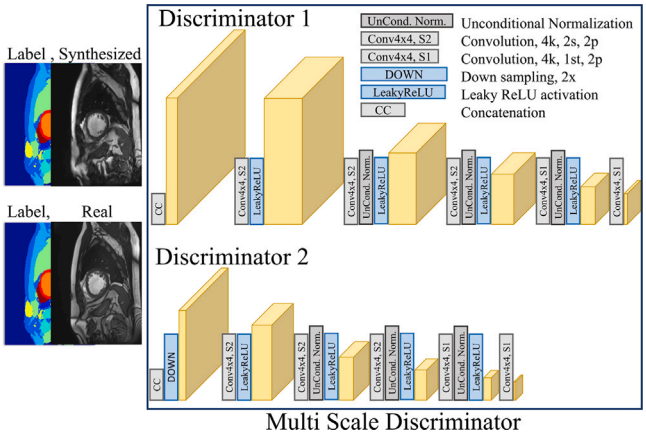

Fig. 3. Multi-scale discriminator architectures including two discriminators, 1 and 2, having identical network structures but operate at two image scales, full scale and down-sampled by a factor of 2.

3.4. Objective function

In the supervised conditional GANs for translating semantic segmentation maps to realistic images, the training set of paired images x_j with corresponding maps s_j is available as (x_j, s_j) during training. The conditional distribution of real images given the input segmentation maps is learned through alternate optimization of the discriminator's objective function $\mathcal{L}_D$ and the generator's objective function $\mathcal{L}_G$. The so-called Hinge loss, introduced by Lim and Ye (2017) is given as:

$$\begin{aligned}\mathcal{L}_D = &- \mathop{\mathbb{E}}_{(x,s)\sim p_{data}(x,s)}[min(0, -1 + D(x, s))] \\ &- \mathop{\mathbb{E}}_{s\sim p_{data}(s), z\sim p(z)}[min(0, -1 - D(G(s, z), s))]\end{aligned} \tag{5}$$

$$\mathcal{L}_G = - \mathop{\mathbb{E}}_{s\sim p_{data}(s), z\sim p(z)} D(G(s, z), s) \tag{6}$$

The GAN is trained via the following modified *minmax* game when using two discriminators D_1 and D_2 operating at different image scales:

$$\min_{G} \max_{D_1,D_2} \sum_{k=1,2} \mathcal{L}_{GAN}(G, D_k) \tag{7}$$

Inspired by many successful GANs for image synthesis (Park et al., 2019; Wang et al., 2018b; Zhu et al., 2020), we employ an enhanced version of the *adversarial loss* in Eq. (7) that is improved using a *feature matching loss* based on the discriminator for stabilizing the training of the generator. Feature maps from multiple layers of the discriminator are extracted for real and synthesized images for the *feature matching loss* calculated as follows:

$$\mathcal{L}_{FM}(G, D_k) = \mathbb{E}_{(x,s)\sim p_{data}(x,s)} \sum_{i=1}^{T} \frac{1}{N_i} \left[\left\| D_k^{(i)}(s, x) - D_k^{(i)}(s, G(s, z)) \right\|_1 \right] \tag{8}$$

where D_k^i denotes the ith layer of the discriminator D_k with the total number of T layers and N_i elements in each layer.

Another well-established practice to overcome the training stability of the GAN models and increase the quality of generated images is to employ a perceptual loss for training the generator, in addition to the adversarial loss (Chen and Koltun, 2017). Training with a perceptual loss achieves superior performance for various image transformation tasks (Ledig et al., 2017; Wang et al., 2018a; Dosovitskiy and Brox, 2016) and its importance has been acknowledged in the domain of medical imaging, as it captures the perceptual quality of the small structures in images (Armanious et al., 2020; Liang et al., 2021).

Let $\phi_i(x)$ be a VGG19 pre-trained network for extracting feature maps from its ith layer when passing the image x, M_i be the number of elements, and N be the total number of layers used to calculate the loss. The *perceptual loss* is defined as the mean absolute error between feature representations of the real and generated images, calculated as:

$$\mathcal{L}_P(G) = \sum_{i=1}^{N} \frac{1}{M_i} \left\| \phi^{(i)}(x) - \phi^{(i)}(G(s, z)) \right\|_1 . \tag{9}$$

The extracted features in the *perceptual loss* are based on the VGG19 pre-trained model rather than the discriminator model in the *feature matching loss*, therefore independent from discriminator training. The overall objective function, composed of the weighted sum of the above mentioned components with hyperparameters of $\lambda_1 = 10$ and $\lambda_2 = 10$ for the contribution of each loss term, is minimized during the training process:

$$\min_{G} \left(\left(\max_{D_1,D_2} \sum_{k=1,2} \mathcal{L}_{GAN}(G, D_k) \right) + \lambda_1 \sum_{k=1,2} \mathcal{L}_{FM}(G, D_k) + \lambda_2 \mathcal{L}_P(G) \right). \tag{10}$$

4. Material and experiments

Simulated and real CMR images are utilized for training the multi-tissue segmentation and synthesis modules.

4.1. Simulated cardiac MR image database

For training the supervised multi-tissue segmentation network of our first module, we require images with labels. We leverage the results from our concurrent project of cardiac MR image simulation and utilize our publicly available generated images provided in the context of the openGTN project[2] including 100 virtual subjects with anatomical and contrast variations. The anatomies of virtual subjects are derived from the 4D XCAT phantoms (Segars et al., 2010) and the images are simulated via a physics-based simulation tool that implements the Bloch equations for cine study. Each simulated image is provided with its corresponding ground truth label map including all simulated organs and tissue types. We create the multi-tissue segmentation map for the simulated images by combining labels for similar-looking organs to have a simplified representation of anatomies visible in the field of view. The segmentation map is comprised of separate labels for left and right ventricles, myocardium, lung, skeletal muscle, skin fat and abdominal organs. An example can be found in Fig. 1.

4.2. Real cardiac MR image database

For training the image synthesis network, the public dataset from the Multi-Center, Multi-Vendor & Multi-Disease (M&Ms) challenge[3] (Campello et al., 2021) is used. The M&Ms data include patients and healthy controls with hypertrophy and dilated cardiomyopathy scanned at clinical centers in three different countries using MR scanners of different vendors, referred to as vendor A1 (Philips scanner center 1), B2 (Siemens scanner center 2), B3 (Siemens scanner center 3). Throughout our experiments, we employ the training subset of data from these two scanner vendors and three clinical centers while the testing subset is completely unseen during the synthesis experiments and it is only used for the final evaluation of the segmentation performance. Expert annotations are only available for left and ventricles, and myocardium, provided for 75 subjects of vendor A1, 50 subjects of vendor B2 and 25 subjects of vendor B3, for end-diastolic (ED) and end-systolic (ES) phases of the heart. We apply our trained multi-tissue segmentation network on all data, but only utilize the data from the first two vendors (A1, B2) for training our image synthesis model, while the last one (B3) is used for evaluation of the synthesis quality of the generated images.

4.3. Data preparation

Before training the segmentation network, all simulated CMR images are re-sampled to 1.25×1.25 mm across short-axis slices, cropped to the same size of 256×256 and normalized with a mean of 0 and standard deviation of 1. Standard data augmentations such as random scaling and rotations, mirroring and horizontal/vertical flips are applied during training. Before training the synthesis network, all real images are pre-processed by re-sampling them to 1.33×1.33 mm in-plane resolution and taking a central crop of the images with 256×256 pixels. The pixel intensity value of images is first clipped between 0 and 4000, assuming 12-bit value, and then normalized between the range of $[-1, 1]$. We only apply random horizontal and vertical flips to images during training with 0.5 probability of the image being flipped. All synthesis models for experiments in this paper are trained with the same training parameters using 3 NVIDIA TITAN Xp GPUs. To have more accurate labels for three heart classes, we correct the raw predicted labels of the first network by substituting the heart annotations provided by the challenge. Furthermore, we remove the slices above basal and below apical locations of the heart that do not contain any heart labels. Therefore, every slice seen during synthesis contains at least one of the heart labels.

4.4. Analysis of the framework

We conduct experiments to qualitatively (noted as **Ql**) and quantitatively (noted as **Qt**) analyze different elements of the framework and their effects on the final outcome:

Ql — Effects of multi-tissue vs. cavity-tissue labels: We evaluate the need for utilizing the multi-tissue segmentation module with an ablated version of our framework, where we train the image synthesis model with only cavity labels. We compare the results with the network trained with full labels, with identical training parameters.

[2] Simulated data can be found at https://opengtn.eu/.

[3] M&Ms data can be found at https://www.ub.edu/mnms/.

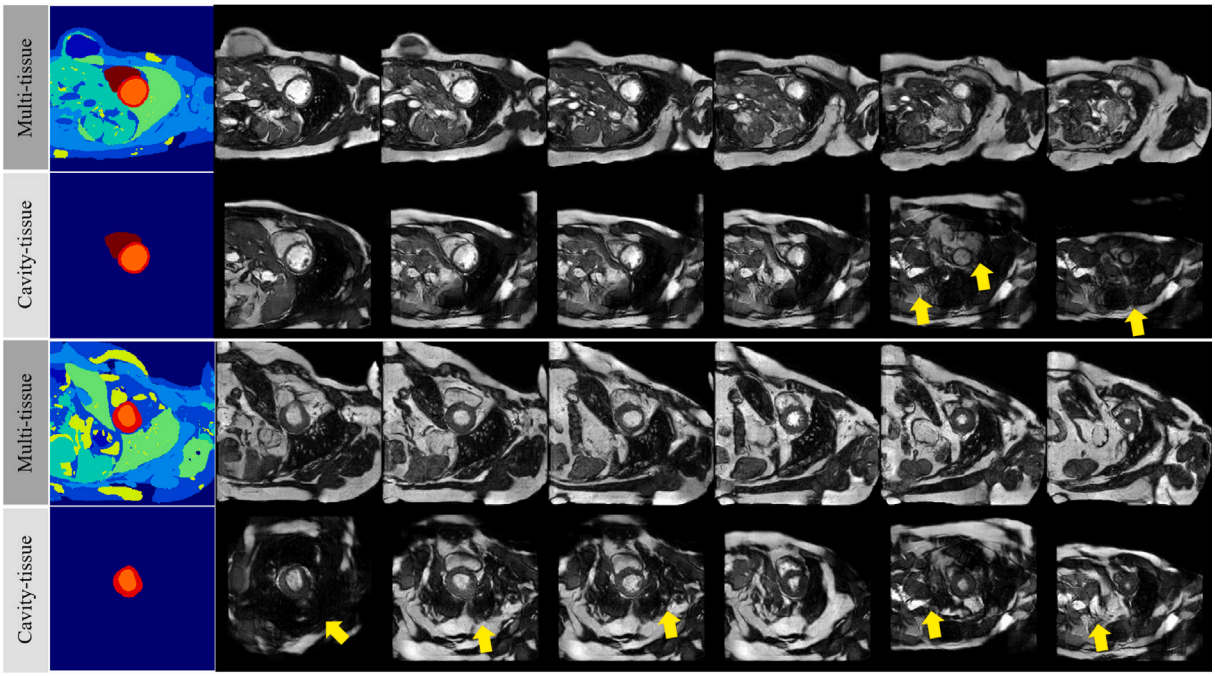

Fig. 4. Visual comparison of synthesis results for two patients for the cases of using multi-tissue labels and only cavity labels. The first column shows the corresponding segmentation of the second column and the following columns show image slices from basal location to the apex of the heart. Distorted parts of anatomies in the latter ones are depicted with yellow arrows. The image quality and volumetric consistency of the generated anatomy have improved in the synthesized images with multi-tissue labels. An animated GIF version of two examples is available in the supplementary material.

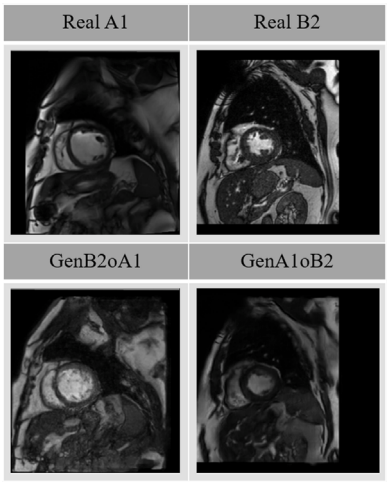

Fig. 5. Cross-synthesis between vendors A1 and B2. GenB2oA1 image is synthesized using the model trained on vendor B2 data and evaluated on the multi-tissue label derived from images of vendor A1, and vice versa for GenA1oB2.

Ql — Effects of Multi-vendor Data: We train two identical multi-tissue generator using training images from the M@Ms challenge dataset (Campello et al., 2021), acquired at different clinical centers by different scanner vendors, to explore the possibility of learning a scanner-specific style of CMR images and the generator's ability to synthesize multiple appearance on a given set of labels. The generator trained on data of vendor A1 and B2 are named GenA1 and GenB2, respectively. When the labels from the other data is used for synthesis, we name this cross-vendor synthesis between (e.g. GenA1oB2 is the result of synthesizing using GenA1 model on labels from vendor B2 images). One example is shown in Fig. 5.

Qt — Synthetic Image Quality Metrics: For quantitative assessment of the synthesis quality, we compute structural similarity index (SSIM), peak signal-to-noise ratio (PSNR), and normalized root mean squared error (NRMSE), between the synthesized images using GenA1 and GenB2 models on the common labels of the B3 (n = 50) and the corresponding real images, completely unseen during training.

Ql, Qt — Effects of Amount of Data: We perform a data reduction experiment to investigate the effect of the number of images used during the training of the image synthesis module on the quality of the synthesized images at inference time. We train different models using a fraction of real training data and compare the quality of the generated images using multi-tissue or cavity-tissue generators.

4.5. Human visual scoring of synthetic images

The quality and realism of the synthesized cardiac MR images are evaluated using a visual scoring tool. We present to the evaluators the results of the cross-synthesis experiment shown in Fig. 5 which includes images of GenA1oB2 (n = 30) and GenB2oA1 (n = 30) together with the real counterpart images of RealA1 (n = 20) and RealB2 (n = 20) in a randomized order, and the evaluator is unaware whether the displayed image is real or synthesized. To include sufficient heart coverage in each displayed image, the mid ventricular slice of each patient is chosen.

One set of experiments with three phases and three questions with multiple choices is designed. To avoid exhausting the evaluators by one time-consuming experiment, the same experimental setup is repeated with different subsets of images (maximum of 40 images for each round). Training phase includes 20 cardiac MR images to get the evaluator familiarized with the scoring setup, questions, and the types of images they can expect to score. Testing phases 1, 2, and 3 including 40, 40 and 20 cardiac MR images that serve as the main experiments for evaluating and scoring the quality of the cardiac MR images. For each displayed image, we ask three questions with multiple choice options: (i) How do you evaluate the overall quality of the image? (ii) How do you evaluate the image realism with focus on the heart? with a scoring on a scale of one to five, with one being worst score (very poor quality) and five being best score (very good quality), and (iii) How confident are you that the image is real or synthesized? (Synthesized, Maybe synthesized, Cannot tell, Maybe real, Real)

4.6. Utilization of synthetic data and experimental design

We employ synthesized images for training a DL model to segment left and right ventricular cavities and myocardium of the heart. The purpose of these experiments is to investigate two aspects of the synthetic data: (i) its ability to mimic the distribution of the real training data and therefore, its effectiveness in training a well-performing segmentation algorithm, which we refer to as a *replacement* experiment and (ii) its usability for augmentation purposes, where we quantify the effect of augmenting the real MR data during training with synthetic data, referred to as an *augmentation* experiment. To this end, we utilize the results of a cross-vendor experiment explained earlier, which aims to transfer the appearance of the images acquired using the scanner of one vendor on the anatomical masks derived from the subjects scanned using another vendor. Additionally, we vary the heart shape during synthesis by applying random elastic deformation and morphological dilation on heart masks (just before image generation) to introduce anatomical variation in the synthetically generated data. An animated GIF of applying multiple deformations on the heart labels during synthesis is available in the supplementary material.

The baseline model is a 2D U-Net trained with a combination of real images of vendor A1 and B2 (**Real**). We compare the performance of this model with four identical ones trained with: (i) only synthetic data with cavity-tissue labels (**Synth-Cavity**), (ii) only synthetic data with multi-tissue labels (**Synth-Multi**) for the *replacement* experiment, (iii) augmenting the real data with cavity-tissue synthesis (**Aug-Synth-Cavity**), and (vi) augmenting the real data with multi-tissue synthesis (**Aug-Synth-Multi**) for the *augmentation* experiment.

All networks consist of six downsampling and upsampling convolutional blocks with five max-pooling operations. Each convolutional block contains 3×3 kernel convolutional layers, batch normalization and leaky ReLU activation function. We additionally apply dropout regularization, with a rate of 0.5, after each concatenating operation in the up-sampling path of the network. To increase robustness of the networks, we augment the training set by applying random vertical and horizontal flips (p = 0.5), random rotation by integer multiples of $\frac{\pi}{2}$ (p = 0.5), random translations (p = 0.3) and mirroring (p = 0.5), as well as random elastic deformations (p = 0.4). We do not apply any contrast transformations, to better inspect the influence of synthetic data. All augmentation operations are applied on the fly during training. At inference time, we only apply normalization and in-plane re-sampling.

To train the network, we use a weighted sum of the categorical cross-entropy and Dice loss. We use Adam for optimization, with an initial learning rate of 5×10^{-5} and a weight decay of $3 \times e^{-5}$. During training, the learning rate is reduced by a factor of 5 if the validation loss does not improve by at least 5×10^{-3} for 50 epochs. We apply early stopping on the validation set to avoid over-fitting and select the model with the highest accuracy. As a post-processing step, we perform connected component analysis on the predicted labels, where we remove all but the largest connected component for each class and therefore, remove large false positive predictions.

5. Results

5.1. Ql — Effects of multi-tissue vs. cavity-tissue labels

Utilizing multi-tissue segmentation labels substantially improves the quality of the synthesized images by providing more guidance to the generator. Fig. 4 depicts the results of utilizing cavity- tissue compared with the use of multi-tissue labels during synthesis, for two synthetic subjects for six slices covering the heart. Although organ labels are not highly accurate for the case of multi-tissue labels, the generator is able to synthesize realistic anatomies. Plausible organs are generated around the heart when we use more labels for image synthesis. Furthermore, we observe from consecutive slices of one subject (An animated GIF is available in the supplementary material) that the generated anatomy is consistent in 3D, making results suitable for 3D medical image analysis. We provide a video in the supplementary material showing 3D image generated using multi- and cavity tissue labels for all three projections (xy, xz, yz) to visualize the consistency between individual slices. Note that the synthesis network is 2D, taking one slice at a time for image generation. Distortions in the generated images for cavity-tissue are highlighted with yellow arrows.

5.2. Ql — Effects of multi-vendor data

Fig. 6 shows the generation results on common labels derived from images of vendor B3, for the generator trained with images of vendor A1 *(GenA1oB3)*, and the generator trained with images of vendor B2*(GenB2oB3)*. From the visual comparison of synthesis results, we observe that the main characteristics of the training images are captured by the generators, allowing them to synthesize multiple appearances on a given label. Vendor-specific characterizations, such as the darkness and blurriness of the images acquired from vendor A1, the sharpness of the edges and the noisiness observed in images from vendor B2, and myocardium-to-blood contrast, are learned by the generators and manifested in the synthesized images.

5.3. Qt — Synthetic image quality metrics and effects of amount of data

The images are generated on *Cavity* or *Multi* tissue labels using models trained with combined data of vendor A1 (n = 150) and B2 (n = 100), *GenA1B2*, separate data of vendor B2, *GenB2*, 50% of vendor B2 data, *Gen50%B2*, 25% of vendor B2 data, *Gen25%B2*, and 12% of vendor B2 data, *Gen12%B2* as depicted in Fig. 7. The results of the generator utilizing multi-tissue labels achieve the best score for all metrics, suggesting that the better quality is achieved for multi-tissue synthesis, despite the reduction in the training data. Interestingly, we observe that the *Multi-Gen12%B2* generator, which uses only 12 volumes for training, scores significantly higher across all metrics than even the *Cavity-GenA1B2* that uses all 250 real images during training, suggesting the benefit of multi-tissue labels when limited data is available during synthesis.

Fig. 8 depicts synthesis results on the same subjects generated by the models trained with a fraction of the training data of vendor B2. We observe that the synthesis quality of the multi-tissue generator is retained even when only 12% of the data (12 volumes) is used for training, compared to the generator models trained with the full available data-set. On the other hand, repeating the same experiment on the generator trained with cavity-tissue labels produces images of impaired quality and severe distortion.

5.4. Human visual scoring of synthetic data

Two independent raters completed the experiments. Evaluator 1 was an imaging scientist with ample experience in CMR application and processing, who was involved in this image synthesis research and has experience looking at GANs-generated medical images, and evaluator 2 was a CMR clinical expert and consultant cardiologist, who was not involved in this project. Fig. 9 depicts the scores of the evaluators for testing phases. The evaluation results for two sources of the real (RealA1, RealB2) and two sources of the synthesized images (GenA1oB2, GenB2oA1) are combined. The results for the total of 40 real and 60 synthesized images are shown in this figure. The horizontal axis shows the level of the image quality score, and the vertical axis shows the percentage of the total images rated with that image quality score. e.g. around 70% of the synthetic data compared to around 60% of real images are perceived as *Good* in terms of image realism with focus on the heart (scored 4). Surprisingly, lower overall quality of the real images is scored by the evaluator 1, having more real images with scores of 1 *very poor* and 2 *poor*, while same percentage scored 4 *Good*, and slightly more synthesized scored 3 *Mediocre*. The

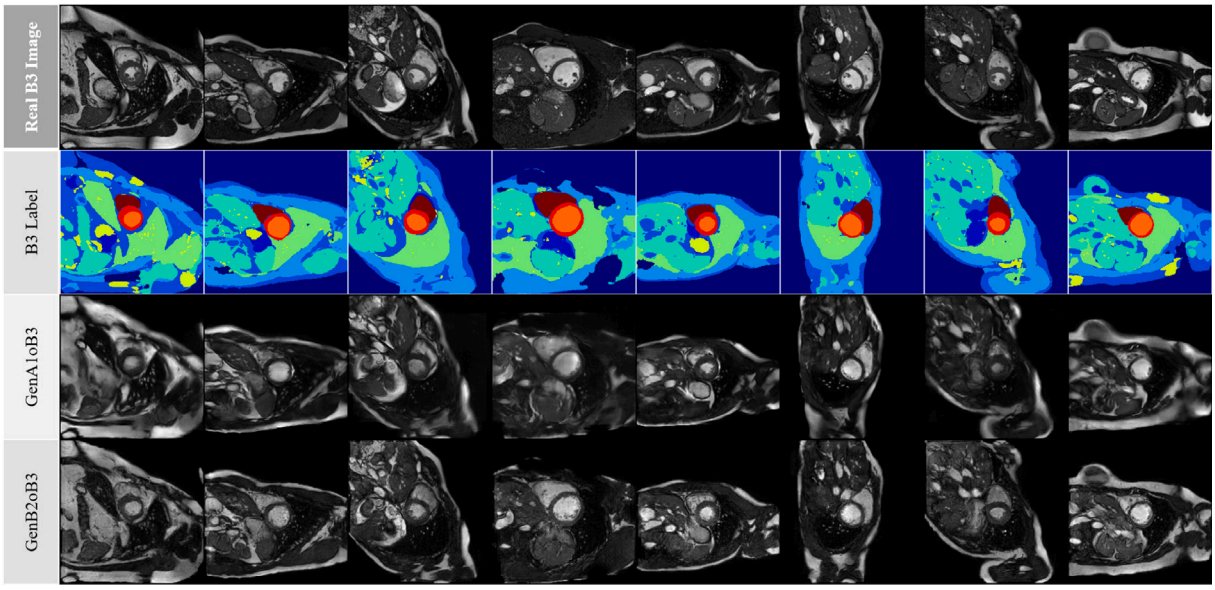

Fig. 6. Synthesis results on labels derived from vendor B3 data for model trained on data of vendor A1 (GenA1oB3) and the same model trained on data of vendor B2 (GenB2oB3). The difference in the appearance of the synthesized images demonstrates that the generator learns the vendor-specific features of the training images.

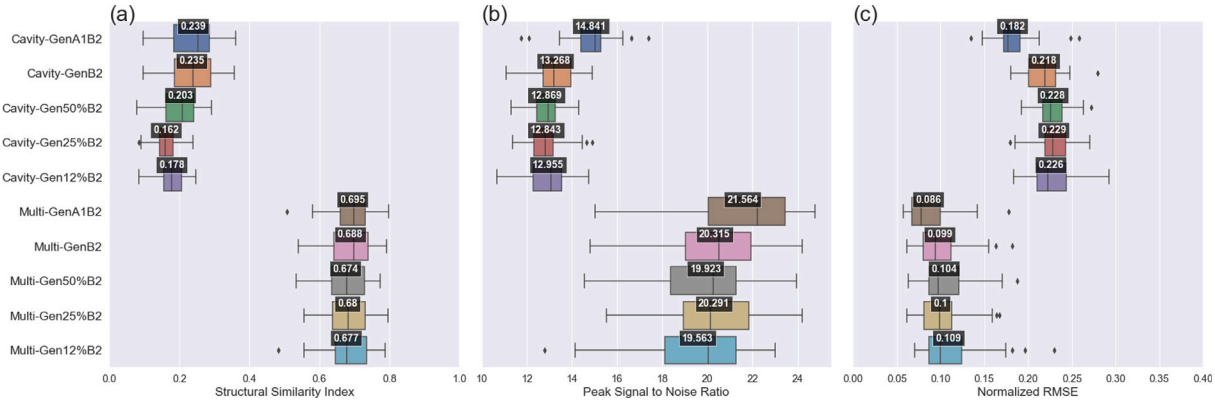

Fig. 7. Quantitative evaluation metrics for comparing the quality of the synthesized images using models trained with multi-tissue (*Multi-*) and cavity-tissue (*Cavity-*) labels. Structural similarity index (a), peak signal to noise ratio (b) and normalized root mean squared error (c) are calculated between real images and corresponding synthesized counterparts. *GenA1B2* and *GenB2* indicate that the synthesized data are generated by the generator trained on the combination of data from vendor A1(n = 150) B2(n = 100) and only vendor B2 (n = 100), respectively. 50%B2, 25%B2, and 12%B2 indicate models trained with fraction of the data from vendor B2 (n = 50, 25, 12). All models are tested on the unseen data of vendor B3 (n = 50).

Table 1
The performance of segmentation models for data *replacement* and *augmentation* experiments.

	Vendor A						Vendor B					
Segmentation models	Dice(LV)	Dice(RV)	Dice(Myo)	HD(LV)	HD(RV)	HD(Myo)	Dice(LV)	Dice(RV)	Dice(Myo)	HD(LV)	HD(RV)	HD(Myo)
Real	0.932	0.894	0.847	8.32	12.44	10.70	0.911	0.896	0.860	8.75	9.89	11.78
Synth-Cavity	0.890	0.845	0.780	14.75	23.86	19.44	0.880	0.842	0.794	13.19	17.22	17.01
Synth-Multi	0.912	0.877	0.819	9.98	13.89	12.92	0.888	0.898	0.842	12.10	11.50	10.91
Aug-Synth-Cavity	0.938	0.907	0.869	6.99	10.72	8.48	0.926	0.926	0.871	8.85	9.51	11.20
Aug-Synth-Multi	**0.945**	**0.920**[a]	**0.877**[a]	**5.83**	**9.22**[a]	**6.83**[a]	**0.932**	**0.929**[a]	**0.883**[a]	**5.24**[a]	**7.47**[a]	**8.85**[a]

[a]Numbers have p-value of $p < 0.01$ against the Real model according to the Wilcoxon signed-rank test and numbers with Bold text are the best scores.

overall quality of both images are higher for the evaluator 2. From the last question, the evaluator 1 is confident that only just 10% of the synthesized images are *"Synthesized"*, around 55% *"Maybe Synthesized"* and the rest are either unidentifiable or rated as *"Maybe Real"*. For the evaluator 2, interestingly, the majority of the synthesized cases are rated either *"Cannot Tell"* or *"Maybe Real"*. One conclusion to draw is that the synthesized images are on par with real images in terms of visual quality, sometimes perceived as better quality, and they are indistinguishable from the real counterparts even by experienced evaluators.

5.5. Utilization of synthetic data

We utilize synthetic CMR images with the aim to investigate the effectiveness and utility of such data in training a deep-learning algorithm for segmentation, evaluated by carrying out the *replacement* and *Augmentation* experiments. Hence, we provide a quantitative, application-based evaluation of the usability of synthesized images for a clinical task to segment myocardium (MYO), left ventricle (LV), and right ventricle (RV). All models are tested on the same set of images, acquired from two vendors and unseen during the training

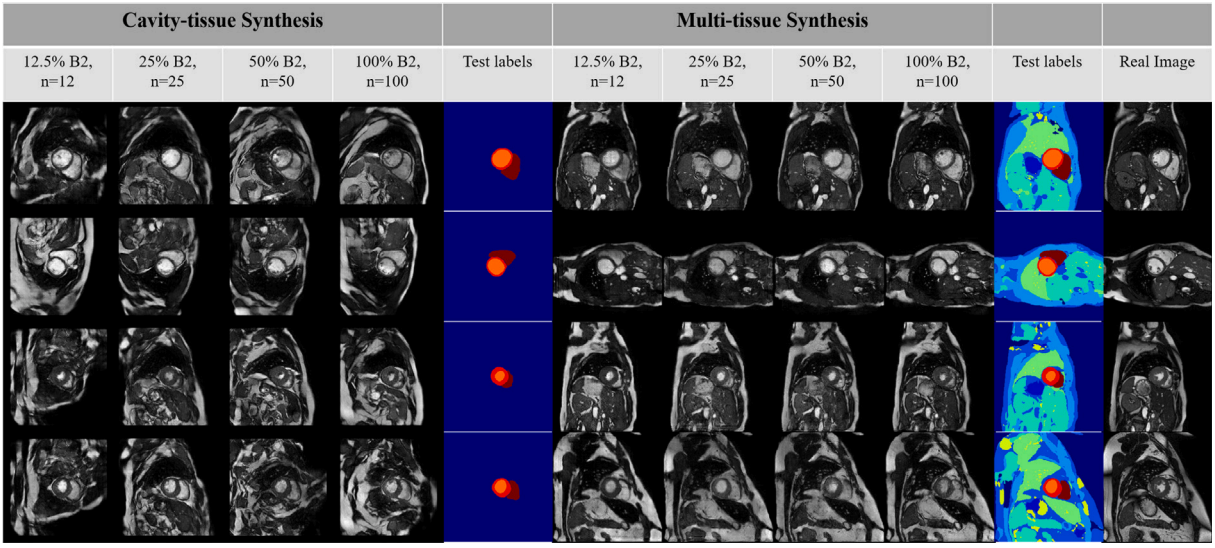

Fig. 8. The data reduction results for the case of multi-tissue and cavity-tissue generators trained using the data from vendor B2 (center 2, n = 100) and tested on the unseen ground truth labels of B3 (center 3, n = 50). The corresponding real testing images are shown in the first row.

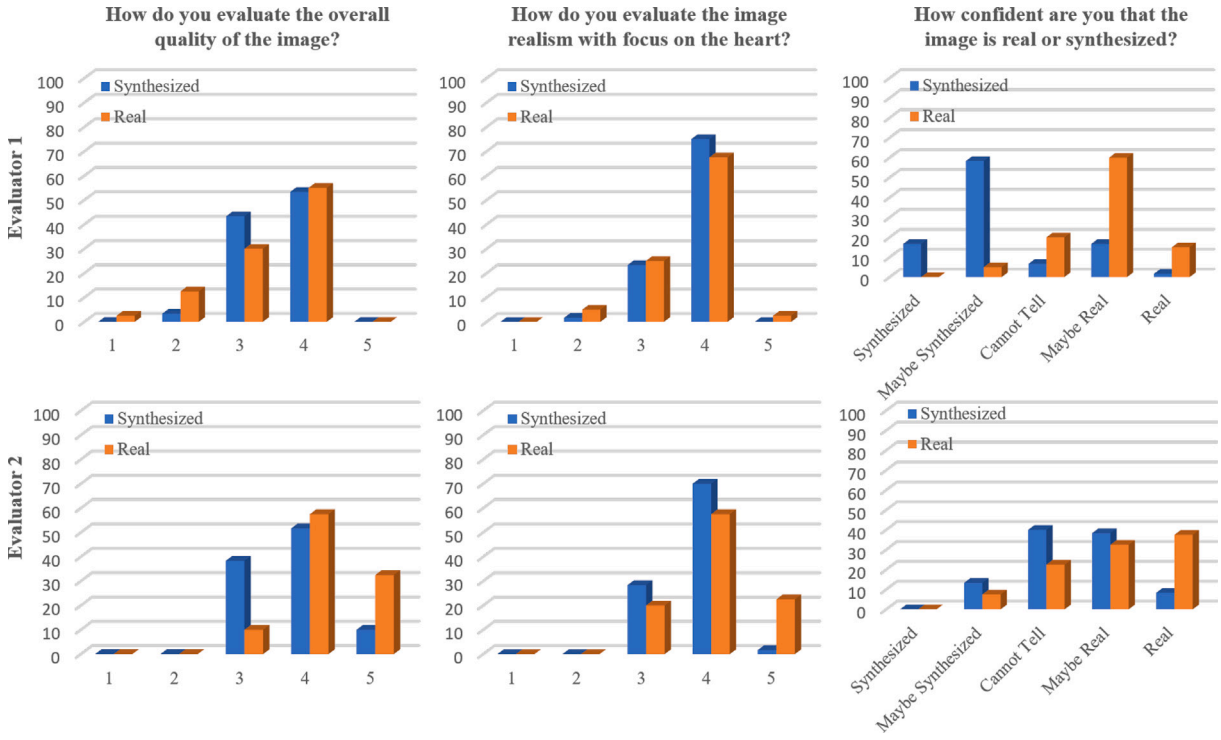

Fig. 9. The visual quality scoring results for two evaluators (1: Scientist, 2: CMR specialist). The outcome of all testing experiments for two sources of the real (RealA1, RealB2) and two sources of the synthesized images (GenA1oB2, GenB2oA1) are combined. This includes the total of 40 real and 50 synthesized images. The scores correspond to; 1: very poor, 2: poor, 3: mediocre, 4: good, and 5: very good quality.

and development of the models. We utilize standardized metrices for evaluation of segmentation performance in the literature, namely the Dice coefficient and the Hausdorff Distance (HD) score.

We start by evaluating the *replacement* experiment, where we study the extent to which the data synthesized in this study can be used to train a model for segmenting the heart cavity tissues in real CMR images. By running this experiment, we can additionally quantify the realism of such data and possibly identify points of improvement for future work. The results in this experiment are shown for separated classes in Fig. 10(a) and (b). First, we observe that both models trained with synthetic data exhibit a drop in performance compared to the model trained with real data (*Real*). This result was expected due to the nature and characteristics of GAN-based conditional synthesis. However, we see a consistent and significant improvement in the performance of the *Synth-Multi* model compared to the *Synth-Cavity* model, where in some cases *Synth-Multi* performs almost at par with the *Real* model, across both vendors and all tissues. We hypothesize this is due to better image quality achieved through synthesizing with multi-tissue labels, where tissue boundaries are much clearer, with better contrast achieved between the tissues. Moreover, this indicates a better resemblance of the *Synth-Multi* data features to the real data. Using the *Synth-Multi* model, we notice a significant improvement in reducing the appearance of outliers, but also improving the average Dice score across all patients. Most of these outlier predictions tend to

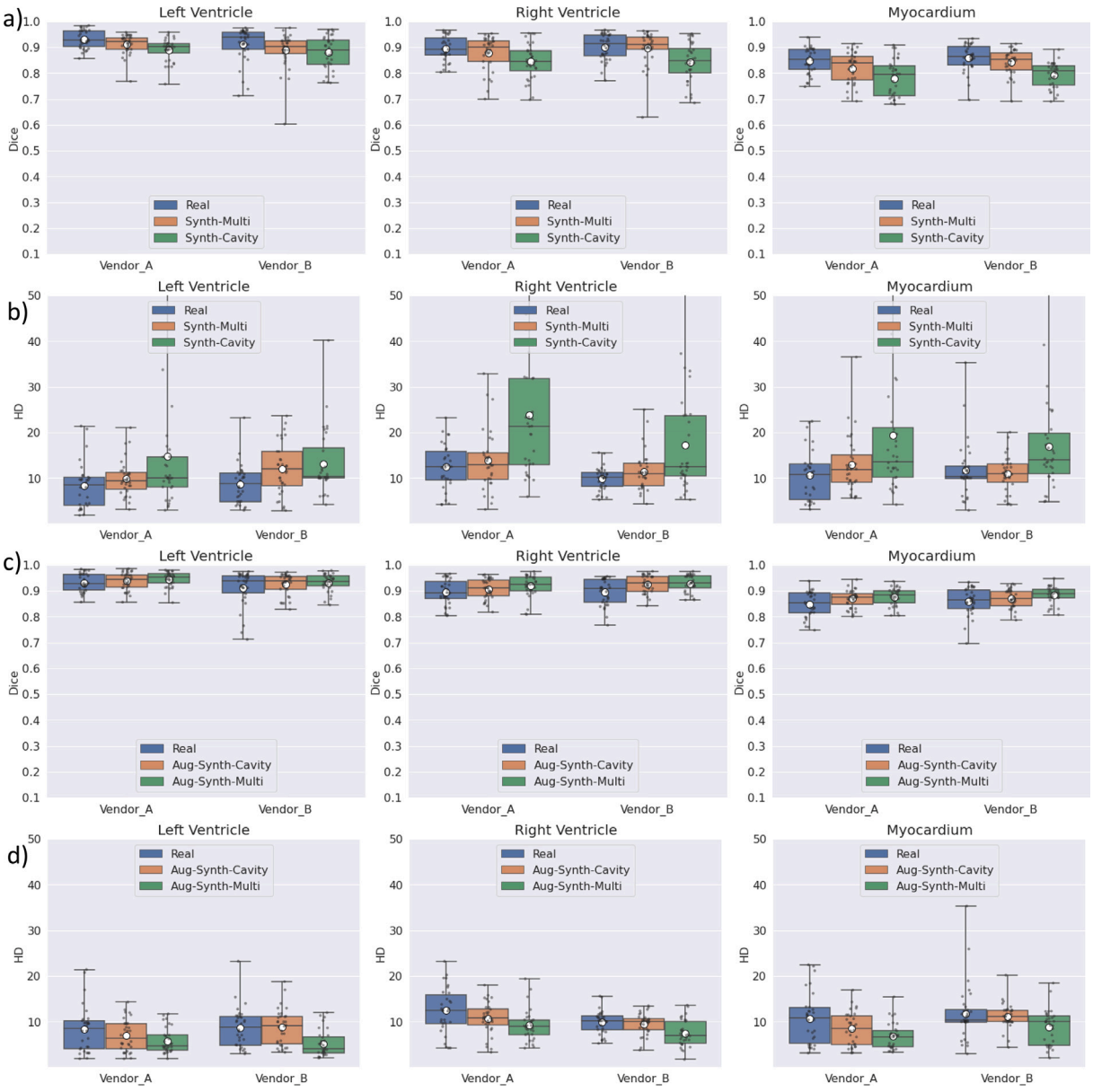

Fig. 10. Quantitative usability evaluation of the synthesized data for two experiments of data *replacement* (a,b) and data *augmentation* (c,d). Identical 2D segmentation networks are trained using different training data combinations for the task of cardiac segmentation. The baseline model, *Real*, is trained using all available real training examples of vendor A1 and B2 (n = 250); *Synth-Cavity* and *Synth-Multi* are models trained completely using synthetic images when cavity-tissue and multi-tissue labels are employed during the GANs training, respectively. *Aug-Synth-Cavity* and *Aug-Synth-Multi* are models trained with augmented data of real with cross-synthesis results. We apply geometrical deformations on heart labels during synthesis to introduce more anatomical variations. The Dice coefficient, *Dice* (higher better), and the Hausdorff distance (pixel), *HD* (lower better) are reported for segmentation performance.

be located across basal and apical slices, which can be partly attributed to the removal of unlabeled slices above the base and below the apex of the heart during synthesis. We additionally observe false positive predictions generally appearing across all slices using the *Synth-Cavity* model, where tissue similar in shape, size or appearance to the heart cavity is often falsely identified as a part of the heart. These errors are reduced by utilizing images of better quality for training the *Synth-Multi* model. Being sensitive to false positive predictions, the HD scores for the *Synth-Cavity* model tend to be higher with a higher percentage of outlier cases across all vendors and tissues.

The results for the *augmentation* experiment are shown in Fig. 10 in rows (c) and (d). This experiment shows that the addition of synthetic data to the training set containing real data improves the segmentation performance across both vendors and all tissues. This happens for both cases (*Aug-Synth-Cavity* and *Aug-Synth-Multi*). However, adding *Synth-Multi* images to the training is shown to be consistently better than *Synth-Cavity* both in terms of Dice score and HD score, especially when it comes to reducing the variance and outliers of the model performance. This is particularly evident in scores acquired across all vendor B tissues, as well as in vendor A myocardium, which are significantly higher compared to the *Aug-Synth-Cavity* model ($p < 0.01$ according to the Wilcoxon signed-rank test). In fact, visual observations indicate that the *Aug-Synth-Multi* model consistently yields segmentation improvements due to:

- the reduction of false positive and false negative predictions around the base and apex of the heart, respectively;
- better segmentation of cases with a thicker myocardium, often appearing in ES images, which we hypothesize is the influence of wider anatomical and contrast variations introduced to the training through synthetic images;

- consistent improvements in the segmentation of the RV around the base of the heart;
- better segmentation robustness in images with visible acquisition artefacts and noise and better adaptation to cases with poor tissue contrast.

Moreover, models trained on real images only are prone to under-segmentation in the occurrence of hyperintensities, which is largely alleviated with the addition of synthetic data.

Some of the above-mentioned improvements are also present in the *Aug-Synth-Cavity* model. However, the under- and over-segmentation of tissues in apical and basal slices still remains a problem. Despite *Synth-Cavity* images being of lower quality, we hypothesize that an improvement in performance compared to the *Real* model is partly achieved as such images present harder examples during the training procedure, thus improving network regularization and model generalization. A detailed summary of the experiments is presented in Table 1.

5.6. Conditional GANs comparison

We compare the proposed label conditional GANs model (based on the SPADE normalization layers) with pix2pix (Isola et al., 2017) and pix2pixHD (Wang et al., 2018b) models in terms of synthesized image quality. All three models are trained with the same hyper-parameters using the full data of vendor A1 and B2 with multi-tissue labels and evaluated on unseen data from vendor B3. We can observe from the synthesis results shown in Fig. 11 that more anatomical distortions are produced by pix2pix and pix2pixHD models (indicated by yellow arrows), especially in the heart area. The SPADE-based generator can better preserve the content of the input labels, resulting in more accurate generation of the heart. The calculated image quality metrics (as described in 4.4 **Synthetic Image Quality Metrics**) are shown in Fig. 12, confirming the superiority of the SPADE-based generator, strongly for the structural similarity index.

6. Discussion

6.1. Multi-tissue vs. cavity-tissue labels for synthesis

For synthesizing realistic cardiac MR images we conduct experiments to investigate the effects of using multi-tissue segmentation compared with using only cavity-tissue labels during synthesis. The quality of the results achieved by the former setup is significantly better, demonstrating superior image quality with fewer distortions and artifacts. We observe substantial benefits of using a rough segmentation mask for partitioning the input image into multi-tissue labels. Firstly, despite the fact that our framework is two dimensional (2D), i.e. it takes 2D labels as the input and produces 2D images, the global consistency for volumetric image generation seems to be improved, as shown in the supplementary material. Our results are consistent in the slice direction and therefore suitable to generate data for 3D image analysis algorithms. Secondly, we observe that the training time and stability is improved when training the synthesis module with multi-tissue labels, resulting in a fewer number of epochs required and a faster learning procedure. Finally, our experiments show that even with as few as 12 real volumes accompanied with rough multi-tissue labels, we can synthesize realistic images, despite the loss high-frequency details of small structures. From the visual assessment of the results we observe wrong anatomical positioning in some synthesized images on cavity-tissue labels, i.e the cavity-tissue generator is not able to correctly locate anatomical structures, such as lung and abdominal organs surrounding the heart, as seen in Fig. 8 second and fourth rows for the cavity-tissue generator. The correct heart position, with respect to lung and other organs, is preserved for the multi-tissue generator, resulting in synthesis of plausible anatomies. Note that we train our multi-tissue segmentation model using simulated cardiac MRI data while one could replace that module with another approach or manually segment the images.

6.2. Synthesis quality and the amount of data

We conduct the data reduction experiment to assess the effects of the number of training images on the synthesis quality, comparing the results of generating on multi-tissue and cavity-tissue labels. The quantitative metrics shown in Fig. 7 suggest that the quality of the synthesized images on multi-tissue labels is significantly better than synthesized images on cavity-tissue labels, increasing the structural similarity index from 0.239 to 0.695 for the case of utilizing all images from vendor A1 and B2 (total of 250 volumes). Gradually reducing the training data causes drastic degradation in the quality of synthesized images on cavity-tissue labels, whereas it does not severely affect the global quality of the synthesized images on multi-tissue labels. For example, the quality of multi-tissue generated images for the case of using only 12%B2 (n = 12) constantly outperforms the cavity-tissue generated ones using all 250 volumes of A1 and B2. This suggests that there are significant advantages of putting effort into obtaining multiple tissue labels when there is a limited amount of data available for training, as shown in Fig. 7. Furthermore, as seen in Fig. 8, while in images obtained by the cavity-tissue generator the anatomy of organs and structures are distorted, we observe that only the high-resolution details of the image, such as the papillary muscles and trabeculation of the heart are missing in images generated by the multi-tissue generator for the cases of reducing the number of data during training.

6.3. On the usefulness of the synthetic data

Despite recent attempts to synthesize realistic cardiac magnetic resonance images, the analysis of how useful the generated data is for tackling a real world task remains largely unexplored. Our experiments show that there is a performance drop when real data is completely replaced by synthesized data for the task of cardiac cavity segmentation. This implies that some of the features of the real data are not fully captured during synthesis. However, the addition of the synthetic data can substantially improve the model performance, indicated by its benefits for data augmentation. We observe a maximum increase of 4% for Dice and a maximum reduction of 40% for HD in the presented experiments. The HD score is significantly improved when real data is augmented with multi-tissue synthetic data, suggested by the reduction in the outlier predictions and an overall improved segmentation accuracy of tissue boundaries. Based on visual inspection of the segmentation results, aiding the training with multi-tissue synthetic images yields a better segmentation performance around the apex and base of the heart and more accurate segmentation of the right ventricular cavity and myocardium. Our observations suggest that the improvements could be attributed to model robustness to artefacts, noise and poor tissue contrast, but further investigation is needed, which is planned for future work. Despite the fact that we observe better performance when using synthetic multi-tissue data, the cavity-tissue images with lower visual quality are still found to be helpful for data augmentation. This could imply that by seeing somewhat more distorted generated images, the model can learn to deal with some aspects of the data such as unclear tissue boundaries, blurriness, and other imaging artifacts, which are unrepresented in the real training data. Similar results have been reported in Karimi-Bidhendi et al. (2020), where the segmentation performance can be improved with visually low-quality generated cardiac MR images.

6.4. Limitations and future research

In this work, we synthesize cardiac MR images for cine study with imaging features learned according to the acquisition protocol of the training data. However, multi-modality image synthesis can be achieved by attaching an encoder network to the generator for capturing the modality-specific style of images, e.g. image appearance of the late gadolinium enhancement. As a result, cross-modality image

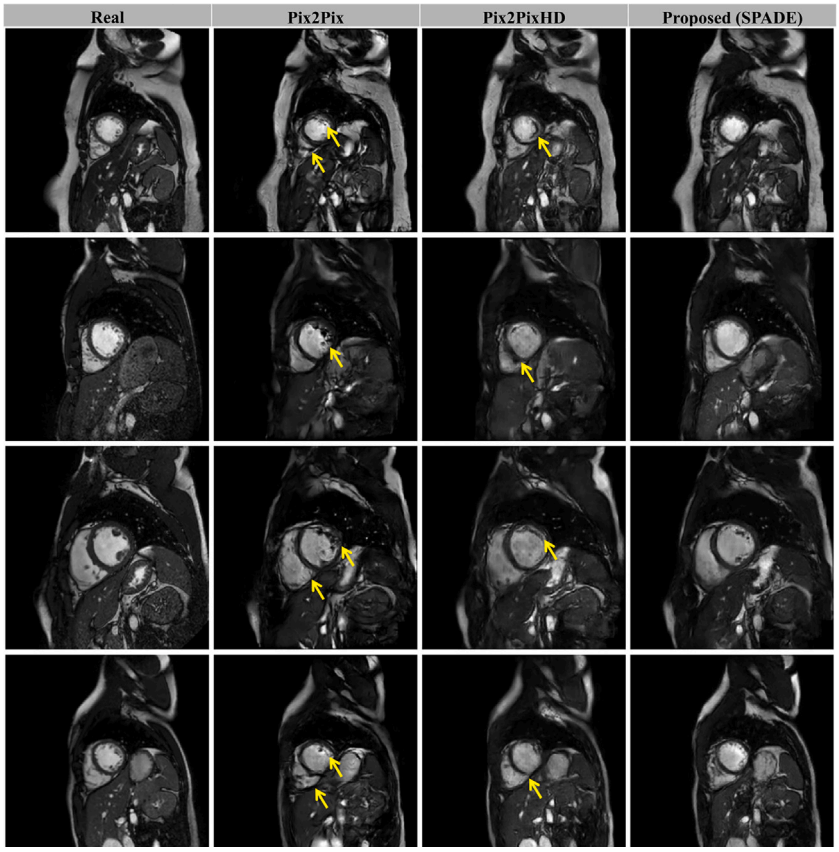

Fig. 11. Visual comparison between different generator models trained with the same hyper-parameters using the combined data from vendor A1 and B2 and evaluated on the unseen data from vendor B3. Yellow arrows indicate geometrical distortions in the heart area for pix2pix (Isola et al., 2017) and pix2pixHD (Wang et al., 2018b) models.

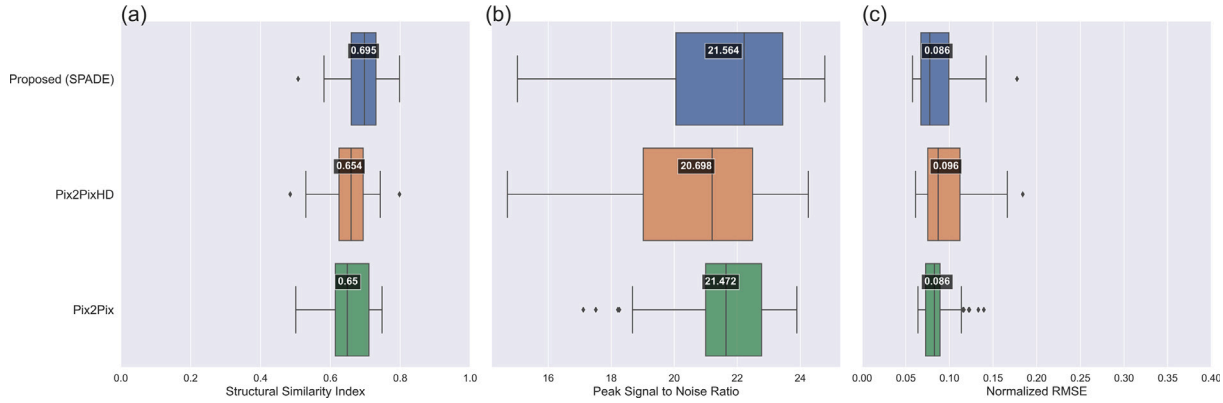

Fig. 12. Quality metrics for quantitative evaluation of the results for the proposed synthesis model in this study (SPADE), pix2pix (Isola et al., 2017) and pi2pixHD (Wang et al., 2018b) models. We compare the generated image quality based on (a) structural similarity index (higher better), (b) peak signal to noise ratio (higher better) and (c) normalized root mean square error (lower better) between the generated and real images for unseen data from vendor B3(n = 50 subjects).

synthesis, defined as domain translation from cine to late enhanced cardiac MR images (or to cardiac CT image), could potentially address the challenge of data availability in a wider medical imaging context. Furthermore, with the ability to manipulate labels for synthesizing a new image, it is possible to generate a population of virtual subjects with heart diseases, as long as the disease can be represented in the tissue mask. We show that the generated image quality increases by utilizing multi-tissue labels even with noisy and inaccurate anatomy. However, to synthesize highly detailed anatomies, one could work on the accuracy of the multi-tissue segmentation network to provide more accurate labels for better guidance of the synthesis network. In fact, the multi-tissue segmentation module could be replaced with any available method to obtain detailed segmentation maps of both the heart and the tissues surrounding the heart. To evaluate the usefulness and effectiveness of synthetic data, we perform experiments for the task of medical image segmentation; however, other medical image analysis tasks can also benefit from the achieved results in this work. Finally, more research should be focused at addressing the synthesis of artifacts and other specific characteristics of MR images that can potentially present a larger feature variability that DL-based models, trained for different medical imaging analysis tasks, can learn from.

7. Conclusion

In this paper, we propose a framework for cardiac MR image segmentation and synthesis with the aim of generating high-quality

images with informed anatomical representation and learned imaging features. Leveraging Label-conditioned normalization layers throughout the generator architecture allows for the preservation of content information, while at the same time accurately transferring the image characteristics of real data. Furthermore, one of the main findings of this work is the importance of introducing detailed labels in the form of multi-tissue maps for generation of highly realistic images with accurate and 3D-consistent anatomies even with a significantly smaller number of training data, compared to utilizing only cavity-tissue labels during training. The effectiveness and usability of synthetic images for the task of cardiac segmentation was evaluated, demonstrating that data augmentation with synthetic data can substantially boost the segmentation performance in terms of both Dice score (maximum increase of 4%) and Hausdorff Distance (maximum reduction of 40%).

CRediT authorship contribution statement

Sina Amirrajab: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft. **Yasmina Al Khalil:** Conceptualization, Methodology, Software, Validation, Investigation. **Cristian Lorenz:** Writing – review & editing, Supervision. **Jürgen Weese:** Writing – review & editing, Supervision. **Josien Pluim:** Writing – review & editing, Supervision. **Marcel Breeuwer:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition, Project administration.

Declaration of competing interest

One or more of the authors of this paper have disclosed potential or pertinent conflicts of interest, which may include receipt of payment, either direct or indirect, institutional support, or association with an entity in the biomedical field which may be perceived to have potential conflict of interest with this work. For full disclosure statements refer to https://doi.org/10.1016/j.compmedimag.2022.102123. Marcel Breeuwer is an employee of Philips Medical Systems B.V. Cristian Lorenz and Jürgen Weese are employees of Philips GmbH Innovative Technologies.

Data availability

Data will be made available on request.

Acknowledgments

This research is a part of the openGTN project, supported by the European Union in the Marie Curie Innovative Training Networks (ITN) fellowship program under project No. 764465.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compmedimag.2022.102123.

References

Abbasi-Sureshjani, S., Amirrajab, S., Lorenz, C., Weese, J., Pluim, J., Breeuwer, M., 2020. 4D semantic cardiac magnetic resonance image synthesis on XCAT anatomical model. In: Medical Imaging with Deep Learning. PMLR, pp. 6–18.

Abraham, N., Khan, N.M., 2019. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE, pp. 683–687.

Amirrajab, S., Abbasi-Sureshjani, S., Al Khalil, Y., Lorenz, C., Weese, J., Pluim, J., Breeuwer, M., 2020. XCAT-GAN for synthesizing 3D consistent labeled cardiac MR images on anatomically variable XCAT phantoms. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 128–137.

Anwar, S.M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., Khan, M.K., 2018. Medical image analysis using convolutional neural networks: a review. J. Med. Syst. 42 (11), 1–13.

Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S., Yang, B., 2020. MedGAN: Medical image translation using GANs. Comput. Med. Imaging Graph. 79, 101684.

Baur, C., Albarqouni, S., Navab, N., 2018. MelanoGANs: high resolution skin lesion synthesis with GANs. arXiv preprint arXiv:1804.04338.

Beers, A., Brown, J., Chang, K., Campbell, J.P., Ostmo, S., Chiang, M.F., Kalpathy-Cramer, J., 2018. High-resolution medical image synthesis using progressively grown generative adversarial networks. arXiv preprint arXiv:1805.03144.

Bermudez, C., Plassard, A.J., Davis, L.T., Newton, A.T., Resnick, S.M., Landman, B.A., 2018. Learning implicit brain MRI manifolds with deep learning. In: Medical Imaging 2018: Image Processing, Vol. 10574. International Society for Optics and Photonics, p. 105741L.

Campello, V.M., Gkontra, P., Izquierdo, C., Martín-Isla, C., Sojoudi, A., Full, P.M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., et al., 2021. Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge. IEEE Trans. Med. Imaging.

Chartsias, A., Joyce, T., Dharmakumar, R., Tsaftaris, S.A., 2017. Adversarial image synthesis for unpaired multi-modal cardiac data. In: International Workshop on Simulation and Synthesis in Medical Imaging. Springer, pp. 3–13.

Chen, Q., Koltun, V., 2017. Photographic image synthesis with cascaded refinement networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1511–1520.

Chuquicusma, M.J., Hussein, S., Burt, J., Bagci, U., 2018. How to fool radiologists with generative adversarial networks? a visual turing test for lung cancer diagnosis. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, pp. 240–244.

Dosovitskiy, A., Brox, T., 2016. Generating images with perceptual similarity metrics based on deep networks. arXiv preprint arXiv:1602.02644.

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., Greenspan, H., 2018. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing 321, 321–331.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial networks. arXiv preprint arXiv:1406.2661.

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, X., Liu, M.-Y., Belongie, S., Kautz, J., 2018. Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 172–189.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. PMLR, 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.

Karimi-Bidhendi, S., Arafati, A., Cheng, A.L., Wu, Y., Kheradvar, A., Jafarkhani, H., 2020. Fully-automated deep-learning segmentation of pediatric cardiovascular magnetic resonance of patients with complex congenital heart diseases. J. Cardiovascul. Magn. Reson. 22 (1), 1–24.

Karras, T., Aila, T., Laine, S., Lehtinen, J., 2017. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196.

Kazeminia, S., Baur, C., Kuijper, A., van Ginneken, B., Navab, N., Albarqouni, S., Mukhopadhyay, A., 2020. GANs for medical image analysis. Artif. Intell. Med. 101938.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al., 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.

Liang, Y., Lee, D., Li, Y., Shin, B.-S., 2021. Unpaired medical image colorization using generative adversarial network. Multimedia Tools Appl. 1–15.

Lim, J.H., Ye, J.C., 2017. Geometric gan. arXiv preprint arXiv:1705.02894.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Liu, M.-Y., Breuel, T., Kautz, J., 2017. Unsupervised image-to-image translation networks. arXiv preprint arXiv:1703.00848.

Liu, Y., Chen, A., Shi, H., Huang, S., Zheng, W., Liu, Z., Zhang, Q., Yang, X., 2021. CT synthesis from MRI using multi-cycle GAN for head-and-neck radiation therapy. Comput. Med. Imaging Graph. 91, 101953.

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

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

Radford, A., Metz, L., Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

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.

Segars, W.P., Sturgeon, G., Mendonca, S., Grimes, J., Tsui, B.M., 2010. 4D XCAT phantom for multimodality imaging research. Med. Phys. 37 (9), 4902–4915.

Singh, N.K., Raza, K., 2021. Medical image generation using generative adversarial networks: A review. Health Inf.: Comput. Perspect. Healthc. 77–96.

Ulyanov, D., Vedaldi, A., Lempitsky, V., 2016. Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., Catanzaro, B., 2018b. 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, C., Xu, C., Wang, C., Tao, D., 2018a. Perceptual adversarial networks for image-to-image transformation. IEEE Trans. Image Process. 27 (8), 4066–4079.

Wolterink, J.M., Dinkla, A.M., Savenije, M.H., Seevinck, P.R., van den Berg, C.A., Išgum, I., 2017. Deep MR to CT synthesis using unpaired data. In: International Workshop on Simulation and Synthesis in Medical Imaging. Springer, pp. 14–23.

Yi, X., Walia, E., Babyn, P., 2019. Generative adversarial network in medical imaging: A review. Med. Image Anal. 58, 101552.

Zhang, G., Chen, K., Xu, S., Cho, P.C., Nan, Y., Zhou, X., Lv, C., Li, C., Xie, G., 2021. Lesion synthesis to improve intracranial hemorrhage detection and classification for CT images. Comput. Med. Imaging Graph. 90, 101929.

Zhang, Z., Yang, L., Zheng, Y., 2018. Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9242–9251.

Zhu, P., Abdal, R., Qin, Y., Wonka, P., 2020. Sean: Image synthesis with semantic region-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5104–5113.

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.