Medical Image Analysis 84 (2023) 102688

Contents lists available at ScienceDirect

Medical Image Analysis

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

On the usability of synthetic data for improving the robustness of deep learning-based segmentation of cardiac magnetic resonance images

Yasmina Al Khalil [a,*,1], Sina Amirrajab [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 magnetic resonance image
CMR synthesis
Domain adaptation and generalization
Image segmentation

ABSTRACT

Deep learning-based segmentation methods provide an effective and automated way for assessing the structure and function of the heart in cardiac magnetic resonance (CMR) images. However, despite their state-of-the-art performance on images acquired from the same source (same scanner or scanner vendor) as images used during training, their performance degrades significantly on images coming from different domains. A straightforward approach to tackle this issue consists of acquiring large quantities of multi-site and multi-vendor data, which is practically infeasible. Generative adversarial networks (GANs) for image synthesis present a promising solution for tackling data limitations in medical imaging and addressing the generalization capability of segmentation models. In this work, we explore the usability of synthesized short-axis CMR images generated using a segmentation-informed conditional GAN, to improve the robustness of heart cavity segmentation models in a variety of different settings. The GAN is trained on paired real images and corresponding segmentation maps belonging to both the heart and the surrounding tissue, reinforcing the synthesis of semantically-consistent and realistic images. First, we evaluate the segmentation performance of a model trained solely with synthetic data and show that it only slightly underperforms compared to the baseline trained with real data. By further combining real with synthetic data during training, we observe a substantial improvement in segmentation performance (up to 4% and 40% in terms of Dice score and Hausdorff distance) across multiple data-sets collected from various sites and scanner. This is additionally demonstrated across state-of-the-art 2D and 3D segmentation networks, whereby the obtained results demonstrate the potential of the proposed method in tackling the presence of the domain shift in medical data. Finally, we thoroughly analyze the quality of synthetic data and its ability to replace real MR images during training, as well as provide an insight into important aspects of utilizing synthetic images for segmentation.

1. Introduction

Deep learning (DL) methods have made a tremendous impact on a variety of visual tasks across many fields, including medical imaging, particularly in medical diagnostic and prognostic tasks (Lundervold and Lundervold, 2019). These methods have the ability to automatically model high-level discriminatory data features, crucial for object detection. DL models are data-driven, relying on a significant amount of annotated data with sufficient variation in relevant distinguishable image factors for training (Nalepa et al., 2019). Adequate variation in data ensures that the model captures a wide range of probable alterations and does not simply "memorize" the data seen during training. In fact, an effective DL model should perform robustly in the presence of unseen data, such that no unexpected increase in the testing error is observed compared to the training error (Abdollahi et al., 2020). However, the requirement of large and variable data-sets remains a significant obstacle for adaptation of DL models in medical image analysis domain.

Acquiring high-quality ground-truth data annotated by experts is a time-consuming process prone to human errors, as well as inter- and intra-annotator variability, but is also liable to constrained sharing policies (Yi et al., 2019; Hussain et al., 2017). Despite ongoing efforts across multiple healthcare institutions to develop a large open access

* Corresponding author.
E-mail addresses: y.al.khalil@tue.nl (Y. Al Khalil), s.amirrajab@tue.nl (S. Amirrajab), 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] Contributed equally.

https://doi.org/10.1016/j.media.2022.102688
Received 21 December 2021; Received in revised form 14 September 2022; Accepted 7 November 2022
Available online 17 November 2022
1361-8415/© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

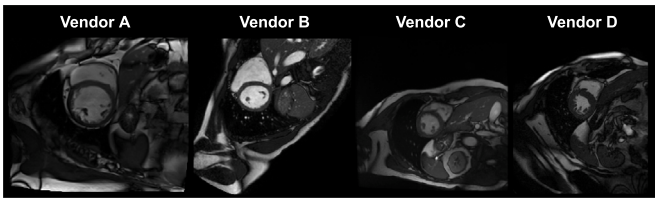

Fig. 1. Examples of variation observed in MR images acquired from different scanner vendors, taken directly from the data-set used in this study.

database, the above issues still hold and constrain the number of accessible images and annotations to researchers (Hosny et al., 2018). Consequently, the application of DL methods in clinically realistic environments results in poor generalization and performance, despite the expert-level performance achieved during development (Yasaka and Abe, 2018). A major reason for this is the existence of a domain shift (Kondrateva et al., 2021) in data acquired across different hospitals and scanners, such as the short-axis CMR images used in this study (Fig. 1).

Several approaches have been proposed so far to tackle the generalization and adaptation of DL models in the presence of limited data, including transfer-learning (Cheplygina et al., 2019; Ghafoorian et al., 2017), domain adaptation (Tzeng et al., 2017) and data augmentation (Nalepa et al., 2019). While transfer-learning in the form of fine-tuning a portion of pre-trained networks has shown significant improvements in the tasks involving natural images, it is limited in the medical imaging domain due to the lack of pre-trained models developed on large sets of medical data (Zhang et al., 2020). Domain adaptation addresses the development of models that can generalize to known target domains whose annotations are unknown or limited during training. However, the assumption is that examples from the target domain are available, which is not often the case with medical data (Choudhary et al., 2020; Zhang et al., 2020). On the other hand, observed domain shift properties can be "simulated" by applying a variety of data augmentation approaches in the image space, which has been shown across many fields (Zhang et al., 2016; Xu et al., 2020; Cubuk et al., 2019). Recent work performing latent space augmentation has also showed promising results in tackling the data domain shift (Chen et al., 2021; Jeong and Lee, 2021; Liu et al., 2018). However, we focus this work on image space augmentation.

Generation of synthetic images using generative adversarial networks (GANs) has recently emerged as a potential approach to data augmentation (Yi et al., 2019). A number of works have demonstrated their impressive capability to transfer appearance (style) from a set of images belonging to one domain to another domain (Frid-Adar et al., 2018; Chuquicusma et al., 2018; Wolterink et al., 2017; Chartsias et al., 2017), with most notable results in CT to MRI style transfer. More recent approaches, synthesizing realistic tissue appearance from the provided labels as input, referred to as conditional image synthesis (Mirza and Osindero, 2014; Qasim et al., 2020), have already shown a significant value in both computer vision and medical imaging. Using conditional synthesis we can artificially generate large data-sets of medical images, with enough variation to train robust models, while avoiding the problem of data anonymization. However, the ability of GAN synthesis approaches to represent more diverse tissue patterns, especially pathology, as well as plausible anatomy variations, has so far been limited.

In this work, we investigate the effectiveness and usability of a diverse synthesized database of realistic CMR images for MRI cardiac segmentation. The synthesized images are derived from anatomically plausible labels using a conditional GAN architecture, which leverages segmentation masks to guide the generation process and preserve the anatomical information contained in real images. Once trained, the model can synthesize realistic appearance on any given set of segmentation masks and generate a diverse set of realistic MR images. We present a detailed investigation of the quality and usability of such images with the aim to (1) handle data scarcity through either training a model with synthesized data only or through data augmentation with synthetic data and (2) observe the ability of such data to improve the generalization and adaptation of the model to variations appearing in multi-vendor and multi-center data. Through this, we want to understand how well can synthetic data replace real MRI data, but also gain insight in the current limitations of conditional GAN-based synthesis.

Compared to previous work in this area, we are among the first to address the benefits of utilizing synthetic images for CMR segmentation on images acquired across different scanner vendors and institutions and demonstrate its potential in addressing the domain shift occurring due to changes in acquisition. This is achieved through optimizing the realism, diversity and quality of synthetic images by utilizing an approach previously presented in Amirrajab et al. (2020a), extended with multi-tissue semantic segmentation module guiding conditional image synthesis. We additionally show that the generated synthetic images successfully replace missing data and display good potential to overcome challenges in medical image data scarcity. The main contributions of this paper are:

1. We present an optimized framework for MRI cardiac segmentation, which utilizes image synthesis to target segmentation generalization. The synthesis module is a substantial extension of the one presented in Amirrajab et al. (2020a). We introduce a heart region detection module to restrict the field of view (FOV) of images used for training the segmentation module, optimized for heart cavity segmentation and aided by the generated synthetic images.
2. We showcase the importance of handling images of the varying fields of view for improving the segmentation performance and eliminating false positive predictions.
3. We extensively assess the performance of our proposed model across multi-vendor and multi-site data and demonstrate the benefits of training with synthetic images when handling images that exhibit a domain shift due to differences in acquisition.
4. We quantify the effectiveness of synthetic images when used alone for training, as well as for augmentation.
5. Finally, we assess synthetic data usability for domain adaptation, where we replicate the style of unlabeled images and improve the segmentation performance on test data coming from the same source as the unlabeled data.

2. Related work

2.1. Image synthesis applied to medical image segmentation

While there are diverse approaches available in the literature addressing the task of medical image synthesis, this review focuses on GAN-based methods, as most relevant to the work presented in this paper. GANs show a strong potential to alleviate data scarcity and class imbalance limitations by generating realistic-looking images from a distribution that closely resembles the distribution of real data. The image generation is usually performed through either unconditional or conditional synthesis approaches. Unconditional approaches resemble the original GAN models, which are unsupervised in nature and typically generate data from a noise vector with limited influence on the output. Such approaches are replaced by conditional ones in medical imaging, as they allow infusing some useful prior information into the generation process and thus, provide more control over the generation procedure, producing more realistic images.

Following the success of GANs for synthesizing medical images, a number of works have attempted to utilize synthetic data in the tasks of classification and segmentation. Chartsias et al. (2017) uses a conditional GAN to synthesize CMR images from CT images and demonstrates that using synthetic data alongside real images results in a

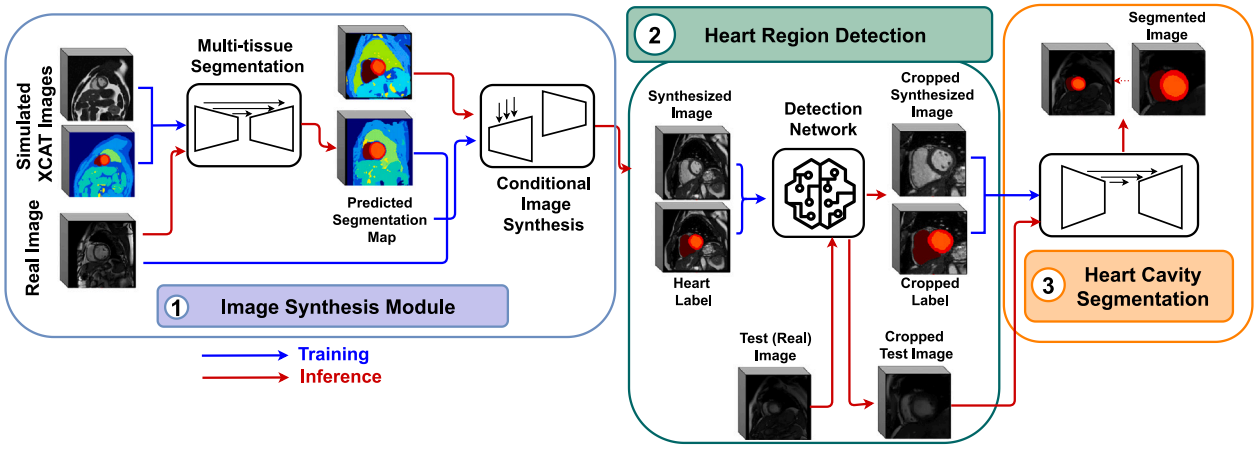

Fig. 2. An overview of the proposed synthesis and segmentation pipeline. The image synthesis module is based on a conditional image synthesis framework, utilizing both cavity labels and labels of the tissue surrounding the heart, generated by the multi-tissue segmentation network, to guide the synthesis and style transfer. A heart region detection module can be applied to detect a bounding box around the heart and crop the images accordingly. This step can boost the performance in terms of generalization. Finally, the synthesized images are used to train a network for the task of heart cavity segmentation. Note that the heart region detection module can be turned off, which means that the input images for training the segmentation module contain the original acquisition FOV of the heart.

better segmentation performance. Similarly, Zhang et al. (2018) utilize GANs for volume-based synthesis of MR volumes from corresponding CT volumes and vice versa and show that synthetic volumes obtained using their approach improve the segmentation performance on cardiovascular MRI. Other approaches (Costa et al., 2017; Vemulapalli et al., 2015; Yi et al., 2019) demonstrate that utilizing unsupervised translation can generate data that can overcome the problem of insufficient labeled data. Combined with transfer learning, unsupervised multi-modal translation has also been successfully utilized for adapting a pre-trained segmentation network to the data from a different modality, without any available annotations during training (Chen et al., 2019).

One of the main challenges observed in image synthesis, despite prior conditioning, is the fact that semantic information and spatial relations between different classes is often not retained (Park et al., 2019). This results in generated images that lack in realism, contain blurred regions, and are difficult to synthesize in high resolution. Moreover, the quality of synthetic data is still limited by the number of existing data-sets used for training, which introduces difficulties for generating high dimensional data reflecting realistic motion and volumetric changes. Consequently, synthetic data is only partly used during training — some methods utilize synthetic images for pre-training or weight initialization, while some amount of real images is later used to refine the model (Onishi et al., 2019). Finally, the generator often produces multiples of similar examples, which does not improve generalization, known as the mode collapse problem (Wang et al., 2017).

2.2. Improving generalization and adaptation of DL-based methods on multi-site and multi-vendor CMR images

There have been a number of works aimed at developing sophisticated deep learning approaches tackling CMR image segmentation on specific data-sets (Baumgartner et al., 2017; Bernard et al., 2018; Ammar et al., 2021; Pérez-Pelegrí et al., 2020; Yang et al., 2020). While these models demonstrate high performance on samples extracted from the same data-set, they have not been tested in cross-data settings. Initial approaches tackling generalization typically focus on training a model derived from one set (source domain) and testing it on other data-sets (Tran, 2016; Bai et al., 2018; Khened et al., 2019; Tao et al., 2019). However, these approaches either require re-training or fine-tuning or a collection of a large set of annotated data from multiple vendors and sites for training, which is impractical.

A recently organized M&Ms challenge is the first of its kind to tackle CMR segmentation on data from different centers, vendors, diseases and countries at the same time (Campello et al., 2021). Many approaches tackling generalization have been presented throughout this challenge, including domain adaptation methods (Acero et al., 2020), adversarial approaches (Scannell et al., 2021), disentangled representations (Liu et al., 2021) and utilization of specific processing blocks (Kong and Shadden, 2020) to alleviate the differences among domains. While some approaches have utilized unsupervised GANs for style transfer (Li et al., 2021; Zhang et al., 2021; Kong and Shadden, 2020; Li et al., 2020) to aid with training and transfer images from different domains to the same general style, there are rare attempts to utilize synthetic images directly during training. Moreover, the evaluation of the ability of synthetic data to replace real MR data during training is very rarely discussed, especially in the context of CMR segmentation.

3. Proposed method

A general overview of the complete pipeline proposed in this paper is shown in Fig. 2. The method consists of three main modules, image synthesis, heart region detection and heart cavity segmentation using synthetic data. We use a conditional image synthesis approach to generate realistic short-axis cardiac MR images. The quality of the generated images is improved by utilizing labels of various tissues surrounding the heart, which are typically present in the imaging FOV. These labels are generated by the multi-tissue segmentation network, trained on XCAT phantom-based simulated MR images, described in Section 3.1.1. We then utilize the synthesized images to train a CNN for the task of heart cavity segmentation, with separate segmentation maps of the right ventricle (RV), left ventricle (LV) and myocardium (MYO). To further improve the performance of the cavity segmentation network, we add a heart region detection module, used to detect a bounding box that encompasses the complete heart and accordingly crops the input image and its respective label ensuring that the heart is centralized in the cropped image.

A detailed description of each module is provided in the sections below. The proposed method visualized in Fig. 2 is just a general overview of the whole pipeline. However, in our experiments we utilize the synthesized data in a variety of ways to assess its effectiveness. Since the focus of this paper is on the usability of synthesized images for the task of segmentation, we do not describe the synthesis module in detail. A more comprehensive description can be found in Amirrajab et al. (2020a).

3.1. CMR data

3.1.1. Simulated data from variable XCAT phantoms

The first stage of our image synthesis module consists of a multi-tissue segmentation network that generates segmentation maps of anatomies typically present in the FOV of short-axis CMR images. Since the network is trained in a supervised manner, it requires the same number of tissue labels for corresponding organs in images during training. However, curating a database of real MRI data with such dense labels is a tedious process. Instead, we propose to train this network with the simulated cardiac MR images provided by the openGTN project,[2] consisting of 100 virtual subjects with diverse anatomical and contrast variations (Amirrajab et al., 2020b; Al Khalil et al., 2020a,b). The anatomy of each virtual subject is derived from 4D XCAT phantoms (Segars et al., 2010), while the simulation is based on Bloch equations for cine MR acquisition. Due to the versatility of the simulation process and XCAT phantoms, we generate segmentation maps with separate labels for the LV, RV, myocardium, lung, skeletal muscle, skin fat and abdominal organs. Fig. 2 shows an example of a simulated image and its respective multi-tissue segmentation map.

3.1.2. Multi-center, multi-vendor and multi-disease cardiac image segmentation challenge (M&Ms) data

The M&Ms[3] challenge data-set consists of 350 images from a mix of healthy controls and patients with hyptertrophic and dilated cardiomyopathies. All patients were scanned in clinical centers across three different countries (Spain, Germany and Canada) using four different MRI scanner vendors (Siemens, Philips, General Electric-GE and Canon). The provided training set contains 150 annotated patient scans from two different scanner vendors (Philips and Siemens, 75 each) and 25 un-annotated scans from a third vendor (GE). The in-plane resolution of the training images varies between 1.18 to 1.72 mm, with slice thickness ranging between 9.2 to 10.0 mm. Annotations have been provided by experienced clinicians at both end-diastolic and end-systolic phases, including contours for the left (LV) and right ventricle (RV) blood pools, as well as the left ventricular myocardium (MYO). This amounts to 300 annotated and 50 un-annotated images, taking both phases into consideration. We use the training set images from the M&Ms challenge for training both the synthesis and segmentation models. However, for training the synthesis module, we first generate multi-tissue maps using the multi-tissue segmentation network and combine them with original cavity labels provided in the M&Ms data-set.

For testing we utilize the additional images provided by the M&Ms challenge as a separate test-set. These consist of an additional 50 studies from each of the vendors provided, as well as another 50 studies from a vendor unseen during training (Canon), with in-plane resolution ranging from 0.68 to 1.8 mm. Due to different acquisition sources (centers and vendors), a domain shift between the data is expected. Some of these variations can be observed in Fig. 1 per vendor. In the rest of this paper, we refer to different data domains for data acquired with different scanner vendors, where Philips, Siemens, GE and Canon scanners are denoted as domain or vendor A, B, C and D, respectively.

3.2. Conditional image synthesis module

3.2.1. Multi-tissue segmentation

Synthetic images used in this study are generated through two consecutive units responsible for semantic image segmentation on real cardiac MR images and semantic image synthesis on the produced segmentation maps. We refer to the segmentation unit as a multi-tissue segmentation network, as it generates coarse segmentation maps of cardiac structures, but also of tissues typically surrounding the heart, such as lung, skeletal muscle, skin fat and abdominal organs. The multi-tissue segmentation module is utilized in two ways: (i) to provide labels for training the image synthesis network, and (ii) to provide labels on unseen data for synthesizing new examples using the trained network. Thus, the synthesis network utilizes multi-tissue labels with corresponding real MR images to learn the translation from the segmentation map to realistic MR contrast. While segmentation masks of only cavity tissue (LV, RV and MYO) can be used directly to train the synthesis network, our experiments show that utilizing multi-tissue maps synthesizes images of higher quality and better consistency in appearance for anatomical structures present in the FOV. We propose this module as the most convenient way of obtaining detailed tissue maps on images where such annotations are not available. However, it could be replaced by any other algorithm able to provide the same segmentation masks. Moreover, we hypothesize that introducing even more annotations and improving their accuracy could further improve the quality of synthetic images.

We adopt a U-Net architecture, completely trained on the XCAT simulated data-set (described in Section 3.1.1) with its multi-class ground truth masks. The network structure is similar to Ronneberger et al. (2015), with several changes introduced to optimize the network for the multi-tissue segmentation task. We utilize leaky ReLU and batch normalization (BN) after each convolutional layer to stabilize the training. Moreover, we apply dropout regularization (dropout rate of 0.5) to avoid over-fitting and boost generalization. The network consists of five down-sampling and up-sampling blocks, with a batch size of 32 2D CMR images fed to the input at each iteration, generating pixel-wise predictions for 9 tissue classes, including the background on

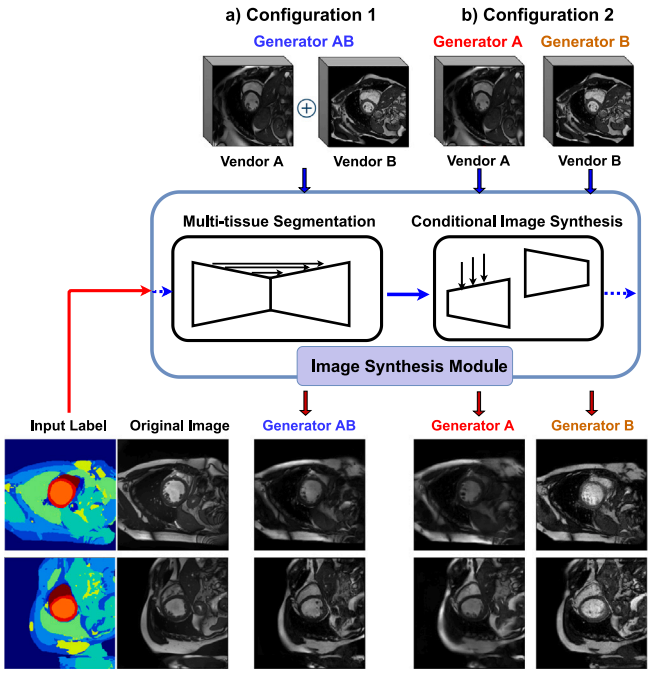

Fig. 3. Utilization of the conditional synthesis pipeline for the generation of synthetic images in this study. We synthesize images through two different configurations based on the domain and appearance of the data seen during the training of the synthesis module. Configuration 1 (a) refers to the training set-up where images from vendors A and B (M&Ms) are both seen during training. Configuration 2 (b) denotes a training set-up where only images from one vendor (A or B) are used during training. We note that synthesis through configuration 1 produces images with a combined style from A and B, while in configuration 2, the style or appearance of synthetic images is more similar to the images used during training.

[2] Simulated data can be accessed at https://opengtn.eu/database/ and https://osf.io/bkzhm/.

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

each slice. All images are resampled to 1.25×1.25 mm^2 across short-axis slices, cropped to the same size of 256×256, according to image center, and normalized with a mean of 0 and standard deviation of 1. Random scaling and rotations, mirroring and horizontal/vertical flips are applied on the fly during training. The network is trained using a Focal Tversky loss (Abraham and Khan, 2019) and optimized using Adam for stochastic gradient descent, with an initial learning rate of 10^{-4}. The choice of the loss function was determined experimentally (Al Khalil et al., 2020a), whereby we observe a significant improvement in performance compared to the standard cross-entropy and Dice losses. We hypothesize this could be due to extensive variation in shape and size of various tissue present in CMR images, while the Focal Tversky loss is specifically designed to handle such cases of class imbalance. We apply early stopping when the learning rate drops below 10^{-6} and train the network for a total of 350 epochs (determined by early stopping) using 200 simulated images for training. At test time, the trained model is applied on real MR images to obtain multi-tissue labels, examples of which can be seen in Fig. 3 marked under *Input Label*. Before segmentation, all real images are histogram matched to simulated images to tackle the presence of domain shift.

3.2.2. Conditional image synthesis

To synthesize data in this work, we utilize a conditional GAN synthesis approach, a mask-guided image generation technique that employs spatially adaptive denormalization (SPADE) layers (Park et al., 2019), reinforcing semantically-consistent image synthesis. The network is trained on paired real images and corresponding multi-tissue labels, estimated from the multi-tissue segmentation module, to learn the underlying modality-specific image characteristics for each tissue label. The main advantage of this approach is provided by the utilization of SPADE layers, as they inject information from the segmentation map throughout the network and thus, guide the generator to correctly learn the translation between the particular tissue class and its appearance in real MR images. More details in regards to the design and training of the conditional image synthesis GAN are provided in Amirrajab et al. (2020a) and Abbasi-Sureshjani et al. (2020).

The trained image synthesis module can be utilized to synthesize images of any style provided during training. Since we focus on investigating whether such approach could contribute to improving generalization and adaptation of segmentation models, we experiment with training the synthesis network in two different configurations, as outlined in Fig. 3. First, we combine acquired images from vendor A with images from vendor B, along with their respective multi-tissue labels predicted by the multi-tissue segmentation network, to train the **generator AB**, capable of synthesizing images with a mixture of style/appearance present in the images from the two vendors. We refer to this as *Configuration 1*. Alternatively, in *Configuration 2*, we train separate generators, **generator A** and **generator B**, to synthesize images of distinctive styles, reflecting the vendor A and vendor B appearances, respectively. Consistent with the training images, **generator A** produces synthetic images similar to real images from vendor A, while **generator B** outputs synthetic images similar to real images acquired from vendor B, which are brighter and with better contrast/quality compared to images from vendor A (see Fig. 1). In the attempt to better understand the distribution and diversity of the generated synthetic data in comparison to real images, we utilize the t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten and Hinton, 2008) for distribution visualization (see Appendix A and Fig. A1).

Note that all conditional image synthesis models are trained on images from the M&Ms data-set, where we remove the slices below apical and above basal location of the heart which do not contain any cavity ground-truth labels. This is done since our experiments show that without the presence of labels to guide the synthesis, the network struggles with synthesizing plausible appearance of anatomy.

3.3. Heart region detection

As shown in Fig. 2, the first step after generating synthetic images is the automatic detection of the heart region, whereby a bounding box is detected that encompasses the LV cavity, myocardium and RV cavity and is used for cropping images at both training and inference time. The training labels defining the bounding boxes are obtained from the ground truth masks in the M&Ms data-set by computing the smallest bounding box that fits the entire heart in the FOV and expanding it by 25 voxels in each dimension to incorporate some background. All bounding boxes are processed in a way to crop images of the size 128×128 voxels. Additionally, all images are resampled to a median spatial resolution of $1.25 \times 1.25 \times 10$ mm^3 before cropping. We use a simple convolutional neural network (CNN) designed for regression, where the output consists of 6 continuous values. The inputs to the network are 2D (256×256) mid-cavity slices extracted from the M&Ms training set, while the outputs consist of parameters defining the bounding box.

Inspired by the approach in Scannell et al. (2020), we first initialize the bounding box around the center of the image, with the assumption that the heart lies within a 100×100 voxel ROI defining this initial bounding box. The CNN is then trained to output the adjustment parameters so that it better fits the whole heart. In other words, the output of the CNN is the displacement in x and y directions of the center of the initialized ROI and its lower left corner, as well as the scaling factors for the width and height of the initial ROI. The CNN consists of five convolutional layers, followed by two fully-connected layers with a linear activation. Each convolutional layer uses 3×3 kernels, followed by a 2×2 max-pooling layer. Batch normalization and leaky ReLU activations are used in each layer, except for the output. Dropout with the probability of 0.5 is used in the fully-connected layers.

The network is trained for 2000 epochs with a batch size of 32 and early stopping (assessed from the validation accuracy), by minimizing the mean squared error between the computed transformation and the actual transformation (estimated from the ground-truth) using the Adam optimizer. We start with an initial learning rate of 0.001 but decrease it by a factor of 0.5 every 250 epochs using a scheduler. Note that all image dimensions and scaling/displacement parameters were normalized in a way to generate translations that are in the range from -1 to 1. Thus, after prediction, all the parameters need to be de-normalized to reflect the original image scale. The CNN for heart region detection is trained with a total of 750 2D mid-cavity slices extracted from the M&Ms train set from both vendor A and B. Input images are normalized to have the intensity values in the range of [0,1]. On-the-fly data augmentation is applied to the training images, consisting of random translation, rotation, scaling, vertical and horizontal flips, contrast augmentation and addition of noise. At inference time, we again use mid-cavity slices from the test images to obtain the adjustment parameters of the ROI. The predicted bounding boxes on mid-cavity slices are then propagated through the whole 3D volume, from which these slices were extracted. To evaluate the performance of the network, we calculate the mean Dice score (DSC) between the detected and manually extracted bounding box for the test set of 120 images. The final selected model performs with a mean DSC of 95.37 and standard deviation of 0.03. The cropped images obtained using the predicted bounding box are post-processed to be of the size 128×128 voxels, without any additional resampling, and used for training the segmentation networks. Expanding the bounding box to this size ensures that we do not crop any part of the actual heart tissue throughout the volume and avoids the need for any additional processing.

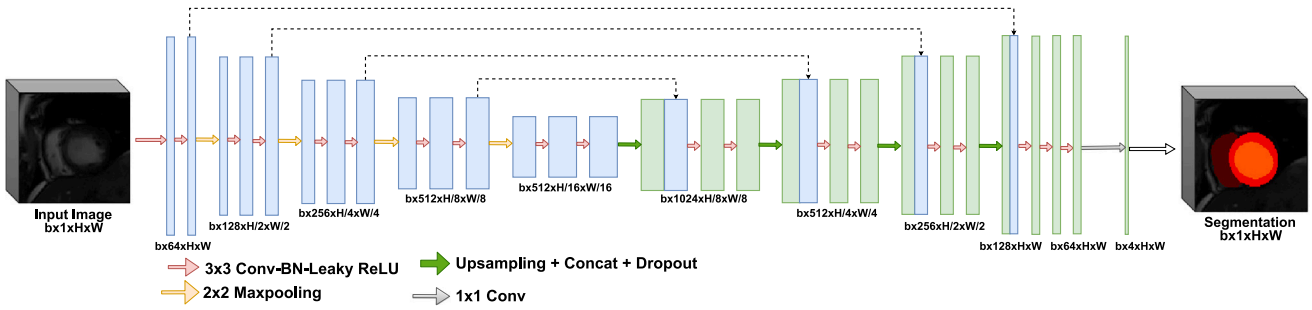

Fig. 4. An overview of the cardiac segmentation network structure. The U-Net takes a batch size of b 2D CMR images as input during each iteration and propagates it through a series of downsampling and upsampling blocks aimed at extracting multi-scale discriminatory features and generating 4 class pixel-wise predictions (backround, LV, MYO, RV). Each block consists of a convolutional layer (Conv), batch normalization (BN) and leaky rectified linear unit (Leaky ReLU). Note that in cases when the heart region detection module is utilized before training/inference, we use the exact architecture as shown above. Alternatively, we add one more block at the downsampling (along with the maxpool operation) and upsampling path to adjust for the increased FOV and image size.

3.4. Cardiac MR segmentation

3.4.1. Network architecture

We adopt a U-Net architecture for a multi-structure segmentation task, designed according to the recommendations in Isensee et al. (2021), illustrated in Fig. 4. We opt for a 2D U-Net for a number of reasons, including: (i) having the ability to work with images irrespective of their slice thickness or severe motion artifacts between slices, as well as unaligned slices, (ii) having a limited amount of data for training, while 3D models are typically data-hungry and (iii) the complexity of training a 3D network, due to its memory and time consumption, as well as a large number of parameters, often difficult to optimize.

The network consists of either five or six convolutional blocks, depending on whether the heart detection module is utilized, where output images are cropped to the size of 128×128 voxels. Such images are processed through five downsampling and upsampling convolutional blocks in total, with 4 max-pooling layers. Alternatively, if region detection is omitted, the input images seen during training are resized to 256×256 voxels and processed through a total of six downsampling and upsampling convolutional layers, with five max-pooling operations. Each convolutional block consists of 3×3 kernel convolutional layers, batch normalization and leaky ReLU activation. We apply batch normalization to improve regularization, but also generalization, as batch normalization on the adequate amount of images fed in one batch has the effect of restricting the distribution of the learned weights and thus, helps the network be less susceptible to noise and intensity variation. Moreover, we apply dropout regularization, with a rate of 0.5, after each concatenating operation to further avoid over-fitting.

3.4.2. Data processing pipeline

The first step of the processing pipeline is image resampling, essential to ensure that the proportion of the heart and the background in all images is relatively consistent. One of the main aspects by which images acquired across various scanners and sites differ is the FOV, causing significant variations in heart sizes. The pixel spacing in the images available for training ranges from 1.18 to 1.72 mm, while the range in the test set spans from 0.68 to 1.8 mm. We choose a value of 1.25×1.25 mm for resampling, across short-axis slices. After resampling, we crop all images to the size of 256×256 voxels. Further cropping is done if the heart region detection module is utilized. We normalize input images at both the training and inference time to an intensity range from [0,1]. This is followed by contrast stretching, which rescales the image intensity levels to include all intensities that fall within the 2nd and 98th percentile.

To increase robustness and cover a wide range of variations in heart pose and size, we augment the training set by applying:

- random horizontal and vertical flips ($p = 0.5$),
- random rotation by integer multiples of $\pi/2$ ($p = 0.5$),
- random scaling (scale factor $s \in [0.85,1.25]$, $p = 0.3$),
- random translations ($p = 0.5$) and random elastic deformations ($p = 0.3$).

All augmentations are applied on the fly during training. Intensity or contrast augmentations (such as random gamma correction) are not applied in this study, in order to better evaluate the variations in contrast supplemented by addition of synthetic data during training. At inference time, we only apply in-plane resampling, center cropping or cropping using the heart region detection, intensity normalization and contrast stretching.

3.4.3. Training

After pre-processing, the network is fed with batches of 58 images for training. We use a validation set to track the training progress and identify overfitting, where the same augmentation approach is applied to the validation set and the mean Dice score is calculated per each epoch. To train the network, we use a weighted sum of the categorical cross-entropy (L_{CE}) and Dice loss (L_{Dice}):

$$L_{CE} = -\sum_k y^k log(p^k), \qquad L_{Dice} = \sum_k 1 - \frac{2|P^k \cap Y^k|}{|P^k| + |Y^k|},$$
$$L_T = \lambda_1 L_{CE} + \lambda_2 L_{Dice}. \tag{1}$$

where k denotes a class, while p^k represents the predicted probability map per class. L_{Dice} measures the similarity between the probability map P^k and the ground truth Y^k. λ_1 and λ_2 are weighting parameters set at 0.6 and 0.4, respectively, to balance the contribution of the two losses. We use Adam for optimization, with an initial learning rate 10^{-4} and a weight decay of $3 * e^{-5}$. During training, the learning rate is reduced by a factor of 5 if the validation loss has not improved by at least $5 * 10^{-3}$ for 50 epochs. We apply early stopping on the validation set to avoid overfitting and select the model with the highest accuracy. All models are trained for a maximum of 1000 epochs, but are shown to converge within 250 to 350 epochs.

3.4.4. Post-processing

We perform a connected component analysis on the predicted labels and remove all but the largest connected component per class. Since test images are both resampled and cropped, we perform bilinear upsampling on the post-processed outputs of the network to recover the resolution back to the original.

4. Experiments

To assess the proposed segmentation pipeline and evaluate the usability of synthetic data generated in this study, we perform a series of experiments assessing its performance in challenging settings containing multi-domain and scarce data:

- First, we assess the quality of synthetic MR images by evaluating their ability to replace real images during training of segmentation models. To this end, we train a model with real MR images (denoted as **Real**), which we compare with a model trained using synthesized images, **Synth**. To study the effect of adding heart region detection to the training, we train two additional models **Real BB** and **Synth BB**, where **BB** stands for bounding box. More details about the experiment can be found in Section 4.1.
- In Section 4.2, we study the effects of augmentation with synthetic images (model **Synth Aug**) on segmentation robustness. We compare this to the baseline trained with standard data augmentation (**Baseline**), as well as the models trained with severe contrast transformations (denoted as **Style Aug** and **Synth + Style Aug**).
- Next, we evaluate the effectiveness of synthetic data for domain adaptation, where available unlabeled images (vendor C) are used to train a synthesis module to replicate their style (or contrast) on any given labels. Synthesized images with the style of vendor C are then used to augment the regular training set and train a model **Synth-C Aug**. The performance of this model is compared to models trained using histogram standardization (**Hist-C Aug**) and severe contrast transformations (**Style Aug**). Detailed experiment description is available in Section 4.3.
- Finally, in Section 4.4, we compare the performance of the models trained only with images acquired from vendor A or vendor B (**Real A** and **Real B**, respectively) to models where synthetic images generated in the style of vendor B or A are added to real vendor A or vendor B images (**Real_A_Synth_B** and **Real_B_Synth_A**, respectively). Our aim is to study the effectiveness of synthetic data in reducing the drop in segmentation performance when data is limited or data sharing is constrained.

4.1. Cardiac MR segmentation using images generated through conditional synthesis

In this experiment, we quantitatively evaluate the quality and usability of generated images for segmentation. First, we train two models without the heart region detection module:

- **Real**: a segmentation network trained on 300 real images from the M&Ms data-set, with half of the images acquired from vendor A and the other half from vendor B.
- **Synth**: a segmentation network trained on 300 synthesized images using two separate conditional GAN models per vendor, trained on the same data used for training the **Real** model. In other words, one generator synthesizes images with the style of vendor A on labels acquired from vendor B, while the other generator synthesizes the style learned from vendor B images on the labels from vendor A.

Next, we retrain **Real** and **Synth** with the addition of heart region detection as an extra pre-processing step (**Real BB** and **Synth BB**). Thus, the models are trained using images with a smaller FOV focused on the heart, where the whole heart is placed in the center of the image. All models are trained with the same training hyper-parameters and settings described in Section 3.4, using both ED and ES images. By utilizing the same labels and data sources for both models, we restrict the influence of variation in shape and anatomy during training, but allow for variation in appearance and image quality between synthetic and real images.

We evaluate the trained models on all images available in the test set, where vendor C and D images are not seen during training. This allows us to additionally observe the generalization ability of all models during segmentation. The test images are pre-processed with respect to the pre-processing set-up used during training, where the pipeline differs depending on the usage of the heart region detection module.

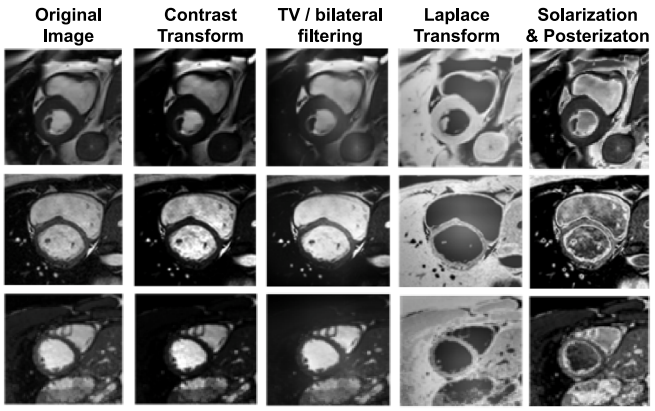

Fig. 5. Style transforms used as augmentation for training the **Style Aug** model (see Section 4.2). These include a combination of random brightness and contrast adjustment (**Contrast Transform**), total variation (TV) and bilateral filtering, Laplace transform and a combination of solarization and posterization.

4.2. Addressing data augmentation using synthesized images

In the next set of experiments, we evaluate whether the addition of synthetic images to the training set containing real MR images can further boost the segmentation performance. For this, we train a baseline model (**Baseline**) using a total of 600 real MR images, with 300 acquired from vendor A and the other 300 from vendor B. Note that for both vendors, we double the number of images by utilizing 150 images from each vendor twice in order to retain the same number of images used during training in all experiments. However, to ensure the network is not over-fitting and seeing the same images multiple times, we apply heavy data augmentation. In particular, we apply random horizontal and vertical flips, as well as rotation and translation with the probability of 0.7 instead of 0.5, we increase the probability of random scaling from 0.3 to 0.5 with a wider range for scale factor s ([0.60, 1.50]) and increase the probability of applying random elastic deformations from 0.3 to 0.5. Finally, we apply gamma transformation with a probability of 0.5 and γ randomly chosen from the range of [0.5,1.8].

We further train a model similar to the baseline, where we again use 600 images but with half of the images pre-processed differently before training. We refer to this as style augmentation and the model as **Style Aug**, where the focus is on introducing contrast diversity in the training set using different image processing techniques and thus, focus the optimization of the model towards the fundamental geometry features of the target tissue. Transformation methods are chosen arbitrarily with the main aim to increase the variety of training images. For this experiment, we selected a combination of random brightness and contrast adjustment, posterization, solarization, total variation (TV) and bilateral filtering, as well as the Laplacian transformation. While these methods have a potential to aid the robustness of segmentation models, they are mostly suitable under the assumption that the target shape is fairly consistent. Examples of transformed images can be observed in Fig. 5.

The previously described models are compared to a model augmented by adding a set of 300 synthesized images generated by a conditional GAN model trained using the setup proposed under configuration 1 (see Fig. 3) on both vendor A and B images (**generator AB**). In other words, to the existing training set of 300 vendor A and B real MR images, we add 300 synthesized images by the generator AB on labels extracted from vendors A and B. We refer to this model as **Synth Aug**. Finally, we train a model similar to **Synth Aug**, but with half of the train set augmented using the style augmentation techniques utilized for training the **Style Aug** model. By training such a model (**Synth + Style Aug**), we wish to observe the influence of

both augmentation methods and see if together they could add an additional benefit to the training and model generalization. Note that all models are trained using a heart region detection module in both the training and testing pipeline and evaluated on the same test set as in the previous experiments.

To evaluate the generalization and adaptation capabilities of the proposed pipeline to larger, more heterogeneous data-sets, we evaluate the method on other open-source CMR data-sets, recognized by the medical imaging community, which include the data acquired from the ACDC (Bernard et al., 2018) and M&Ms-2 challenges (Campello et al., 2021). The detailed description of the experiment, as well as the results, are available in Appendix D. Both data-sets contain a significant number of pathological abnormalities, typically challenging to segment, allowing us to evaluate whether the additional heterogeneity added by synthetic data allows for better adaptation during segmentation. However, please note that all evaluated models (**Real, Real BB and Synth Aug**) have not been re-trained with new data and thus, are expected to exhibit a performance drop due to the domain shift in comparison to M&Ms data.

4.3. Domain adaptation using synthesized images

Unlike domain generalization, domain adaptation assumes the ability to leverage some extent of the target information, which typically comes in the form of unlabeled data. We utilize 25 unlabeled images from unlabeled vendor C available during training in the M&Ms challenge and train a conditional generator to synthesize the style learned from vendor C (**generator C**) to labels acquired from vendors A and B. Note that the labels for vendor C images used during the synthesis pipeline are acquired solely from the multi-tissue segmentation module. Thus, we train a model **Synth-C Aug** using 300 images from vendor A and B with the addition of 150 synthetic images generated in the style of vendor C on randomly selected 75 masks from vendors A and B, respectively.

We compare the performance of the model adapted to vendor C using a synthesis pipeline to a model that is trained in a similar fashion, where instead of generated synthetic data, we utilize a histogram standardization approach to mimic the average intensity distribution of vendor C (model **Hist-C Aug**). Histogram standardization has become a common approach to tackle the domain shift appearing in medical images of the same modality, but acquired from different vendors and centers (Kushibar et al., 2021; Campello et al., 2021). For this purpose, we utilize a landmark-based histogram standardization approach proposed in Nyúl et al. (2000). For our application, we use 25 unlabeled images from vendor C to create a standardized landmark set according to which a randomly selected set of 75 images per vendor (A and B) are standardized and added to the training set containing 300 images from vendor A and B.

Finally, we train an additional model using 300 real images from vendor A and B, with the addition of 150 images undergoing severe contrast transformations used for training the **Style Aug** model, in order to observe if utilizing adaptation methods to target data in the two approaches proposed above truly has a benefit over training without the target data. We refer to this model as **Contrast Aug**. Our baseline is **Real BB**, first introduced in Section 4.1. The evaluation is done on vendor C and D test sets, but we pay special attention to the segmentation performance on vendor C. All models are trained using the same pre-processing and augmentation pipeline, as well as by utilizing the heart region detection module.

4.4. Addressing data scarcity using synthesized images

In theory, GANs enable the generation of both anonymous and potentially infinite data-sets based on a small number of available medical images (Singh and Raza, 2021; Yi et al., 2019). A commonly proposed set-up considers training specific generators at each site and sharing them between sites to synthesize data on labels available at each site. However, many argue that generators could learn patient-sensitive data at each site, which could further be extracted from the learned weights and thus, present a privacy issue (DuMont Schütte et al., 2021). An alternative approach is to restrict both the generator training and synthesis processes to each site, while sharing only the already synthesized data between sites to train a common segmentation model. The assumption here is that there exists a common data-set, accessible by all sites, which is devoid of any privacy concerns. Labels acquired from this common data-set would be shared across sites for the generators to synthesize on. To avoid the anatomical similarities of such images, we additionally randomly deform the labels.

In the first experiment under this paradigm, we assume that data from vendor A (150 labeled images) is a publicly available data-set, while data from vendor B is private, located at another site. A generator is trained at the site with private data (**generator B**) to synthesize images of style B on labels acquired from vendor A data-set. Using this data, we train:

- **Real_A**: trained on 150 images acquired from vendor A.
- **Real_A_Synth_B**: trained on 150 images acquired from vendor A and 150 synthetic images generated in the style of the images acquired from vendor B.

Furthermore, we reverse the experiment above and assume that vendor B images are publicly available, while vendor A images are private. Thus, we train the additional two models:

- **Real_B**: trained on 150 images acquired from vendor B.
- **Real_B_Synth_A**: trained on 150 images acquired from vendor B and 150 synthetic images generated in the style of the images acquired from vendor A, where we utilize all 150 deformed vendor B label masks for synthesis.

We hypothesize that models trained only with data acquired from one vendor will exhibit a significant drop in performance on data from multiple vendors. However, our goal is to assess if synthetic data generated in this work has the ability to replace real data in scenarios where data sharing is constrained.

5. Results

5.1. Cardiac MR segmentation using images generated through conditional synthesis

There are several observations we can derive by inspecting the results of utilizing synthetic data through the proposed pipeline, observed in Fig. 6 and Table B.1 (Appendix B). First, **Real** models tend to outperform the models trained on synthetic data and are generally more consistent in their predictions, as suggested by smaller standard deviation and less outliers. These outliers are further reduced by the introduction of the heart region detection module. In fact, heart region detection improves the performance across all models. Evaluation on images containing a large FOV, which cover tissues such as lung and abdominal area, often tends to produce false positive predictions, largely due to the presence of tissues similar in shape and appearance to heart cavity. Constraining the FOV reduces the impact of such tissue and further benefits the generalization of the model in the presence of domain shift. Given that heart region detection proves to be quite an easy task for a CNN to learn, while not significantly increasing the prediction time, we apply the module in all further experiments.

Despite **Real** models performing better, models trained on synthetic data show quite a remarkable ability to accurately segment real MR images across multiple sites/vendors. We observe that most of the errors appearing in **Synth** predictions stem from basal and apical slices, partly due to unlabeled basal and apical slices removed during training of these models, as explained in Section 3.2. While we observe an

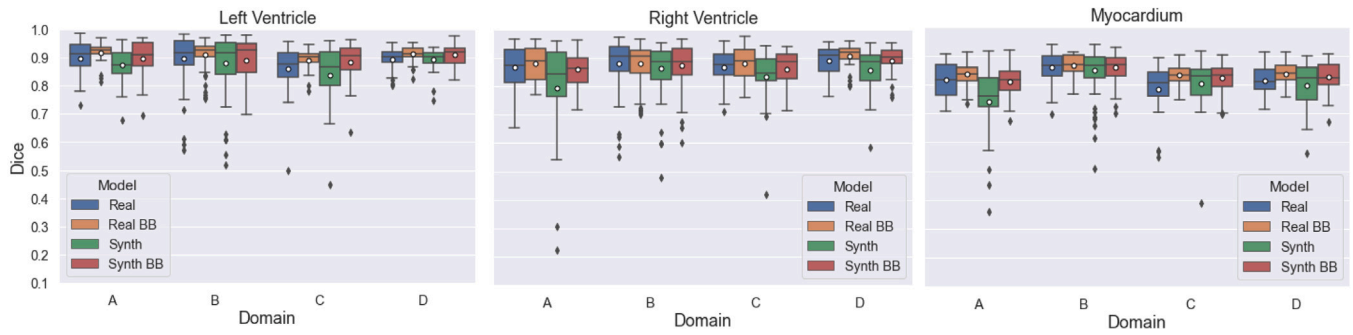

Fig. 6. Performance of the models trained with data synthesized on the M&Ms ground truth labels (**Synth**, n = 300) compared to the baseline trained on real M&Ms data (**Real**, n = 300), with and without the region detection module. Results of the models utilizing region detection are marked with **BB**. All models are evaluated on all three cavity tissues (left ventricle, myocardium and right ventricle).

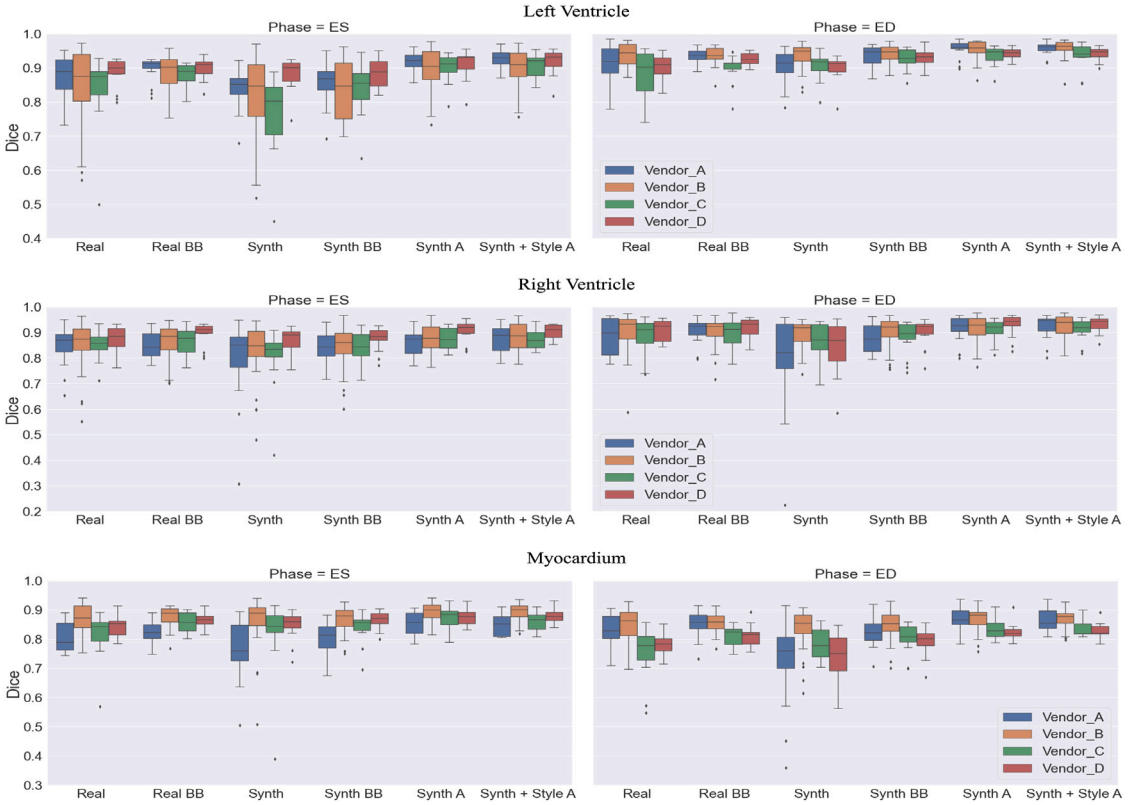

Fig. 7. Average Dice score for models trained on real (**Real and Real BB**) and synthetic data only (**Synth** and **Synth BB**) at end-systole (ES) and end-diastole (ED), per tissue and vendor. In addition, the performance of the models using synthetic data as augmentation (see Section 4.2) is shown, where Synth A stands for **Synth Aug** model and Synth + Style A for **Synth** + **Style Aug** model.

improvement in performance by introducing a heart region detection module, slices around the base and apex of the heart still remain the largest source of errors. Obtained results suggest that **Synth** models perform better on images acquired from vendors B and D, particularly in the case of RV and MYO segmentation, in comparison to their performance on other vendors and the LV. This could be a result of better contrast these images exhibit, with clear delineations and a quite homogeneous appearance per each cardiac tissue. However, **Real** models also under-segment such images, often performing worse than the **Synth** models. We hypothesize this is caused by more images subject to poor contrast, blurring effects and artifacts in synthetic datasets, which is a characteristic consequence of the synthesis and style transfer process.

We additionally investigate the contribution of ED and ES images on segmentation (see Fig. 7), where we observe that:

1. The segmentations at ED are more accurate for LV and RV, but not for MYO, with a better performance at ES across all models. This is partly due to myocardium becoming thicker at ES and thus, easier to segment.
2. Segmentation performance of **Synth** models at ED seems to significantly drop compared to ES, impacting the average performance of these models reported in Table B.1. However, **Synth BB** models improve this difference significantly, resulting in higher accuracy compared to ES that positively contributes to the overall performance.
3. The performance of the **Synth BB** model at ED for LV is at par or slightly better compared to **Real BB**. Similar can be observed for ES myocardium (vendor D), ED right ventricle (vendor B) and ED myocardium (vendor C).

Finally, both **Real** and **Synth** models are susceptible to errors in cases of myocardium thickening, as well as other pathological cases,

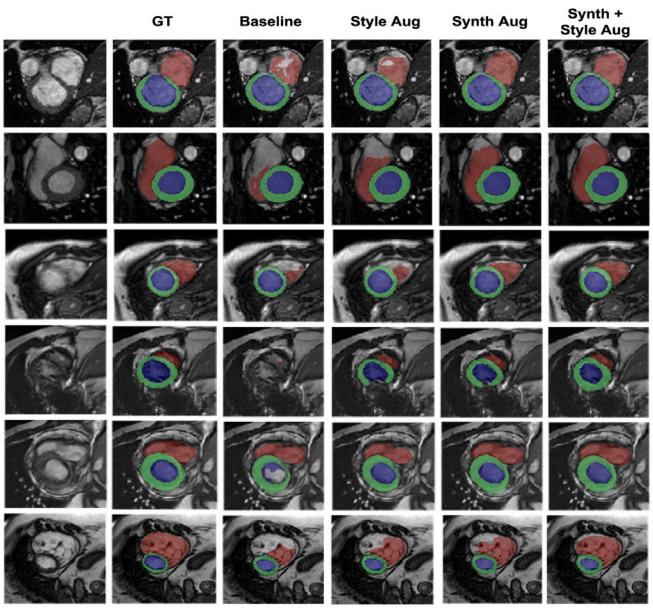

Fig. 8. Comparison of segmentation results in challenging cases with poor contrast and quality, as well as artifacts across four different testing vendors. We observe that RV is the most challenging to segment, which is especially problematic for slices around the base of the heart. However, the segmentation performance is consistently improved by augmenting the real MRI data with synthetic data (**Synth Aug**), with additional improvements obtained by adding style transformed data in **Synth + Style Aug**.

with **Real** models performing slightly better. This might be caused by an insufficient representation of such cases during the training of both synthesis and segmentation models. All in all, this experiment shows a strong indication that synthetic MR images generated in this study can replace real images during training of DL-based segmentation.

5.2. Addressing data augmentation using synthesized images

In this experiment we evaluate the influence of synthetic images utilized for data augmentation in a setup described in Section 4.2. From results in Fig. 9, we observe that the addition of synthetic data to the training has a positive impact on segmentation performance, both in terms of DSC and HD scores. While the **Style Aug** model also shows improvement, it is not able to reduce outlier examples compared to both **Synth Aug** and **Synth + Style Aug** , with **Synth + Style Aug** performing slightly better. However, according to the Wilcoxon signed-rank test, the differences between the two are not significant, while both are significantly better for almost all cases compared to the baseline and **Style Aug** (except for the myocardium DSC) in vendors C and D. Prominent differences are particularly seen in HD scores, caused by the reduction of false positive predictions, but also false negatives related to the right ventricle. Notably, augmentation shows the most prominent impact on the segmentation of the RV blood pool, which has also been visually confirmed, with some examples depicted in Fig. 8. Fig. 8 shows that both **Style Aug** and **Synth + Style Aug** models compensate for under-segmented areas, with **Synth + Style Aug** exhibiting better ability to handle tougher cases (see Fig. 8 rows 4 and 6), particularly around the base and apex of the heart. Other commonly occurring cases of under-segmentation improved by augmentation with synthetic data include areas affected by artifacts and blur, particularly impacting the LV due to blurring between the MYO and the blood pool (causing the cavity to appear smaller), in patients with thickened myocardium and in areas affected by brightness heterogeneity and lack of contrast. We hypothesize that the added synthetic images contain more examples of such cases, allowing the network to learn how to adequately handle a part of these issues. Similar behavior is confirmed when evaluating **Style Aug** and **Synth + Style Aug** models per ED and ES phases of the heart, shown in Fig. 7. We further deduce that **Synth + Style Aug** shows a steeper increase in performance compared to **Style Aug** at ES, while at ED the performance mostly remains the same. Interestingly, for myocardium segmented in vendors C and D, the addition of heavy style transformations reduces the performance at both ES and ED phases. Visual observations of the predictions for these cases indicates the over-segmentation of both LV and RV, which impacts the boundary areas of the myocardium, especially in the presence of trabeculations. Compared to the **Real** model, the biggest improvements in Dice scores when augmenting with synthetic data (**Synth Aug** model) are observed for LV, RV and MYO in vendors C, D and A with an increase up to 4.5% (0.88→0.92), 3.3% (0.90→0.93) and 3.6% (0.83→0.86), respectively. Similarly, the biggest reduction in Hausdorff distance is obtained for LV, MYO and RV in vendor D with a percentage decrease by 40.9% (12.2→7.2), 38.1% (18.1→11.2) and 39% (14.1→8.6), respectively. Appendix C and Fig. C.2 contain additional analysis in terms of two automatically derived clinical parameters with reference to manually derived ones, namely, the Bland–Altman plots of predicted and ground truth LV and RV ejection fraction (EF). Per-subject, the Bland–Altman analysis shows an improved agreement between the manual and automatically quantified EF using the proposed pipeline (**Synth Aug** model), compared to other approaches evaluated in this study.

We additionally perform a small study focused on bench-marking our approach to the state-of-the-art medical image segmentation methods available in the literature. In particular, we replace the segmentation network used in this study with the nnU-Net (Isensee et al., 2021; Full et al., 2020) model and evaluate the effect of adding different stages of the pipeline to nnU-Net training. In addition, by utilizing both 2D and 3D nnU-Net models, we gain additional insight on the effects of augmentation with synthetic data proposed in this study on 3D segmentation networks. We aim to follow the same training approach as proposed in Full et al. (2020). However, we do not apply ensembling of 2D and 3D models to obtain final results.

In total, we train 6 additional nnU-Net models that are directly compared to the results obtained by previously reported results obtained from **Real**, **Real BB** and **Synth-Aug** models. These include the **nnUnet Real 2D** and **nnUnet Real 3D** models, trained using the same data as the **Real** model (300 real CMR images acquired from vendors A and B) using 2D and 3D architecture, respectively; the **nnUnet BB 2D** and **nnUnet BB 3D** models, trained on the same data as **Real** and **Real BB** models with the addition of the heart region detection module; and the **nnUnet Synth Aug 2D** and **nnUnet Synth Aug 3D** models, trained by utilizing the heart region detection module and augmented with 300 synthetic images in the same manner as the **Synth Aug** model. The resulting Dice scores of all models are visualized in Fig. 10, acquired by evaluating the networks on the test set across all four vendors and cardiac tissues. We can make the following observations from the attained results: (i) both 2D and 3D nnU-Net models exhibit improvement when trained on images cropped around the heart area, where we observe a significant reduction in false positive predictions, (ii) the addition of synthetic data positively impacts the performance of both 2D and 3D nnU-Net models, whereby the 3D model seems to perform slightly better overall, although statistical significance is not observed, (iii) in some cases, nnU-Net models outperform the models belonging to the original pipeline, but the improvements in performance are not statistically significant when comparing the final models augmented with synthetic data.

Finally, the results obtained by evaluating the **Synth Aug** model on larger, external data-sets with more inherent heterogeneity, shown in Appendix D and Tables D.2 and D.3 suggest that augmentation with synthetic images significantly improves model generalization. In fact, this happens in spite of the model being trained on a completely different data-set than the testing set, suggesting that synthetic images add a significant amount of heterogeneity and variability to the training.

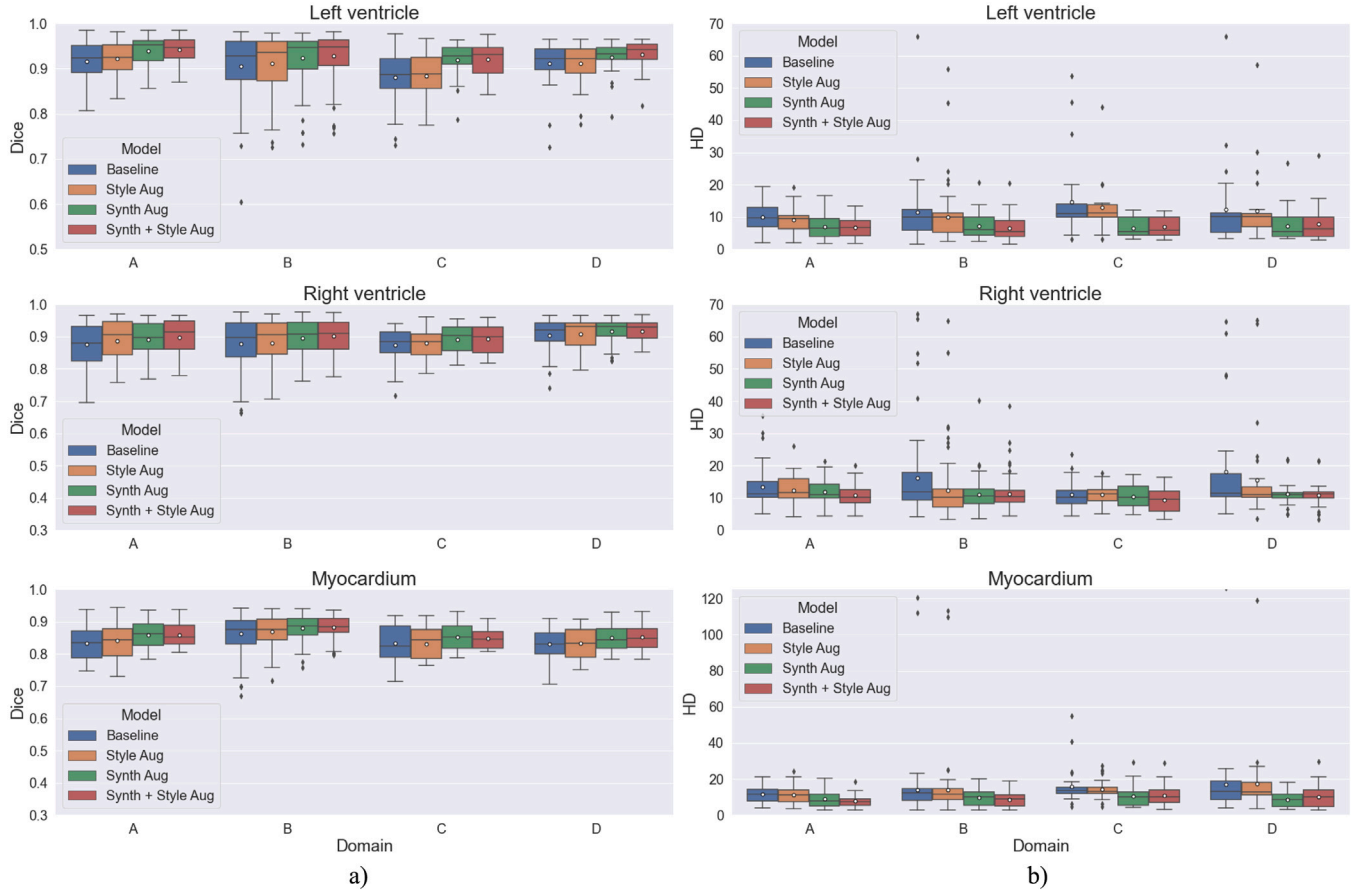

Fig. 9. Performance of the models augmented with synthetic data per tissue (**Synth Aug** and **Synth + Style Aug**) in terms of (a) Dice and (b) HD scores, compared to the baseline trained with classical data augmentation and a model trained with the addition of severe style transformations to the images (**Style Aug**). All models are tested on unseen data acquired from four different domains/vendors. Mean DSC and HD scores are indicated per model.

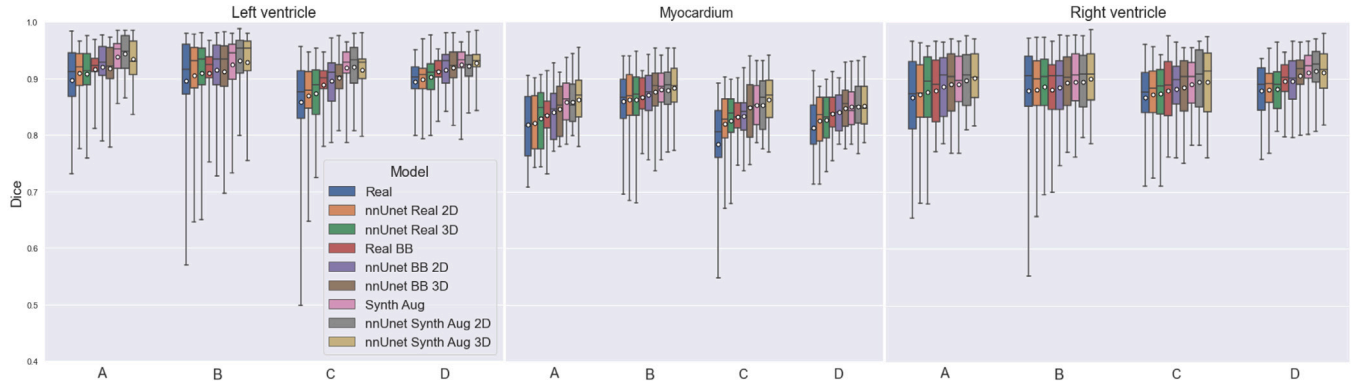

Fig. 10. Comparison in performance of the baseline model (**Real**), the model trained on images cropped around the heart region (**Real BB**) and the model augmented with synthetic data (**Synth Aug**) with the 2D and 3D nnU-Net segmentation models trained under the same conditions, as described in Section 5.2. The Dice score per model is reported across all four vendors and three cardiac tissues, obtained by performing the evaluation on the test set.

5.3. Domain adaptation using synthesized images

The evaluation of synthetic data utilization in a domain adaptation setup for segmentation, described in Section 4.3, can be observed in Table 1. Obtained results suggest that the addition of synthetically generated data with the style transferred from vendor C, outperforms the rest of the models by a significant margin, especially when evaluated on test images acquired from vendor C. While **Hist-C Aug** model introduces improvement in terms of both Dice and HD across vendor C, it does not add significant differences in terms of overlap for vendor D, compared to **Synth-C Aug**. We hypothesize that this is due to a more extensive variability added to the training when utilizing synthetic data, which introduces additional diversity in term of contrast and anatomy. Despite the synthesis procedure being mainly focused on generating accurate heart cavities, the surrounding tissue changes accordingly and introduces additional variation that helps with network regularization. This is especially pronounced when blur and other artifacts occur around tissue boundaries, as such examples are typically considered tough for segmentation, but benefit the training. Examples in Fig. 11 indicate that **Synth-C Aug** boosts the performance on challenging cases,

Table 1
Segmentation performance of models in a domain adaptation scenario, across test images acquired from vendors C and D, where a small set of unlabeled images from vendor C are available during training. A baseline model trained with classical augmentations is compared to a model trained with extensive style and contrast transformations (**Contrast Aug**), a model augmented with data histogram-matched to vendor (**Hist-C Aug**) and a model augmented with synthetic data generated using a generator trained to produce images of style C (**Synth-C Aug**). Last row presents the p-values obtained by a paired t-test between the **Synth-C Aug** and **Hist-C Aug** model, run under the null hypothesis that the **Synth-C Aug** model performs significantly better.

Model	Vendor C						Vendor D					
	LV		MYO		RV		LV		MYO		RV	
	DSC	HD	DSC	HD	DSC	HD	DSC	HD	DSC	HD	DSC	HD
Baseline	0.89 (0.04)	10.2 (7.5)	0.83 (0.04)	13.0 (8.3)	0.88 (0.06)	11.7 (5.0)	0.91 (0.03)	7.1 (3.2)	0.84 (0.04)	10.4 (4.8)	0.91 (0.04)	11.3 (3.4)
Contrast Aug	0.90 (0.06)	9.3 (6.6)	0.84 (0.03)	12.6 (7.1)	0.88 (0.06)	11.1 (7.2)	0.91 (0.04)	7.2 (3.2)	0.83 (0.04)	10.5 (4.6)	0.90 (0.04)	11.0 (3.2)
Hist-C Aug	0.91 (0.05)	7.9 (4.5)	0.84 (0.04)	12.5 (7.2)	0.89 (0.05)	10.8 (5.7)	0.91 (0.04)	6.8 (3.0)	0.83 (0.04)	9.9 (3.4)	0.91 (0.05)	10.8 (3.3)
Synth-C Aug	0.92 (0.03)	6.2 (2.6)	0.86 (0.03)	10.9 (7.9)	0.90 (0.03)	10.1 (4.1)	0.92 (0.02)	6.2 (3.2)	0.85 (0.03)	9.1 (3.1)	0.91 (0.04)	10.2 (2.9)
p-value	<0.01	<0.01	<0.01	<0.01	<0.01	0.029	<0.01	0.023	<0.01	0.031	0.131	0.027

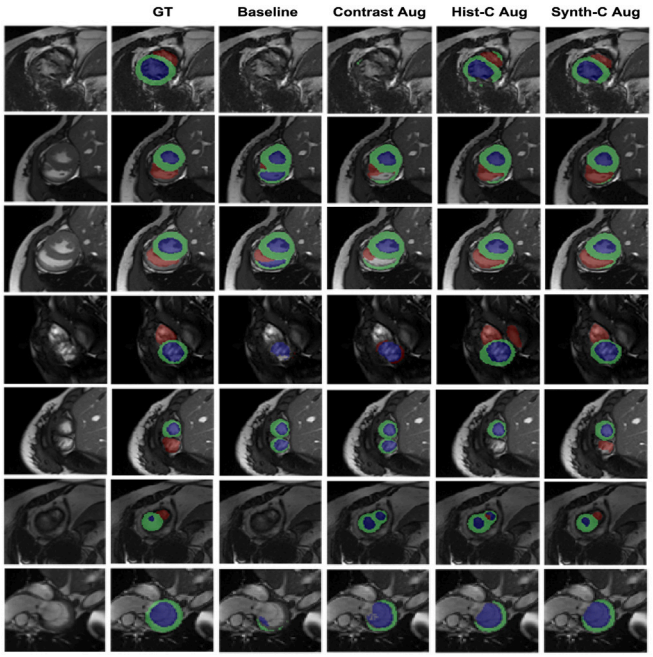

Fig. 11. Performance of models trained in a domain adaptation scenario on challenging cases from vendor C test set. For most cases, **Synth-C Aug** outperforms other models, while reducing both false positive and negative predictions.

where other models under-segment, as well as in slices around the base or apex of the heart.

5.4. Addressing data scarcity using synthesized images

This experiment focuses on assessing the capability of synthetic data to replace real data in settings when data is limited or privacy protection restricts data sharing. Results in Fig. 12 show that the addition of synthetic images generated from the style not originally included in the training set can significantly boost the model performance when evaluated across multiple domains ($p < 0.01$), except for vendor B left ventricle DSC scores. In particular, we note a significant drop in performance when **Real B** model is tested on vendor A images, especially for right ventricle and myocardium segmentation. The addition of synthetic data (**Real_B_Synth_A**) is able to compensate for this drop and reduce the number of outlier predictions, which negatively impacts both the DSC and HD scores. Such drop in performance is not observed for the **Real A** when evaluated on images from vendor B, suggesting the presence of a significant domain shift between the two data-sets and the lack of generalization capability of models trained with vendor B images only. These results suggest that utilizing synthetic data can aid the performance of the models across different acquisition sites and scanners, avoiding significant performance degradation.

6. Discussion

In this paper, we explore the usability of synthesized short-axis cardiac MR images, generated through the conditional synthesis pipeline (Section 3.2). Recent success of GANs for image synthesis and style transfer has been recognized and extended to the domain of medical imaging, as it holds a lot of promise for the generation of missing data and addressing the limitations of scarce data in most image analysis settings. Deficits in generalization to real-world data-sets with moderately different characteristics (distribution-shifts) represent one of the most common hurdles appearing due to scarce data. These deficits significantly affect the adoption of deep learning methods in clinical environments. However, realistically generated synthetic images could address these deficits, particularly when it comes to anonymization, protection of patient information and decreasing the cost of data collection.

While significant work has been done so far on the development of synthesis methods for medical images, there is a lack of research discussing the usability and benefits of the generated data. We show that by utilizing the method for image synthesis, previously proposed in Abbasi-Sureshjani et al. (2020) and optimized in Amirrajab et al. (2020a), we can significantly aid the performance of DL-based segmentation models in the task of heart cavity segmentation across data acquired from a variety of scanners and vendors. We run extensive experiments exploring the extent to which synthesized data can replace real data during training of a deep learning-based segmentation model in a challenging setting where the testing data is unseen during the training process. Our experiments show that synthesized data exhibits a strong potential to aid and replace real data during training, which could be an important way to tackle data scarcity due to data sharing restrictions. In fact, our results demonstrate an improvement in Dice and Hausdorff distance scores up to 4% and 40%, respectively, across augmentation and domain adaptation experiments performed in this work.

Given that the synthesis pipeline can be trained to generate images of any style or appearance, the generated images can be utilized to serve as a more efficient way of augmenting data and tackling the problem of varying appearance due to differences in acquisition, common in MR images. Consequently, this makes the models more robust and able to generalize to unseen data. We show this through experiments in Sections 4.2 and 4.3, where we explore the effect of boosting the training set with synthetic data in comparison to other commonly used approaches for both augmentation and adaptation. A similar behavior is observed when augmenting the training of state-of-the-art segmentation models, such as the 2D and 3D nnU-Net. Statistically significant improvements are further observed on other external data-sets, such as the ACDC (Bernard et al., 2018) and M&Ms-2 (Campello et al., 2021) data, as shown in Appendix D, where the variability provided by synthetic images improves the adaptation of the network to pathological tissue. Another application of such data could be in the domain of federated learning, where instead of acquiring real data

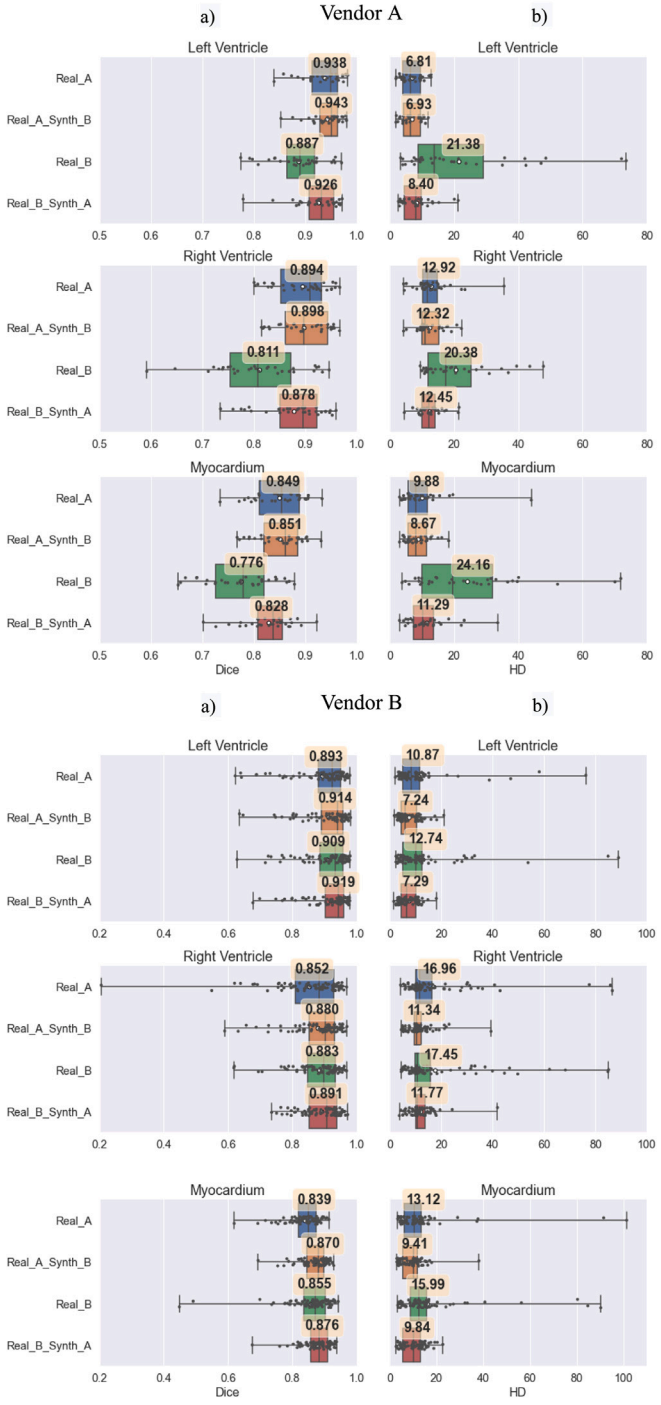

Fig. 12. Utilizing synthetic data to compensate for data scarcity. Models **Real A** and **Real B** are compensated with data synthesized with style generated from unavailable vendor B and vendor A data, respectively (**Real_A_Synth_B** and **Real_B_Synth_A**). The performance is reported across test data from vendors A and B in terms of (a) Dice and (b) HD scores for all cardiac tissues. Mean DSC and HD scores per model are reported.

for training from multiple sites, a synthesis module could be deployed to generate synthetic data of a similar style present in each site. At inference time, different models could be used to synthesize images of different appearance on an available set of labels and thus avoid the problem of data anonymization and privacy protection. A small example of how this could work is shown in Section 4.4.

A limitation of this work is that we cannot assess how well synthesized images can tackle cases containing pathology, in spite of some positive results obtained on other data-sets containing pathology. Similarly, we have not yet attempted to synthesize images containing pathology, due to the lack of labels and pathological data. However, since we show that images of good quality can be synthesized provided there are enough labels present during training, we strongly believe that this method could be extended to cases containing pathology as well and enrich the scarce data-sets currently available. Furthermore, we note that the current segmentation pipeline often shows a degradation in performance at the slices around the base and apex of the heart. This is a common problem in all heart cavity segmentation models, but is a bit more prominent in our case since we were not able to synthesize accurate appearance and anatomy in those slices. We hypothesize this is due to the lack of ground truth labels for the tissues surrounding the heart in those slices, as well as inter- and intra-examiner disagreement in those areas, common across different data-sets.

In addition, while we provide extensive experiments and report commonly used segmentation metrics, we are limited by the lack of confidence or uncertainty evaluation of the proposed method. However, since this is still an open-research question in the medical imaging community (Mehrtash et al., 2020), we plan to focus some of our future work on understanding how to properly assess the uncertainty of the proposed method and augmentation with synthetic data, in general. This also involves a proper assessment of the quality of generated images, which we currently evaluate through their usability in a medical image analysis task. While a number of measurements have been so far utilized for the quality assessment of medical images, such as the Inception score, Frechet Inception Distance, structural similarity index, normalized root mean squared error (Skandarani et al., 2021; Tronchin et al., 2021), as well as scoring based on human visual perception, it is still difficult to objectively measure the quality of GAN-derived medical images. Finally, while the extraction of heart region proves to be a very important step in tackling the generalization issue of segmentation models, especially in cases where the FOV varies significantly, it is not yet fully beneficial when applied on synthetic images. We attribute this to the artifacts and blurring effects appearing at the edges between cavity tissue, which are not that apparent when the FOV is considerably larger. This is indicated by the visual assessment of the results, where the performance of models trained on synthetic data is degraded at the edges between the myocardium and the left ventricle, as well as the myocardium and the right ventricle.

Future work includes tackling some of the above-mentioned issues, as well as adapting the whole pipeline on cardiac images containing pathology. We hypothesize that current results can be further improved by introducing additional augmentation and processing steps, aimed at decreasing the susceptibility to variations in appearance and shape. Moreover, introducing plausible anatomical variations to the tissue masks at the input of the synthesis generator could additionally contribute to the diversity of generated images. This could be done through geometrical transformations applied on the masks, but also through more informed approaches, such as auto-encoders, able to capture the realistic variance in the anatomy of cardiac tissue. On the other hand, we plan to focus on better inference of what data should be synthesized based on the characteristics of the data distribution in the training set or testing set, as well as prior knowledge of the segmentation task at hand. An example of this is addressing basal and apical slices of the heart, which are typically under-represented compared to other slices across the heart volume. Furthermore, we wish to extend this method to other modalities and organs, as well as repeat our experiments on a larger clinical data-set, undergoing a wider range of variation. To do this, we plan to increase the resolution and representational power of the conditional synthesis approach presented, in order to generate higher quality and more diverse synthetic images. One way to possibly tackle this is by fusing the synthesis and segmentation approaches and train those modules, at least partly, together.

Finally, as shown in Sun et al. (2022) and Hu et al. (2022), 2D synthesis approaches can be potentially extended to generate true 3D

volumes, while retaining memory efficiency and the ability to generate high resolution images. However, this might not be straightforward for CMR synthesis, as we hypothesize that the target image shape and acquisition type highly influence the choice of the network architecture. Cardiac cine short-axis and long-axis images are characterized by large slice thickness and inter-slice gaps and are often affected by breath-hold related motion artifacts between consecutive slices. This could hamper the effectiveness of 3D approaches, particularly caused by the lack of true volumetric information in such images. Moreover, conditional approaches typically rely on tissue labels for both training and inference. In images with large slice thickness, the labels are often patchy and suffer from ”blocky” effects when observed in 3D, which we hypothesize would have more detrimental effects on the training of 3D models. Increasing the through-plane resolution in such images could mitigate this. Thus, while an extension to 3D synthesis using GANs is essential due to the prevalence of 3D imaging techniques, their application needs to be carefully studied in data such as short-axis CMR images.

7. Conclusion

In this paper, we show that synthetic images generated through the proposed conditional synthesis framework can benefit medical image analysis, especially in cases where data is limited or missing. In particular, we use these images to aid the training of a deep learning-based method for the task of heart cavity segmentation from short-axis cardiac MR images. We are able to generate high quality, semantically consistent and anatomically plausible images by utilizing a method that benefits from segmentation-conditioned normalization layers. We first demonstrate that a model trained with synthetic images only is able to achieve competitive performance when evaluated on a test set of real cardiac MR images and compared to the models trained with real data only. Furthermore, we show that utilizing synthetic images aids network generalization and adaptation to data from varying scanners and vendors. Synthetic images show a strong potential to address privacy issues with respect to data sharing (e.g. in federated learning).

This approach has a potential to tackle challenging cases with low quality and poor contrast, as well as pathological cases. Finally, we demonstrate that synthetic data can successfully tackle data scarcity and achieve competitive performance combined with data acquired from a single site or scanner only. Although models trained with the addition of synthetic images still undergo a decrease in performance on apical and basal slices containing complex structures and extremely tiny objects, as well as images with significant deformations, they demonstrate a strong potential for resolving major issues deep learning models are facing in the domain of medical image analysis.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Marcel Breeuwer is an employee of Philips Medical Systems B.V. Cristan 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.media.2022.102688.

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.

Abdollahi, B., Tomita, N., Hassanpour, S., 2020. Data augmentation in training deep learning models for medical image analysis. In: Deep Learners and Deep Learner Descriptors for Medical Applications. Springer International Publishing, Cham, pp. 167–180. http://dx.doi.org/10.1007/978-3-030-42750-4_6.

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.

Acero, J.C., Sundaresan, V., Dinsdale, N., Grau, V., Jenkinson, M., 2020. A 2-step deep learning method with domain adaptation for multi-centre, multi-vendor and multi-disease cardiac magnetic resonance segmentation. In: International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, pp. 196–207.

Al Khalil, Y., Amirrajab, S., Lorenz, C., Weese, J., Breeuwer, M., 2020a. Heterogeneous virtual population of simulated CMR images for improving the generalization of cardiac segmentation algorithms. In: International Workshop on Simulation and Synthesis in Medical Imaging. Springer, pp. 68–79.

Al Khalil, Y., Amirrajab, S., Lorenz, C., Weese, J., Breeuwer, M., 2020b. Simulated CMR images can improve the performance and generalization capability of deep learning-based segmentation algorithms. In: Proceedings of the 28th Annual Meeting ISMRM 2020.

Amirrajab, S., Abbasi-Sureshjani, S., Al Khalil, Y., Lorenz, C., Weese, J., Pluim, J., Breeuwer, M., 2020a. 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.

Amirrajab, S., Al Khalil, Y., Lorenz, C., Weese, J., Breeuwer, M., 2020b. Towards generating realistic and hetrogeneous cardiac magnetic resonance simulated image database for deep learning based image segmentation algorithms. In: Proceedings of the 12th Annual Meeting ISMRM Benelux Chapter 2020. p. 077.

Ammar, A., Bouattane, O., Youssfi, M., 2021. Automatic cardiac cine MRI segmentation and heart disease classification. Comput. Med. Imaging Graph. 88, 101864.

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

Baumgartner, C.F., Koch, L.M., Pollefeys, M., Konukoglu, E., 2017. An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation. In: International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, pp. 111–119.

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., et al., 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Trans. Med. Imaging 37 (11), 2514–2525.

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, C., Hammernik, K., Ouyang, C., Qin, C., Bai, W., Rueckert, D., 2021. Cooperative training and latent space data augmentation for robust medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 149–159.

Chen, C., Ouyang, C., Tarroni, G., Schlemper, J., Qiu, H., Bai, W., Rueckert, D., 2019. Unsupervised multi-modal style transfer for cardiac MR segmentation. In: International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, pp. 209–219.

Cheplygina, V., de Bruijne, M., Pluim, J.P., 2019. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med. Image Anal. 54, 280–296.

Choudhary, A., Tong, L., Zhu, Y., Wang, M.D., 2020. Advancing medical imaging informatics by deep learning-based domain adaptation. Yearb. Med. Inform. 29 (1), 129.

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.

Costa, P., Galdran, A., Meyer, M.I., Abramoff, M.D., Niemeijer, M., Mendonça, A.M., Campilho, A., 2017. Towards adversarial retinal image synthesis. arXiv preprint arXiv:1701.08974.

Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V., 2019. Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 113–123.

DuMont Schütte, A., Hetzel, J., Gatidis, S., Hepp, T., Dietz, B., Bauer, S., Schwab, P., 2021. Overcoming barriers to data sharing with medical image generation: a comprehensive evaluation. Npj Digit. Med. 4 (1), 1–14.

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.

Full, P.M., Isensee, F., Jäger, P.F., Maier-Hein, K., 2020. Studying robustness of semantic segmentation under domain shift in cardiac MRI. In: International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, pp. 238–249.

Ghafoorian, M., Mehrtash, A., Kapur, T., Karssemeijer, N., Marchiori, E., Pesteie, M., Guttmann, C.R., de Leeuw, F.-E., Tempany, C.M., et al., 2017. Transfer learning for domain adaptation in mri: Application in brain lesion segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 516–524.

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H., Aerts, H.J., 2018. Artificial intelligence in radiology. Nat. Rev. Cancer 18 (8), 500–510.

Hu, Q., Li, H., Zhang, J., 2022. Domain-adaptive 3D medical image synthesis: An efficient unsupervised approach. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 495–504.

Hussain, Z., Gimenez, F., Yi, D., Rubin, D., 2017. Differential data augmentation techniques for medical imaging classification tasks. In: AMIA Annual Symposium, Vol. 2017. American Medical Informatics Association, p. 979.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18 (2), 203–211.

Jeong, S., Lee, S., 2021. Biased extrapolation in latent space forimbalanced deep learning. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 337–346.

Khened, M., Kollerathu, V.A., Krishnamurthi, G., 2019. Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. Med. Image Anal. 51, 21–45.

Kondrateva, E., Pominova, M., Popova, E., Sharaev, M., Bernstein, A., Burnaev, E., 2021. Domain shift in computer vision models for MRI data analysis: an overview. In: Thirteenth International Conference on Machine Vision, Vol. 11605. International Society for Optics and Photonics, 116050H.

Kong, F., Shadden, S.C., 2020. A generalizable deep-learning approach for cardiac magnetic resonance image segmentation using image augmentation and attention U-net. In: International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, pp. 287–296.

Kushibar, K., Salem, M., Valverde, S., Rovira, À., Salvi, J., Oliver, A., Lladó, X., 2021. Transductive transfer learning for domain adaptation in brain magnetic resonance image segmentation. Front. Neurosci. 15.

Li, H., Zhang, J., Menze, B., 2020. Generalisable cardiac structure segmentation via attentional and stacked image adaptation. In: International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, pp. 297–304.

Li, L., Zimmer, V.A., Ding, W., Wu, F., Huang, L., Schnabel, J.A., Zhuang, X., 2021. Random style transfer based domain generalization networks integrating shape and spatial information. In: International Workshop on Statistical Atlases and Computational Models of the Heart. pp. 208–218.

Liu, X., Thermos, S., Chartsias, A., O'Neil, A., Tsaftaris, S.A., 2021. Disentangled representations for domain-generalized cardiac segmentation. In: International Workshop on Statistical Atlases and Computational Models of the Heart. pp. 187–195.

Liu, X., Zou, Y., Kong, L., Diao, Z., Yan, J., Wang, J., Li, S., Jia, P., You, J., 2018. Data augmentation via latent space interpolation for image classification. In: 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, pp. 728–733.

Lundervold, A.S., Lundervold, A., 2019. An overview of deep learning in medical imaging focusing on MRI. Z. Med. Phys. 29 (2), 102–127.

Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T., 2020. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Trans. Med. Imaging 39 (12), 3868–3878.

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

Nalepa, J., Marcinkiewicz, M., Kawulok, M., 2019. Data augmentation for brain-tumor segmentation: a review. Front. Comput. Neurosci. 13, 83.

Nyúl, L.G., Udupa, J.K., Zhang, X., 2000. New variants of a method of MRI scale standardization. IEEE Trans. Med. Imaging 19 (2), 143–150.

Onishi, Y., Teramoto, A., Tsujimoto, M., Tsukamoto, T., Saito, K., Toyama, H., Imaizumi, K., Fujita, H., 2019. Automated pulmonary nodule classification in CT images using a deep convolutional neural network trained by generative adversarial networks. BioMed Res. Int. 2019.

Park, T., Liu, M.-Y., Wang, T.-C., Zhu, J.-Y., 2019. Semantic image synthesis with spatially-adaptive normalization. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, Los Alamitos, CA, USA, pp. 2332–2341.

Pérez-Pelegrí, M., Monmeneu, J.V., López-Lereu, M.P., Ruiz-España, S., Del-Canto, I., Bodí, V., Moratal, D., 2020. PSPU-net for automatic short axis cine MRI segmentation of left and right ventricles. In: 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE). IEEE, pp. 1048–1053.

Qasim, A.B., Ezhov, I., Shit, S., Schoppe, O., Paetzold, J.C., Sekuboyina, A., Kofler, F., Lipkova, J., Li, H., Menze, B., 2020. Red-GAN: Attacking class imbalance via conditioned generation. Yet another medical imaging perspective. In: Medical Imaging with Deep Learning. PMLR, pp. 655–668.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, Cham, pp. 234–241.

Scannell, C.M., Chiribiri, A., Veta, M., 2021. Domain-adversarial learning for multi-centre, multi-vendor, and multi-disease cardiac MR image segmentation. In: International Workshop on Statistical Atlases and Computational Models of the Heart. pp. 228–237.

Scannell, C.M., Veta, M., Villa, A.D., Sammut, E.C., Lee, J., Breeuwer, M., Chiribiri, A., 2020. Deep-learning-based preprocessing for quantitative myocardial perfusion MRI. J. Magn. Reson. Imaging 51 (6), 1689–1696.

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 Inform.: A Comput. Perspect. Healthc. 77–96.

Skandarani, Y., Lalande, A., Afilalo, J., Jodoin, P.-M., 2021. Generative adversarial networks in cardiology. Can. J. Cardiol..

Sun, L., Chen, J., Xu, Y., Gong, M., Yu, K., Batmanghelich, K., 2022. Hierarchical amortized GAN for 3D high resolution medical image synthesis. IEEE J. Biomed. Health Inf. 26 (8), 3966–3975.

Tao, Q., Yan, W., Wang, Y., Paiman, E.H., Shamonin, D.P., Garg, P., Plein, S., Huang, L., Xia, L., et al., 2019. Deep learning–based method for fully automatic quantification of left ventricle function from cine MR images: a multivendor, multicenter study. Radiology 290 (1), 81–88.

Tran, P.V., 2016. A fully convolutional neural network for cardiac segmentation in short-axis MRI. arXiv preprint arXiv:1604.00494.

Tronchin, L., Sicilia, R., Cordelli, E., Ramella, S., Soda, P., 2021. Evaluating GANs in medical imaging. In: Deep Generative Models, and Data Augmentation, Labelling, and Imperfections. Springer, pp. 112–121.

Tzeng, E., Hoffman, J., Saenko, K., Darrell, T., 2017. Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7167–7176.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (11).

Vemulapalli, R., Van Nguyen, H., Zhou, S.K., 2015. Unsupervised cross-modal synthesis of subject-specific scans. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 630–638.

Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X., Wang, F.-Y., 2017. Generative adversarial networks: introduction and outlook. IEEE/CAA J. Autom. Sin. 4 (4), 588–598.

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.

Xu, J., Li, M., Zhu, Z., 2020. Automatic data augmentation for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 378–387.

Yang, X., Zhang, Y., Lo, B., Wu, D., Liao, H., Zhang, Y., 2020. DBAN: Adversarial network with multi-scale features for cardiac MRI segmentation. IEEE J. Biomed. Health Inf..

Yasaka, K., Abe, O., 2018. Deep learning and artificial intelligence in radiology: Current applications and future directions. PLoS Med. 15 (11), e1002707.

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

Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O., 2016. Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530.

Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., Wood, B.J., Roth, H., Myronenko, A., et al., 2020. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE Trans. Med. Imaging 39 (7), 2531–2540.

Zhang, Y., Yang, J., Hou, F., Liu, Y., Wang, Y., Tian, J., Zhong, C., Zhang, Y., He, Z., 2021. Semi-supervised cardiac image segmentation via label propagation and style transfer. In: International Workshop on Statistical Atlases and Computational Models of the Heart. pp. 219–227.

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.