Reducing segmentation failures in cardiac MRI via late feature fusion and GAN-based augmentation

Computers in Biology and Medicine 161 (2023) 106973

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Reducing segmentation failures in cardiac MRI via late feature fusion and GAN-based augmentation

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 segmentation
Image synthesis
GANs
Late fusion
Outlier failure reduction

ABSTRACT

Cardiac magnetic resonance (CMR) image segmentation is an integral step in the analysis of cardiac function and diagnosis of heart related diseases. While recent deep learning-based approaches in automatic segmentation have shown great promise to alleviate the need for manual segmentation, most of these are not applicable to realistic clinical scenarios. This is largely due to training on mainly homogeneous datasets, without variation in acquisition, which typically occurs in multi-vendor and multi-site settings, as well as pathological data. Such approaches frequently exhibit a degradation in prediction performance, particularly on outlier cases commonly associated with difficult pathologies, artifacts and extensive changes in tissue shape and appearance. In this work, we present a model aimed at segmenting all three cardiac structures in a multi-center, multi-disease and multi-view scenario. We propose a pipeline, addressing different challenges with segmentation of such heterogeneous data, consisting of heart region detection, augmentation through image synthesis and a late-fusion segmentation approach. Extensive experiments and analysis demonstrate the ability of the proposed approach to tackle the presence of outlier cases during both training and testing, allowing for better adaptation to unseen and difficult examples. Overall, we show that the effective reduction of segmentation failures on outlier cases has a positive impact on not only the average segmentation performance, but also on the estimation of clinical parameters, leading to a better consistency in derived metrics.

1. Introduction

Accurate segmentation of cardiovascular magnetic resonance (CMR) images is an essential step for heart structure and function assessment, as well as a reliable diagnosis of major cardiovascular diseases [1]. In current-day clinical practice this procedure is typically performed manually or semi-automatically, requiring significant input and correction from clinicians. However, recent developments in automating this task have achieved a remarkable performance. These include approaches ranging from more classical techniques based on statistical shape models or cardiac atlases to newer deep learning (DL) based models, which have gradually outperformed previous state-of-the-art methods [2]. However, most DL methods proposed in the literature have been trained and evaluated using images acquired from single clinical centers, utilizing similar imaging protocols and hardware. Consequently, such models exhibit a significant drop in performance when evaluated on unseen, out-of-distribution data such as abnormal and pathological cases not included in the training set, often characterized by a considerable amount of outliers [3–5]. While typically defined as erroneous or low quality sample, we refer to outliers as data samples exhibiting rare conditions, under-represented in the training data, which often occur at the deployment time. A rare occurrence of such classes during training negatively affects model's ability to adapt to their appearance at test time, leading to a significant degradation in prediction performance and generalization ability. Although anatomical shape constraints can improve segmentation performance in cardiac anatomy, they may not work well with malformations caused by pathologies. Collecting a large and diverse labeled training set for cardiovascular disease patients to improve model performance is not scalable due to acquisition requirements and patient privacy concerns. As a result, research has shifted to methods that optimize model performance on a target dataset without additional labeling. These methods include domain adaptation and generalization algorithms, which aim to extract

* Corresponding author.
E-mail address: y.al.khalil@tue.nl (Y. Al Khalil).
[1] Contributed equally.

https://doi.org/10.1016/j.compbiomed.2023.106973
Received 16 December 2022; Received in revised form 5 April 2023; Accepted 22 April 2023
Available online 26 April 2023

domain-independent features or data augmentation techniques to extend the data distribution [2,6,7]. Generative models are of particular interest in this field, as they can synthesize missing data and features, thus expanding the training set.

1.1. Challenges of CMR segmentation

A change in acquisition parameters causes CMR images to exhibit a great variability in terms of contrast, texture and noise. On the other hand, variation in patient characteristics causes significant divergence in tissue shapes and sizes. Such diversity of imaging characteristics is further intensified by changes in scanner models or vendors. The appearance of pathology has a significant influence on the ventricle morphological variation, resulting in unique tissue shapes and contrast, often under-represented in datasets available for training. Ventricular remodeling further causes changes in heart mass, geometry, function and respiratory wall motion. Segmentation is most commonly hindered by the appearance of dilated left and right ventricles, causing increased wall thickness and regional wall abnormalities. Diseases such as tetrology of fallot and defects in inter-atrial communication induce challenges such as the overriding aorta, right ventricular outflow tract obstruction and pulmonary stenosis. Moreover, segmentation difficulties are caused by gray level inhomogeneities in the blood flow, as well as the presence of papillary muscles and trabeculations, which exhibit the same intensity levels as the myocardium. Finally, segmentation complexity is largely affected by the slice level of the image, where apical and basal slices are more difficult to segment compared to mid-ventricular slices. Due to low MRI resolution, sizes of small structures at the apex and base are often incorrectly estimated due to the vicinity of the atria. Moreover, while the short-axis image orientation is typically used to develop segmentation algorithms due to its efficiency for analyzing both ventricles, it is not fully optimized for the right ventricle [8–10].

1.2. Related work

Recent attempts to handle issues with model robustness and a large number of outliers propose training with images acquired from multiple large cohorts; however, these do not explicitly evaluate the trained models on completely unseen cohorts from other centers, nor directly address the domain shift between cohorts [11–14]. Other research focuses on image and latent space augmentation techniques on models trained and evaluated mostly using single cohorts, with limited evaluation on unseen cohorts [15–18]. However, such approaches are limited by different standards in annotation operating procedures, experiments conducted on private data, as well as the need for a training set sufficiently large to model immense variability across subjects. As a result, these models may perform poorly in clinical settings when faced with a more heterogeneous subject population, including many outlier cases.

Besides data augmentation, other techniques incorporating modifications in model architectures have been proposed to improve the robustness of DL models [12]. Solutions such as transfer learning [19] have been successful, but are limited by the requirement to perform fine-tuning for each specific domain. Domain adaptation approaches [20] have shown promising results for image analysis applications [21,22], but their effects on out-of-domain data are inconclusive [23]. The M&Ms challenge provides a benchmark for testing CMR segmentation algorithms on data from different centers and scanner vendors [24]. Several approaches presented in the challenge demonstrated improvements in domain adaptation, adversarial training, disentangled representation learning, and augmentation to improve model generalization [24]. However, few studies have evaluated the performance of such methods in the presence of diseased tissue and outlier cases that hinder overall performance.

In some cardiovascular disease assessment and segmentation applications, combining information from multiple modalities has been effective. Using balanced-steady state free precession (bSSFP), late gadolinium enhancement (LGE), and T2-weighted (T2w) contrasts for myocardial pathology segmentation through multi-modal training can reduce information uncertainty and improve clinical diagnosis, given the visual variation of myocardial pathology [25,26]. Feature extraction from different tissue representations can help account for these variations and reduce segmentation failures. However, this approach requires multiple patient scans and large amounts of data for training. One solution is using generative adversarial networks (GANs) designed for image synthesis and style translation. Recent developments in the area of GANs have paved the way towards a number of interesting medical imaging applications, from style transfer to utilizing GAN-like architectures for classification or segmentation [27–30]. However, methods focused on medical image synthesis have captured the most attention due to their ability to generate realistic-looking medical images, thus having a potential to increase and vary the available training data [31–35]. While a lot of research so far has focused on improving the quality of image synthesis, a small amount of work evaluates their applicability across different medical image analysis tasks. Moreover, the application of GAN-synthesized images to address algorithm robustness in the presence of out-of-domain images, as well as on data undergoing variations in size, shape and contrast induced by the presence of pathology, has rarely been explored. Utilizing GANs, in conjunction with a multi-modal training approach that permits diverse visual representations of identical tissue images, shows potential to enhance the precision of cardiac tissue segmentation when dealing with constrained, heterogeneous, and imbalanced datasets lacking comprehensive coverage of all potential cases.

1.3. Our contributions

Our work is motivated by the observed heterogeneity in multi-vendor, multi-center and multi-disease cardiac MRI data, shown to severely impact the accuracy of segmentation models. We identify the main aspects that cause a domain shift between images of different cardiac pathologies from different sources. We then simulate these properties by applying a series of steps proposed in this work, with the aim to improve the robustness of segmentation models to the observed variations. To address variability in the FOV and heart size, we introduce a heart region detection module that constrains visible background tissue and centralizes the heart in the image. Further, we use conditional GANs to augment the training with a large number of highly realistic and diverse synthetic images with corresponding labels, particularly focusing on generating enough pathological examples to balance the ratio between pathological and normal cases. Finally, we improve model regularization and robustness by combining the late fusion segmentation approach with intensity transformations that emphasize tissue shape and provide variation in the visual representation of each imaged tissue. We utilize this through a multi-modal training approach, enabling us to extract complementary information from images that have undergone various transformations, shown effective in reducing the prediction uncertainty and minimizing instances of large segmentation failures.

Compared to our previous work in [36], we extend the proposed method and experiments to (i) analyze the proposed pipeline's effect on model robustness using publicly available M&Ms-2 challenge data[2] [24] and show improved performance on outlier cases; (ii) demonstrate the importance of outlier reduction on segmentation performance and clinically-relevant parameters; (iii) use a conditional image synthesis module to generate diverse images with corresponding labels to address

[2] More information about the M&Ms-2 challenge and the data provided can be accessed at https://www.ub.edu/mnms-2/.

data availability limitations; (iv) utilize a variational auto-encoder to increase pathological example diversity in the training set and address mis-segmentation of diseased tissue; and (v) demonstrate the pipeline's ability to adapt to out-of-domain datasets.

2. Materials and methods

2.1. Method overview

The overview of the proposed pipeline is shown in Fig. 2, consisting of (1) heart region detection module, (2) label-conditional image synthesis with a variational autoencoder (VAE) for label deformation and (3) late fusion-based segmentation module utilizing transformed versions of input images during training. We apply the proposed method to both short-axis and long-axis cardiac MR images. In the following sections, we motivate our design choices and introduce each component of the pipeline, as well as the data used for training and validation.

2.2. Background

Synthesis: Previous works in [37,38] have shown the effectiveness of using SPADE-based generators in translating input segmentation labels to realistic CMR images. These are based on a mask-guided image generation technique that employs spatially adaptive denormalization (SPADE) layers, reinforcing semantically-consistent image synthesis [39]. The network is trained using paired images and corresponding labels, where SPADE layers have the ability to inject information from the segmentation map throughout the network and thus guide the generator to correctly learn the translation between the particular class and its appearance. Provided by labels at the input during inference, the trained generator performs label-to-image translation. However, recent work [40,41] suggests the importance of producing comprehensive labels of all visible tissues in the image for generating realistic MR images. Additionally, deforming these labels can facilitate the creation of novel and previously unseen images, wherein automated models like VAEs can introduce a wide range of plausible and diverse deformations. Concurrently to the presented work, Fernandez V. et al. [42] proposed brainSPADE framework which includes a label generator based on a VAE model coupled with a latent diffusion model [43] and a SPADE-based generator for generating labeled brain data. Moreover, a constrained VAE has been proposed in [44] to learn the latent representation of valid cardiac shapes that can be used as post-processing to correct invalid cardiac shapes.

Late Feature Fusion: Existing methods for processing multi-modal images with CNNs typically use early-fusion, combining modalities from low-level features at the network input [45–47]. However, the detection of inter-relations between low-level features from different modalities is difficult due to the non-linear nature of these relationships and the distinct statistical properties of each modality [48,49].

Recent methods propose deep architectures for multi-modal data that fuse higher-level information from different modalities (late fusion), as high-level representations are assumed to be more complementary [50–52]. This approach can be integrated with state-of-the-art architectures, like the U-Net, using separate encoder layers for each modality to disentangle information that would otherwise be fused early on, allowing the network to capture distinctive inter-modality relationships.

Hyper-dense connections, proposed in [50,53,54], can improve the modeling of relationships between different streams by enabling the learning of complex, discriminative features. They facilitate information and gradient propagation through the entire network, reduce the risk of overfitting, and enhance generalization.

As seen in Fig. 5, previous layer outputs across different streams are concatenated at each subsequent layer per stream. The regularization effect can be increased during training by shuffling densely connected layer feature maps, concatenating them in a different order for each

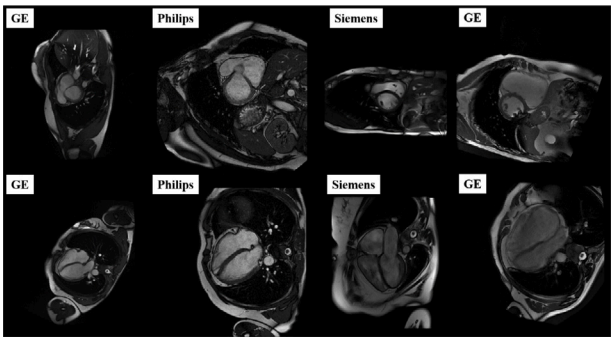

Fig. 1. Variations in field-of-view, image contrast, anatomy, and pathology for SA and LA images in the training set.

branch and layer [55]. Let x_l and F_l denote the output and mapping function (e.g., convolution layer or block with non-linear activation) of the lth layer, respectively. Typically, the output of the lth layer in CNNs is derived by passing the output of the previous layer, x_{l-1}, through a mapping function:

$$x_l = F_l(x_{l-1}). \tag{1}$$

In a densely connected network, this can be extended to

$$x_l = F_l([x_{l-1}, x_{l-2}, x_{l-3}, \dots, x_0]), \tag{2}$$

indicating that all previous feature outputs are concatenated in a feed-forward fashion. By introducing inter-stream hyper-dense connections and feature shuffling, the output of the lth layer in a given stream s, x_l^s, where we consider two streams only, can be defined as

$$x_l^s = F_l^s\left(\phi_l^s\left(x_{l-1}^1, x_{l-1}^2, x_{l-2}^1, x_{l-2}^2, \dots, x_0^1, x_0^2\right)\right), \tag{3}$$

where ϕ_l^s represents a feature map permutation function, responsible for concatenating feature maps in a different order for each branch and layer. Thus, the output of the lth layers in a two-stream network can be represented as

$$\begin{aligned} x_l^1 &= F_l^1([x_{l-1}^1, x_{l-1}^2, x_{l-2}^1, x_{l-2}^2, \dots, x_0^1, x_0^2]) \\ x_l^2 &= F_l^2([x_{l-1}^2, x_{l-1}^1, x_{l-2}^2, x_{l-2}^1, \dots, x_0^2, x_0^1]). \end{aligned} \tag{4}$$

2.3. Data

The M&Ms-2 challenge data is comprised of 360 patients with a variety of right-ventricle (RV) and left-ventricle (LV) pathologies, as well as a control group, distributed as shown in Table 1. The data is acquired using different 1.5T and 3.0T scanners from three different manufacturer vendors (Siemens, GE, and Philips), with variations in contrast and anatomy (see Fig. 1). The in-plane resolution of the provided images varies between 0.78 to 1.57 mm, with slice thickness ranging from 8.6 to 14 mm, resulting in a total number of slices varying between 9 to 13 slices per short-axis image.

The training subset includes 160 cases with expert annotations for RV and LV blood pool, as well as the LV myocardium (MYO). The short-axis and long-axis view is provided for each patient. The training set contains five different types of LV and RV pathologies, as well as healthy subjects. The validation set contains 40 cases with 10 cases of pathologies not present in the training set. The final algorithm is evaluated on a separate test set containing 160 cases as outlined in Table 1. We use the provided validation set for testing, increasing the size of the testing set to a total of 200 patients (or 400 ED and ES SA/LA images) while the evaluation and the development of the algorithm is exclusively done on the training set alone. All images were annotated by two annotators according to the same standard operating procedure (SOP) used for the ACDC MICCAI 2017 challenge [11], while maintaining consistency between short and 4 chambers long-axis in basal and apical regions.

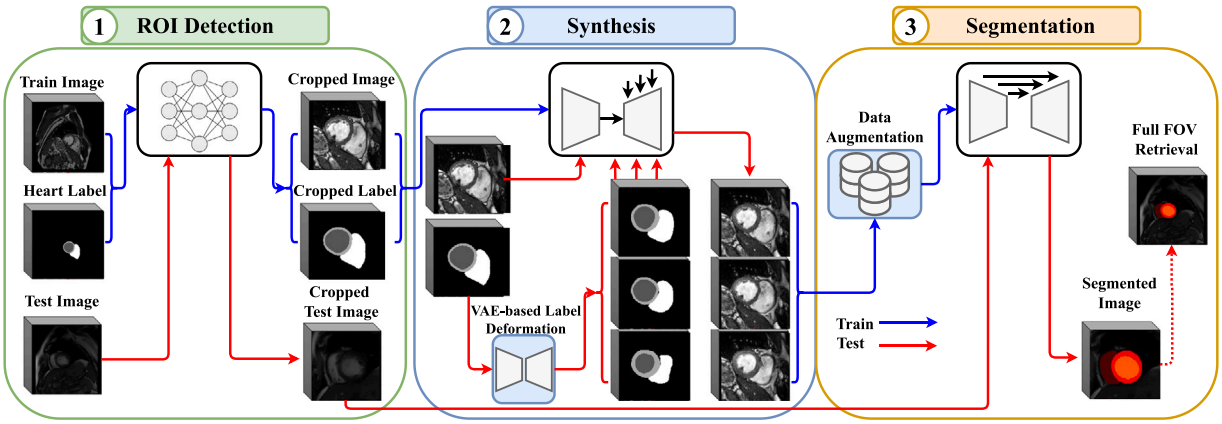

Fig. 2. Proposed pipeline including the ROI detection module (left), image synthesis module (middle) with VAE-based label deformation, and image segmentation module (right).

Table 1
Distribution of the M&Ms-2 challenge data per pathology. Note that all values represent the total number of studies, taken at both ED and ES phases.

Pathology	Training	Validation	Testing
Dilated Right Ventricle (DRV)	0	5	25
Tricuspidal Regurgitation (TRI)	0	5	25
Tetralogy of Fallot (FALL)	20	5	10
Interatrial Communication (CIA)	20	5	10
Congenital Arrhythmogenesis (ARR)	20	5	10
Dilated Left Ventricle (DLV)	30	5	25
Hypetrophic Cardiomiopathy (HCM)	30	5	25
Normal (NOR)	40	5	30

2.4. Heart region detection module

Cardiac MR images acquired at different sites and from varying scanner vendors, typically undergo changes in the acquisition protocol, resulting in images of varying resolution and FOV. This leads to varying heart sizes across different scans, where the heart often takes up only a small portion of the image compared to the background. Experiments show that neural networks trained on images of varying FOV, without ensuring that there is an equal representation of different heart sizes in the training set, often confuse background tissue for cardiac tissue and lead to a large number of false positive predictions.

To address this, we train a regression-based convolutional neural network (CNN), proposed in [36], to automatically detect a bounding box encompassing the heart in both SA and LA images. The detected bounding box is then used for cropping the full FOV images at inference time. The CNN is trained in a supervised manner, with labels obtained from ground truth masks available in the training set by computing the smallest bounding box that fits the entire heart in the FOV and expanding it by 25 voxels to include some background tissue. Before generating the training labels, we resample all SA images to a median spatial resolution of $1.25 \times 1.25 \times 10$ mm^3 and all LA images to a spatial resolution of 1.25×1.25 mm^2.

The inputs to the network are 1000 2D (256×256) mid-cavity SA slices extracted from the training dataset and all LA slices, normalized to intensity values in the range of [0,1]. The cropped SA and LA images using the predicted bounding box are post-processed to the size of 128×128 voxels and 176×176 voxels, respectively. A detailed description of the architecture and training procedure is available in Appendix A.

2.5. Synthesis module

As shown in Fig. 3, the synthesis module encompasses two models; (a) label deformation via latent space manipulation in VAEs and (b) image synthesis via label-conditional GANs. The conditional GANs translate the ground truth labels to realistic images while the VAEs produce new labels with anatomically plausible deformations.

2.5.1. Conditional image synthesis

The image synthesis model is comprised of a ResNet-based [56] style encoder coupled with a label-conditional generator that uses spatially adaptive normalization layers (SPADE) [39] throughout the network architecture. The ResNet encoder is designed to extract style information of the input image and provide it to the generator that preserves the content of the input label map via the conditional SPADE normalization layers. The input image is first fed to the ResNet encoder including a set of convolutional blocks for downsampling and residual blocks followed by two fully connected layers to extract the style information in the bottleneck. This information is then passed to the generator that consists of six SPADE residual blocks, each including SPADE normalization layers that utilize corresponding segmentation mask of the input image for modulating the activation [39].

In contrast to [37,38], our approach alleviates the need for providing multi-tissue segmentation masks for high-quality synthesis by adding the ResNet style encoder network. Moreover, to introduce anatomical variations, random elastic deformation and morphological dilation are applied on the segmentation masks in a previous work [36]. Despite the benefits of label deformation through morphological operations, the heart anatomy of the synthesized subjects is not necessarily anatomically plausible. We tackle this by a VAE-based label deformation approach, described below.

2.5.2. VAE-based label deformation

Instead of elastically deforming the labels as in [36], we propose a deep-learning based label deformation using Variational Autoencoders (VAEs) aiming to learn the underlying factors of heart geometries from the ground truth and in turn provide us with more plausible heart deformations via label encoding and latent space manipulation. The VAE model encodes the shape information of the heart in a compressed manner in the latent space during training. We add random perturbations to the latent code of the original label and then perform label reconstruction by feeding the manipulated latent code to the decoder network. More precisely, to ensure that the latent information is not destroyed, a noise vector is generated from a truncated normal distribution characterized by the statistics of the latent vector (mean, standard deviation, minimum, and maximum). This noise vector is added to the latent vector, resulting in a slight perturbation of the latent vector without compromising its information content. This manipulation of

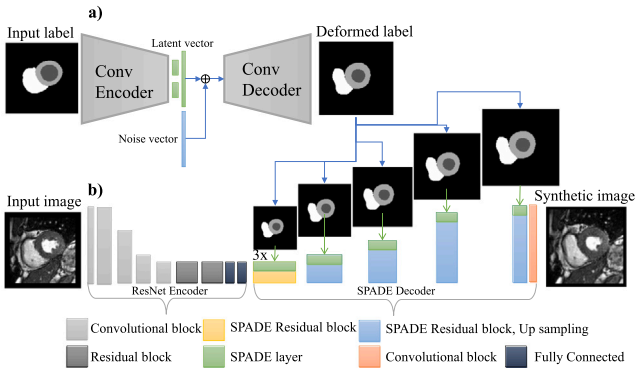

Fig. 3. Synthesis module consisting of (a) a VAE to learn the deformations of the heart shapes to generate a deformed label given the input label and (b) a label conditional GAN model to translate the deformed label to a synthetic image given the input style image.

the latent code changes the heart geometry of the reconstructed label. The rationale behind this approach is that we attempt to directly manipulate the learned geometrical features of the input label in the latent space rather than randomly deform the labels in the image space.

The input of the VAE model is a one-hot encoding version of the label map including four channels for cardiac classes and background. The encoder part of the model includes four convolutional blocks with three convolutional layers each followed by batch normalization (BN) and LeakyReLU activation function. The encoded features are fed to four sequential fully connected layers to output the parameters of a Gaussian prior over the latent representation. The decoder part of the model is comprised of four convolutional blocks each with one up-sampling layer followed by two convolutional layers with BN and LeakyReLU. The last additional block of the decoder includes one convolutional layer followed by BN and another convolution with four channel outputs and Softmax activation function. The VAE model is trained using a weighted combination of cross-entropy loss as the reconstruction loss and Kullback–Leibler divergence (KLD) with a weighting factor of β for regularization of the latent space capacity [57]. We experimentally identify the size of the latent vector ($n_z = 16$) and weight of KLD ($\beta = 15$) by inspecting the quality of the label reconstruction and the outcome of latent code manipulation.

2.5.3. Synthesis strategy

Two identical image synthesis models are trained using LA and SA cardiac MR images. To augment and balance the data using these trained synthesis models, the following strategies are devised. For each vendor-specific subset, the outlier cases are identified based on the end-diastolic or end-systolic volume for the RV calculated using the ground truth label of the SA images. These outlier cases, separated from the rest of the population, are used for image synthesis. For balancing the ratio between outlier cases and the rest of the population, we add different perturbations in the form of Gaussian noise to the corresponding latent space of each label to manipulate labels. This is done in a way that we eventually create roughly 2000 synthesized cases including 50% outliers and 50% the rest of cases. We follow the same strategy for the data from each scanner vendor.

The same strategy is not optimal for LA images as we observe anatomical distortions when noise is added to the latent space of the LA slice. We hypothesize this might be due to having a limited number of LA slices compared to SA stacks and consequently not learning a rich latent space for coherent sampling and accurate reconstruction. Instead, we interpolate between the latent codes from end-diastolic and end-systolic phases of the same subject and feed the interpolated latent codes to the decoder to reconstruct the intermediate shapes. We additionally apply elastic deformation to create more anatomical variations for the LA images.

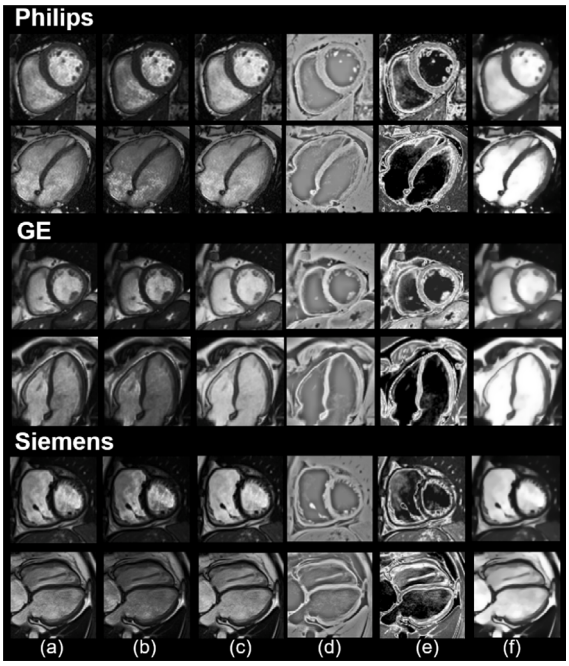

Fig. 4. Examples of contrast-transformed SA and LA training images per vendor (Philips, GE and Siemens). Transformations include: (a) histogram standardization to GE images, (b) histogram standardization to Philips images, (c) histogram standardization to Siemens images, (d) a Laplacian operator, (e) a combination of solarization and posterization and (f) TV-based filtering.

2.6. Segmentation module

2.6.1. Contrast transformations to enhance heart shape features

Segmenting heterogeneous data is challenging due to intensity variations caused by diverse acquisition protocols, signal weighting techniques, and hardware. Applying image appearance transformations during training can introduce contrast diversity, prevent overfitting, and prioritize the model's optimization towards the target tissue's geometry.

We select a set of six contrast transformations per image, each fed into a separate encoding path during training of the late fusion model. First, we match the intensities of images to those representative of each scanner vendor by utilizing histogram standardization [58]. To that end, we generate a standardized set of image histogram landmarks per vendor, used as a reference for matching the histograms of each image at both training and testing time. Next, we apply Total Variation (TV) based denoising [59] to discard high frequency image components and emphasize tissue shape. The scale of the TV filter is controlled by changing the smoothing parameter α, where $\alpha \in [0.1, 15]$. To additionally emphasize tissue edges and flatten the image, while retaining the general appearance, we apply a combination of solarization and posterization. Finally, we calculate the Laplacian of the image to highlight regions of rapid intensity changes and outline major object shapes. The effect of each transformation can be observed in Fig. 4, resulting in a sequence of six augmented images. Note that both real and synthetic images undergo the same procedure during training.

2.6.2. Network architecture

Inspired by late fusion approaches (Section 2.2), we modify the nnU-net [60] architecture to include multiple encoder layers processing each transformed image (Section 2.6.1) fed at the input in a separate path, as shown in Fig. 5. The extracted features by each encoder are then fused at the bottleneck, allowing the network to learn complementary information between different transformations of each image and a better representation of their inter-relationships.

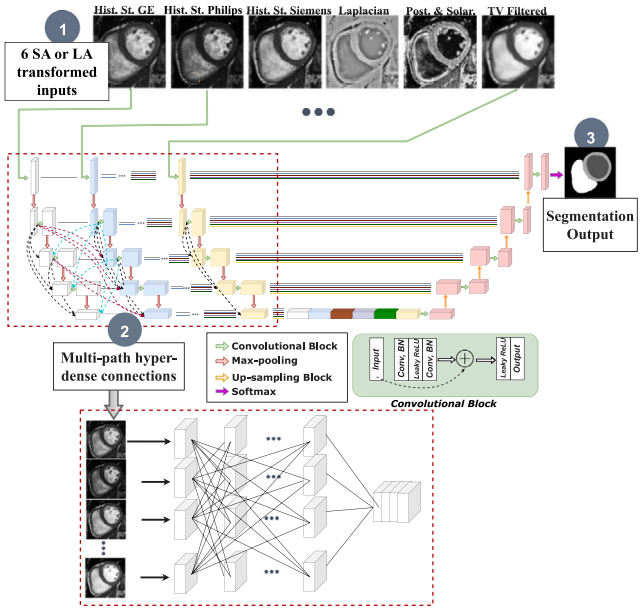

Fig. 5. Overview of the late fusion multi-encoder U-Net proposed in this work. (1) At both training and testing time, the network processed 6 transformed variations of the input SA or LA image through separate encoder paths with (2) hyper-dense dense connectivity used between the layers, within and across paths (dotted lines). Merged features are passed through the decoding path to produce the final LA and SA segmentation maps.

Furthermore, we extend the standard convolutional layers into a convolutional block, consisting of two convolutional and linear units, with batch normalization (BN) applied between each convolution and leaky rectified linear unit (ReLU). We add a short residual connection to sum the input with the output from the second convolutional layer, followed by leaky ReLU to generate the output. Each encoding path consists of five convolutional blocks, with four max-pooling layers. Finally, to improve the modeling of relationships between different streams and promote the learning of highly complex, but more discriminative features, we adopt hyper-dense connections between multiple streams and feature map shuffling. Thus, as discussed in Section 2.2 and shown in Fig. 5, each transformed representation of the input image is processed in a separate path, while dense connections occur not only between the pairs of layers within the same path, but also between those across different paths. As a result, the proposed network has complete flexibility to learn more intricate combinations between transformed images across all levels of abstraction, both within and between them. Note that separate networks are trained for SA and LA segmentation, respectively.

2.6.3. Training procedure

All SA images used for training are first resampled to a median pixel spacing of $1.25 \times 1.25 \times 10$ mm^3. Similarly, a resolution of 1.25×1.25 mm^2 is used for resampling LA images. This is followed by a 98th percentile normalization to an intensity range from $[0, 1]$. All training images are further cropped to reduce the FOV to the region of interest (heart), as described in Section 2.4, while the heart region detection module is used at inference time. This results in all SA and LA images cropped to the size of 128×128 voxels and 176×176 voxels, respectively. Finally, all images are processed to form a set of six different contrast-transformed images (see Section 2.6.1), at both training and testing time. If used for augmentation, synthetic images are pre-processed in the same manner as described above.

After pre-processing, each encoding path is fed with batches of 60 128×128 images for training the SA segmentation model and batches of 20 176×176 images for the LA model. When training with both real and synthetic data, we ensure that each training batch has a minimum of 50% real images to prevent overfitting on the synthetic data. To further increase robustness during training, we employ data augmentation in the form of random vertical and horizontal flips ($p = 0.5$), random rotation by integer multiples of $\frac{\pi}{2}$ ($p = 0.5$), random scaling with a scale factor of $s \in [0.8, 1.2]$ ($p = 0.2$), mirroring ($p = 0.3$) and random elastic deformations ($p = 0.3$). All augmentations are applied on the fly during training.

To train the network, we use a weighted sum of the categorical cross-entropy and Dice loss with Adam optimizer, starting at a learning rate of 1×10^{-4} and a weight decay of $5 \times e^{-5}$. The training of all models converges in 500 epochs, where the initial learning rate is reduced by a factor of 5 if the validation loss does not improve by at least 5×10^{-3} for the last 50 epochs. We apply early stopping on the validation set and select the model with the highest accuracy, to avoid overfitting. We train each model (LA and SA) using a five-fold cross-validation on the training set and use them as an ensemble to produce final predictions on the validation and test sets. The implementation was done in Pytorch using an Nvidia TITAN XP GPU with 12 GBs of RAM.

2.6.4. Post-processing

We perform a connected component analysis on the predicted labels and remove all but the largest connected component per class, which handles most false positive predictions. Since test images are both resampled and cropped, we first restore the original size using the cropping parameters predicted by a heart region detection module and perform bilinear upsampling to recover the original resolution.

3. Experiments

Experiment setup: We train the proposed pipeline on all images provided as a part of the M&Ms-2 training data, consisting of 70, 64 and 26 studies acquired from Siemens, Philips and GE scanners, respectively, and augment the training set with synthetic images generated using the method described in Section 2.5. All studies consist of LA and SA images, at both end-diastolic and end-systolic phases, whereby we train two separate networks per image view (LA vs. SA images). The segmentation performance of the proposed pipeline is compared to the **baseline** model, which is a single-channel nnU-Net [60] combined with heart region detection module. The model is trained on all available real training images from the dataset in a 5-fold cross-validation setup, using the standard augmentation set-up as proposed in [60], without any additional synthetic images.

Overall analysis: The obtained results are evaluated on the unseen test set, containing images acquired across all 3 scanners that have not been previously utilized for the training of any pipeline component. We assess the performance both qualitatively and quantitatively, in terms of standard metrics, such as the Dice score and Hausdorff distance. This is further supported by deriving clinical indicators, such as ventricular volumes and ejection fraction to further assess the benefits of the proposed approach. Detailed discussion is provided in Section 4.2.

Analysis per pathology: Since the major focus of this work is accurate segmentation of cardiac tissue in patients affected by geometrical and textual complexities appearing due to cardiac pathology, we evaluate the proposed pipeline across different diseases available in the test set. In total, we report the results on 160 patient studies, grouped per disease, as well as the normal subjects (seen in Table 1), for both SA and LA images, available in Section 4.3.

We perform additional evaluation on out-of-domain data, namely the short-axis ACDC and M&Ms-1 challenge data (Appendix B) to study the robustness of the proposed method. Detailed results are discussed in Section 4.4. Finally, to study the impact of different pipeline components on segmentation performance, we conduct an ablation experiment by removing one or several elements of the proposed method. All models are evaluated across the whole test set in terms of the Dice score for both SA and LA images, with results available in Section 4.5.

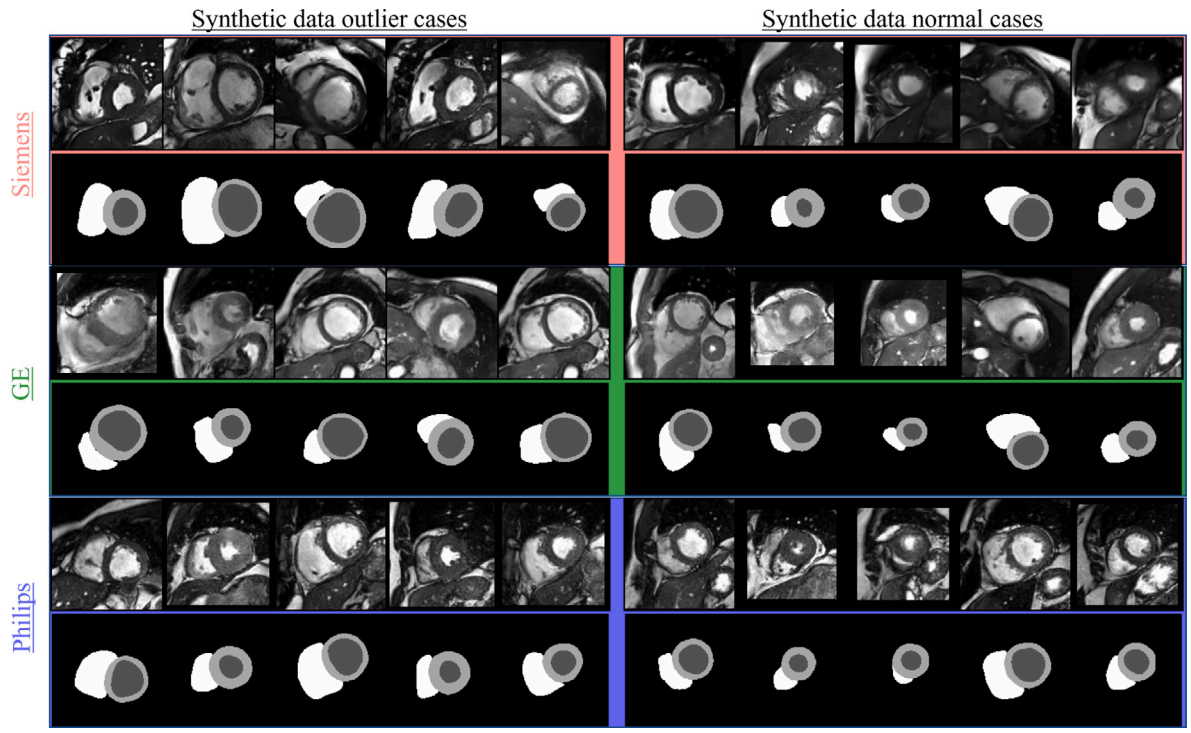

Fig. 6. Random synthetic examples for outlier and normal cases with corresponding labels, stratified into different scanner vendors for short-axis slices. More synthesis examples can be found at https://github.com/sinaamirrajab/CMRSynthVAE.

Table 2
Segmentation performance comparison between the baseline and the proposed model in this work, evaluated on short-axis (SA) and long-axis (LA) test images, across all cardiac tissues. Numbers listed in the table are the means and standard deviations of Dice (DSC) and Hausdorff Distance (HD) scores. DSC and HD values indicated in bold are those which are significantly higher compared to the baseline performance, according to the Wilcoxon signed-rank test for $p < 0.01$.

View	Method	Dice			HD		
		LV	MYO	RV	LV	MYO	RV
SA	Baseline	0.941 (0.05)	0.881 (0.06)	0.923 (0.07)	9.36 (9.22)	13.93 (12.76)	11.71 (10.84)
	Proposed	**0.959 (0.02)**	**0.907 (0.04)**	**0.938 (0.03)**	**6.42 (4.38)**	**9.37 (5.88)**	**8.62 (6.07)**
LA	Baseline	0.947 (0.07)	0.871 (0.08)	0.902 (0.08)	5.04 (4.87)	7.42 (7.21)	7.78 (7.18)
	Proposed	0.958 (0.03)	**0.901 (0.03)**	**0.924 (0.04)**	4.07 (2.09)	**5.27 (3.31)**	**5.81 (3.42)**

4. Results

4.1. Image synthesis results

As discussed in Section 2.5.3, we balance out the number of outlier and normal cases in the final synthetic data by applying a different number of deformations (by adding Gaussian noise to the latent space) on each group. Randomly generated examples from each group are shown in Fig. 6.

Fig. 7 shows the RV and LV volumes at ED and ES phases for the real and synthetic data distribution to inspect how the synthetic population changes the heart cavity distribution of subjects. Each subject is represented as a point with different shape and color for its corresponding scanner vendor. We can observe a gap between subjects in the real data distribution (indicated with black oval) that is filled with generated subjects in the synthetic data distribution as a result of deforming labels and synthesizing more subjects for each real subject. Moreover, the number of samples near the mean of the distribution are increased, resulting in a densely populated area covered by synthetic subjects with normal ranges of ventricular volumes.

4.2. Segmentation results

Table 2 shows the quantitative results in terms of Dice and HD scores obtained by the proposed pipeline compared to the baseline model, across SA and LA images available in the test set. The obtained results suggest significant improvements in segmentation performance across most tissues, except for the LV in LA images, which we ascribe to relative consistency of LV shape over the long-axis view. However, visual observation suggests improvements in patients with dilated left ventricle (LVD) and hypertrophic cardiomyopathy (HCM), which both cause changes in LV shape and appearance. Additional observation of score distribution, depicted in Fig. 8, suggests a significant reduction in the number of outlier predictions by the proposed approach across both SA and LA views. This has a particular impact on the segmentation of the right-ventricular (RV) blood-pool and myocardium (MYO), whereby visual observation of delineations implies that the existing outliers predicted using the baseline mostly relate to false positive predictions, specifically in relation to over-estimation of both the LV and RV blood-pool, which further causes the under-segmentation of the myocardium.

Further inspection suggests that over-segmentation mainly occurs in the basal region of the heart, whereby the baseline model falsely predicts the presence of the RV and other tissues, particularly at the boundary of the pulmonary artery and the right atrium. Under-segmentation by the baseline commonly occurs at the apex of the heart, where endocardium appears smaller and tissue boundaries are less well-defined. The presence of dense papillary muscles at the apex often causes further difficulties for accurate segmentation. The observations obtained by visual inspection are confirmed by quantitative evaluation performed across heart regions, shown in Fig. 9. Compared to mid-ventricular

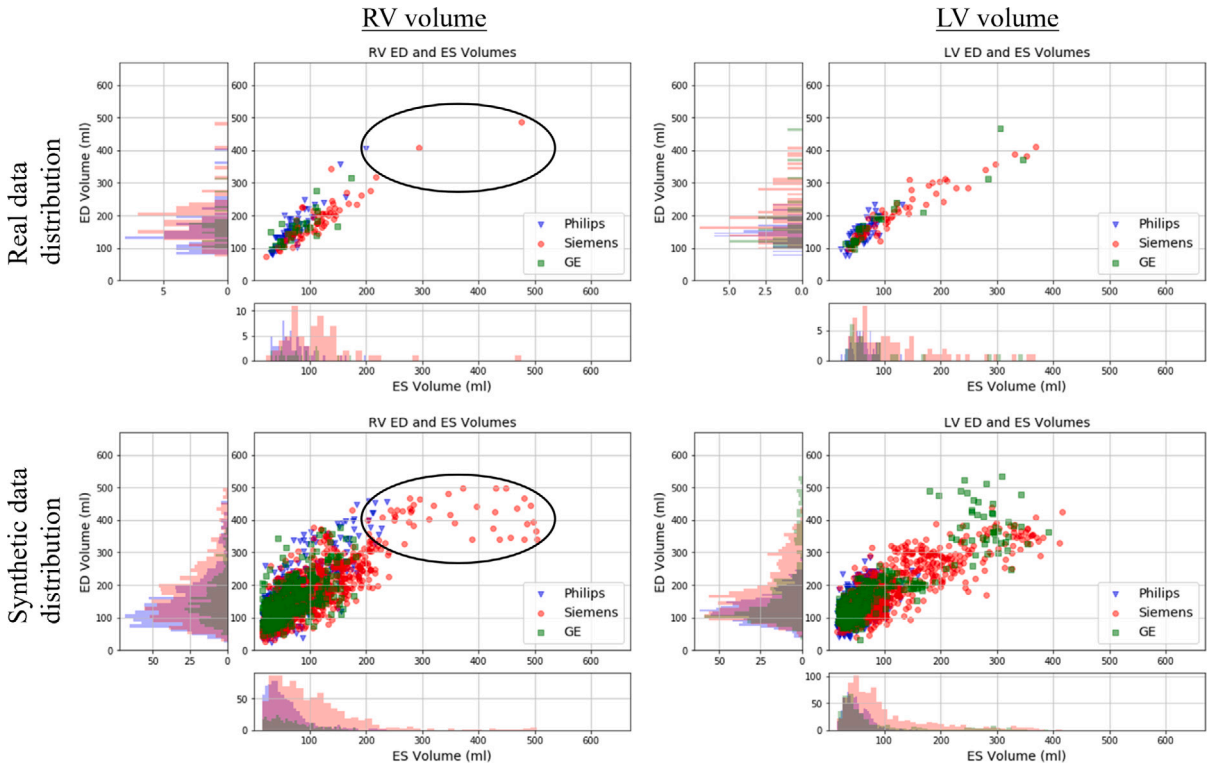

Fig. 7. Distribution of the RV and LV volumes at ED and ES for real and synthetic data. Each subject is represented as a marker (with different colors and shapes indicating the corresponding scanner vendor).

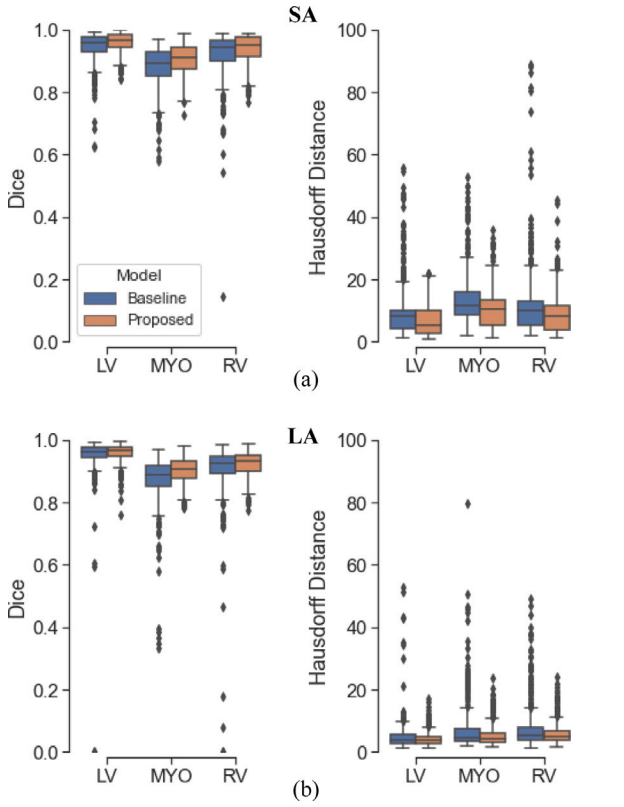

Fig. 8. Segmentation performance of the proposed and baseline models on (a) SA and (LA) images available in the test set, across three cardiac tissues (LV, MYO and RV) in terms of Dice and Hausdorff distance (HD) scores.

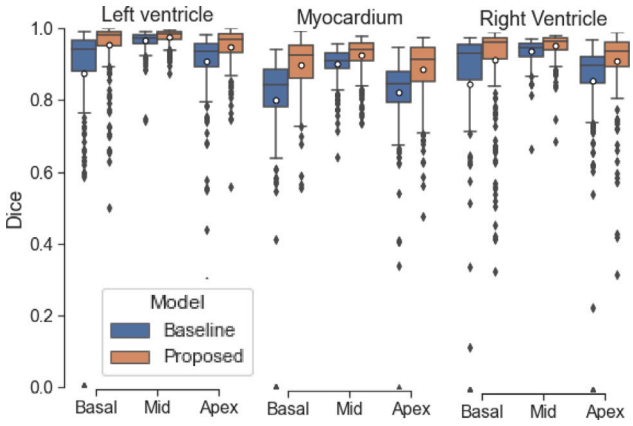

Fig. 9. Segmentation performance of proposed and baseline models in SA images across basal, mid-ventricular and apical heart regions, calculated per cardiac tissue.

slices there is an evident performance drop around both the base and the apex of the heart, which is noticeably improved by the proposed approach. We hypothesize the improvement largely stems from augmentation with synthetic data, where we focus on including a vast array of examples with variable appearance of tissues in the basal and apical regions of the heart, as well as simulating the effects of heart pathology on cardiac tissue appearance and the presence of artifacts. However, a moderate performance drop when segmenting both ends of the heart compared to the mid-ventricular region, obtained by the proposed approach, suggests that under- and over-segmentation at the base and the apex of the heart is still not a completely resolved problem.

To gain more insight into the value and importance of outlier reduction achieved by the proposed segmentation method, we evaluate three automatically derived clinical parameters with reference to manually derived ones using the available ground truth segmentation masks.

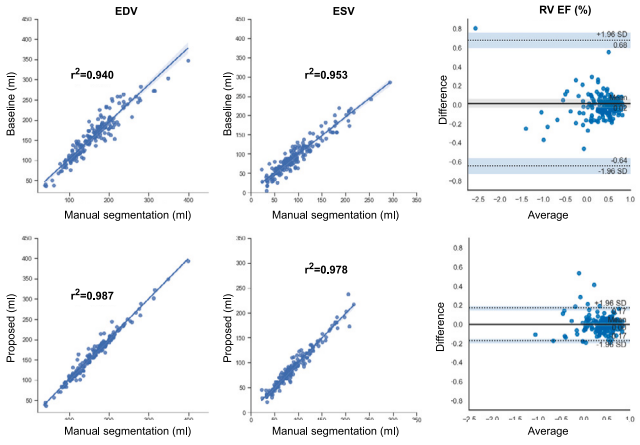

Fig. 10. Correlation and Bland–Altman plots of RV functional parameters generated from manual and automatically predicted segmentation masks using the baseline (first row) and proposed (second row) models.

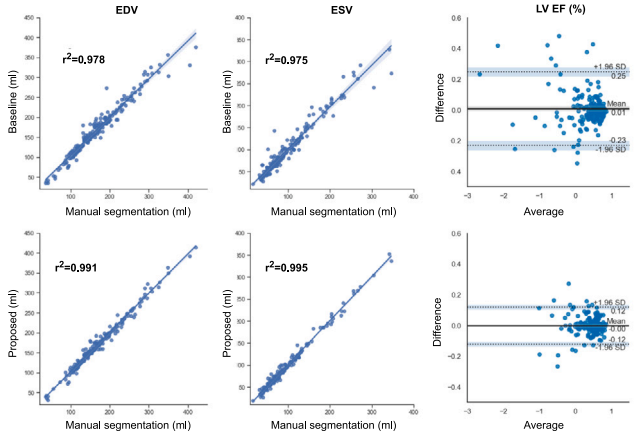

Fig. 11. Correlation and Bland–Altman (rightmost) plots of LV functional parameters generated from manual and automatically predicted segmentation masks using the baseline (first row) and proposed (second row) models.

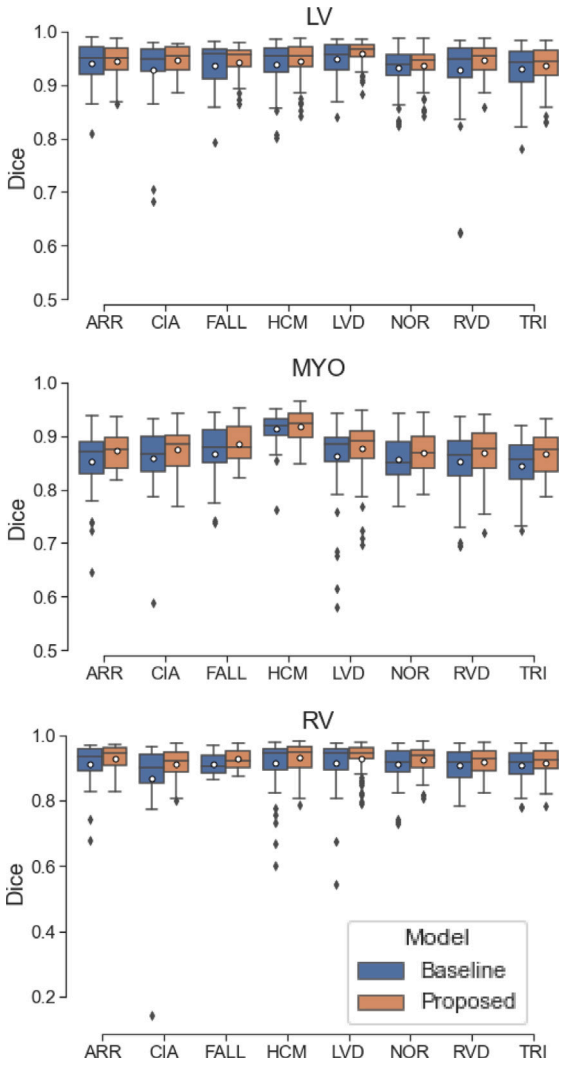

Fig. 12. Average segmentation performance of the baseline and proposed models across SA test images, derived per disease, in terms of Dice score per tissue.

Namely, we derive the RV (Fig. 10) and LV (Fig. 11) end-diastolic (EDV) and end-systolic (ESV) volumes, as well as the ejection fraction (EF) from segmentations obtained by both the proposed and baseline model across the entire test set.

Correlation plots of baseline and proposed ED and ES RV volumes in Fig. 10 show significant improvements in both EDV and ESV correlation acquired from the proposed method, which leads to better agreement between manual and automatically quantified EF compared to the baseline. Similar effects are observed in Fig. 11 for all three clinical parameters related to the LV. These results can largely be attributed to a smaller number of outlier predictions, which in turn decrease the difference between the calculated ED and ES volumes compared to those derived from ground-truth labels. Moreover, the remaining outliers are still relatively close to the acceptable range of deviation, which reduces their overall impact on calculated ED and ES volumes, as well as the ejection fraction.

4.3. Analysis per pathology

To gain additional insight into the performance of the segmentation methods analyzed in this study, we stratify the quantitative analysis per pathology, as shown in Fig. 12. Fig. 12 depicts Dice scores achieved by the baseline and proposed methods, extracted per tissue across SA images. Overall, we note consistent improvements in segmentation performance across all tissue, with more prominent gains in the case of myocardium (MYO) and right-ventricular (RV) blood-pool segmentation. In fact, statistically significant improvements in MYO segmentation, according to the paired Wilcoxon signed-rank test ($p < 0.01$), are obtained for SA cases undergoing defects related to arrhythmogenic cardiomyopathy (ARR), tetrology of fallot (FALL), dilated left ventricle (LVD), dilated right ventricle (RVD), tricuspidal regurgitation (TRI), as well as on healthy patients (NOR). Likewise, statistically significant increase in Dice scores for RV segmentation are observed for patients suffering from inter-atrial communication (CIA), tetrology of fallot (FALL), dilated left ventricle (LVD) and dilated right ventricle (RVD). However, segmentation over LV shows only slight improvements, mostly related to outlier reduction, with statistically significant differences observed for patients with defects in tetrology of fallot (FALL), dilated left ventricle (LVD), as well as dilated right ventricle (RVD).

The scores obtained on RVD and TRI cases are of a particular interest, as these are completely unseen during training, suggesting that the proposed method has the ability to compensate for unseen diseases. The improvement in segmentation of patients undergoing RVD and TRI further leads to enhanced derivation of clinical parameters, as seen in Fig. C.2 and Fig. C.3 for both the LV and RV, respectively (Appendix C). Similar trends are observed in LA images (see Fig. C.1) across RV

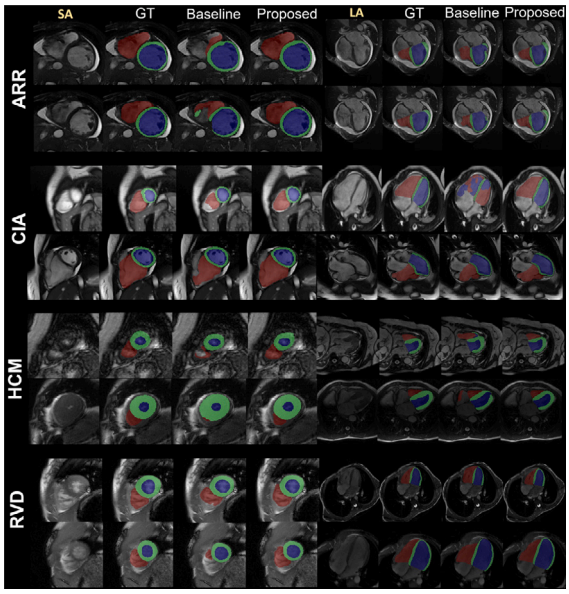

Fig. 13. Qualitative visualization of segmentation results in challenging cases undergoing different cardiac disease, outlining the improvement in segmentation when using the proposed pipeline compared to the baseline. Each row presents a single patient, where we showcase one SA slice, as well as the LA view of the heart corresponding to the same patient. Model predictions are compared to the ground truth, shown in the column marked as GT.

And MYO segmentation, with more moderate improvement across the LV.

Fig. 13 shows some challenging segmentation cases across different pathologies present in the test set for both SA images and their LA counterparts. Patients undergoing arrhythmogenic cardiomyopathy often exhibit right ventricular dilatation and scarring in the myocardial area. This is commonly reflected by difficulties in segmenting both the RV and MYO, as shown in Fig. 13, which can be tackled by utilizing the proposed pipeline. Similar results are observed for patients suffering from interatrial communication defects (CIA), where RV dilatation is typical. Hypertrophic cardiomyopaty (HCM) is often found in the middle septum at the midventricular level, as well as in the inferior region at the apical region [61], as shown in Fig. 13. In these cases, the baseline model often under-segments the RV, but also struggles with over-segmenting the myocardium. Finally, a dilated right ventricle presents another typical case of under-segmentation, especially in slices towards the base and apex of the heart. However, the proposed pipeline shows noticeable improvement in handling such examples.

4.4. Evaluation on external datasets

To demonstrate the robustness of the proposed pipeline, we perform an additional evaluation on a completely different set of out-of-domain CMR images. These include data acquired from ACDC [11] and M&Ms-1 [24] challenges, which we use to directly test both the baseline and the proposed models and report the results in terms of Dice and Hausdorff distance scores. We do this without additionally re-training or adapting the models to new data.

A description of both datasets is available in Appendix B. Since only the training data from the ACDC Challenge dataset is available publicly, consisting of ED and ES images from 100 subjects, we utilize this as our test set. However, the evaluation on the M&Ms-1 data is done on the actual test data provided by the challenge organizers, consisting of 80 ED and ES subjects (a total of 160 images) from four different vendors. It is important to note that the evaluation of this experiment is only performed for SA images, since both challenges do not contain LA data.

While there is an evident domain shift between the ACDC and M&Ms-1 data compared to the M&Ms-2 data used for the training of the complete pipeline, we still observe a significant improvement when utilizing the proposed multi-modal approach and augmenting the training set with synthetic images, as seen in Table 3. A large portion of the ACDC dataset contains pathological cases, where we observe significant improvements in performance, particularly when segmenting RV and MYO. However, we hypothesize that additional improvements in performance could be achieved if the training was adapted to those datasets specifically, especially the synthesis module, as the current approach is specifically tailored to M&Ms-2 data.

4.5. Ablation study

We perform an ablation study to understand the value of different pipeline components on segmentation performance. Therefore, we train the following models: (i) **U-Net** + **BB**, a regular single encoder nnU-Net with Bounding Box (BB) detection, corresponding to the baseline model; (ii) **U-Net** + **BB**+ **IT**, a model similar to (i) but augmented with the same set of intensity transformations (IT) as in the late-fusion model; (iii) **U-Net** + **BB** + **IT** + **Synth**, a model similar to (ii) with added synthetic data (Synth) at training time, generated as described in Section 2.5; (iv) **LF-U-Net** + **BB**, a dense late fusion (LF) approach combined with bounding box detection and (v) **LF-U-Net** + **BB** + **Synth**, a dense late fusion approach proposed in this paper. All models are trained using the procedure in Section 2.6.3. Additional insight into the effects of multi-stream connections and the type of transformations used on the segmentation performance is available in Appendix D, Table 1.

The obtained results across the entire M&Ms-2 test data, for both SA and LA images, are outlined in Table 4. We start observing significant improvements in performance with the addition of synthetic images, generally related to patients with dilated right and left ventricles, hypertrophic cardiomyopathy and arrhythmogenic cardiomyopathy. However, we do not observe any improvement in segmentation among healthy patients, which we hypothesize is due to the fact we focus the augmentation process on diseased patients and abnormal heart shapes. On the other hand, introducing a late-fusion approach, combined with hyper-dense connections, demonstrates some performance improvement in those cases. In LA images, the addition of synthetic images has a significant effect on right ventricle segmentation, where visual observation suggests improvements on patients with severe changes in RV shape due to underlying pathologies.

The late fusion model used in this study leads to more refined segmentations of the LV and MYO in SA and LA images, respectively with consistent improvements in images with visible artifacts, as well as in cases with low contrast between tissues. Adding synthetic data to the late-fusion model (**LF-U-Net** + **BB** + **Synth**) yields further improvements, mostly around the RV area, as well as the myocardium. Augmentation with synthetic data tends to reduce the amount of variation between the predictions, leading to better reliability and stability of segmentations. This is particularly manifested when evaluating the trained models across patients with pathologies unseen during training, such as the tricuspidal regurgitation (TRI) and the dilated right ventricle. Similar results can be observed across LA images, where the proposed model tackles both under- and over-segmentation across all tissues, noticeable in single-encoder models. This is most obvious when studying the delineations across the LV and MYO. Largest improvements in performance are obtained across the patients suffering from dilated RV and TRI — the unseen cases during training, suggesting that both synthetic data and better modeling of relationships between differently transformed images aid with tackling the changes in both heart shape and appearance due to the presence of pathological tissue.

Table 3
Segmentation performance comparison between the baseline and the proposed model in this work, evaluated on short-axis (SA) images acquired from the M&Ms-1[24] and ACDC [11] challenges (Appendix B). The evaluation is performed over all three cardiac tissues, in terms of Dice and Hausdorff Distance (HD) scores. Values indicated in bold are those which are significantly higher compared to the baseline performance, according to the Wilcoxon signed-rank test for $p < 0.01$.

	Model	Dice			HD		
		LV	MYO	RV	LV	MYO	RV
M&Ms-1 (n = 160)	Baseline	0.908 (0.05)	0.799 (0.05)	0.873 (0.07)	12.04 (8.6)	16.04 (9.1)	13.77 (8.1)
	Proposed	**0.925 (0.03)**	**0.821 (0.04)**	**0.901 (0.04)**	**7.81 (3.8)**	**11.34 (5.6)**	11.84 (6.2)
ACDC (n = 200)	Baseline	0.955 (0.03)	0.868 (0.03)	0.922 (0.05)	7.69 (5.2)	9.51 (7.5)	16.49 (15.9)
	Proposed	0.962 (0.01)	**0.891 (0.02)**	**0.934 (0.03)**	**5.62 (3.3)**	**7.61 (5.2)**	**11.45 (4.9)**

Table 4
Segmentation performance comparison between the baseline and the proposed model in this work, as well as the models trained with different elements of the proposed pipeline, according to the ablation experiment described in Section 4.5. Each model is evaluated on original short-axis (SA) and long-axis (LA) test images, across all cardiac tissues. Numbers listed in the table are the means and standard deviations of Dice score. Dice values indicated in bold are those which are significantly higher compared to the baseline performance, according to the Wilcoxon signed-rank test for $p < 0.01$.

Method	Short-Axis (SA)			Long-Axis (LA)		
	LV	MYO	RV	LV	MYO	RV
U-Net + BB (Baseline)	0.941 (0.05)	0.881 (0.06)	0.923 (0.07)	0.947 (0.07)	0.871 (0.08)	0.902 (0.08)
U-Net + BB + IT	0.945 (0.04)	0.886 (0.05)	0.925 (0.05)	0.949 (0.07)	0.874 (0.06)	0.907 (0.07)
U-Net + BB + IT + Syn	0.950 (0.04)	**0.891 (0.06)**	**0.931 (0.04)**	0.951 (0.06)	0.879 (0.05)	**0.911 (0.08)**
LF U-Net + BB	**0.953 (0.03)**	**0.898 (0.05)**	**0.934 (0.05)**	0.954 (0.05)	**0.885 (0.04)**	**0.919 (0.06)**
LF U-Net + BB + Syn (Proposed)	**0.959 (0.02)**	**0.907 (0.04)**	**0.938 (0.03)**	**0.958 (0.03)**	**0.901 (0.03)**	**0.924 (0.04)**

5. Discussion

5.1. Result analysis

In this work, we propose a pipeline designed to tackle the segmentation of pathological CMR images across multiple views (SA and LA) and sources. We provide a comprehensive analysis of the proposed pipeline and compare its performance to the baseline model, widely used in the literature (nnU-Net). The obtained results demonstrate the ability of the proposed pipeline to reduce the performance gap between the outlier cases, riddled by artifacts and shape deformation caused by underlying pathologies, and cases similar to those available at training time.

While outlier cases are typical and commonly found across many medical imaging tasks, they are more prominent in pathological data, owing to limitations in representation and number of cases. This particularly affects data-hungry deep learning algorithms, known to fail on cases poorly represented during training. However, the method proposed in this work can tackle such cases more effectively, leading to a notable decrease in the number of outliers during segmentation. This in turn has a significant impact on not only the average segmentation performance, but also the derivation of clinically relevant metrics characterizing heart function. In fact, we demonstrate significant improvements in levels of agreement and bias reduction for the left- and right-ventricular ejection fraction across all cases available at inference time. Outlier reduction has shown to be consistent throughout all pathologies and cardiac tissues. Further investigation suggests that outliers occurring in this dataset belong to images with large differences in appearance and contrast compared to the majority of images available at training time, as well as those containing higher levels of noise and artifacts. However, cases with severe tissue deformation and occlusion due to the presence of pathologies represent the most challenging cases during segmentation, largely addressed by the proposed approach.

We further show that using conditional GANs holds a lot of promise for the generation of missing data and addressing data scarcity, which is emphasized when dealing with pathological patients. In fact, careful generation of images varying in contrast and tailored to different cardiac diseases can significantly improve model generalization and reduce class imbalance, common in regular datasets skewed towards non-pathological images. We demonstrate the impact of the synthetic images on the distribution of heart cavity samples in the training set, which we hypothesize increases the representation of challenging cases during training and leads to a more stable segmentation performance, even in the presence of unseen diseases. The proposed synthesis approach can be adapted to any type of scarce data, such as rare pathologies, whereby only a small subset of such data is needed for training a synthesis module that can expand the training set with artificial patients of varying appearance.

Although data augmentation with more extreme intensity transformations has recently shown to positively influence the regularization and generalization of DL-based methods, as well as reduce overfitting, we show that combining such transformations using a multi-path approach aids the network with learning complementary information and fosters better data representation, enhancing the networks' discriminative power. Moreover, enhancing the flow of information between the multi-path layers through dense connections shows further benefits in obtaining a more accurate segmentation. Visual observation suggests that this improvement is particularly related to the segmentation of small structures (such as the tissue at the apex of the heart) or at region boundaries, where single encoder networks tend to struggle at differentiating between tissues.

Performance analysis across different pathologies in both SA and LA images reveals additional insights about the behavior of different models evaluated in this study. We observe that baseline models are prone to over-segmentation, particularly in the basal regions, where they falsely predict the presence of either the RV or the whole heart. This is usually caused by blood movement-related artifacts, low tissue contrast or occlusion due to specific diseased tissue. Additional difficulties appear in cases where the myocardial muscle does not completely enclose the blood pool and exhibits variability in shape, becoming non-circular. Furthermore, we note that the proposed method tends to exhibit more significant improvements in terms of the HD scores, which is primarily the result of outlier reduction. Visual observation suggests that cases exhibiting high HD scores when evaluated with a baseline model, contain false positive predictions in the areas outside of the heart, often consisting of tissues similar in appearance and shape to cardiac tissue. Additional errors contributing to segmentation inaccuracies obtained by the baseline include areas with weak or missing edges, artifacts and low signal-to-noise ratio.

In general, we note considerable improvements using the proposed method across most pathological cases, with better adaptation to unseen cases. The obtained results are comparable and even outperform those reported in the M&Ms-2 challenge, ranging from 0.83 to 0.93 and 0.8 to 0.92 in Dice score for RV segmentation across SA and LA images, respectively[3] [62–75]. Moreover, augmenting the training set

[3] Evaluation over LV and MYO is not included in the evaluation procedure of the M&Ms-2 challenge and is not reported by any participants.

with highly diverse data and introducing a more efficient way to extract meaningful features from data leads to improved performance on out-of-domain datasets (Section 4.4). This shows that despite training the model on a completely different set of images, the proposed modules aid in adaptation to the existing domain shift that commonly occurs between different datasets. Finally, we demonstrate the effects on performance when one or more modules are removed from the proposed pipeline and identify the largest sources of improvement in the ablation study (see Section 4.5). We demonstrate that each element of the proposed pipeline adds value to the overall segmentation improvement, but significant differences start appearing when utilizing augmentation with synthetic data, as well as the late fusion approach with dense connections.

5.2. Limitations and future work

Despite the reported improvement in segmentation performance and outlier reduction, the proposed model still has several limitations. The performance drop on ES slices remains higher compared to the ED slices, which further affects the calculation of clinical parameters, such as the ejection fraction. Moreover, basal regions are prone to under-segmentation, followed by a drop in accuracy around the apex of the heart, mainly due to its small size compared to the rest of the cavity. While we manage to partly handle some of these issues, they consistently remain the biggest sources of errors, which is in agreement with findings reported by similar work in the literature on other datasets. This implies that special attention should be placed on addressing these regions, which we plan to focus on in future work. Additionally, the provided LA images could help with extracting the inter- and intra-view information from the complementary SA and LA images, allowing for both the localization of the basal plane and possibly better segmentation of the basal slices. Thus, integration of the proposed modules into a truly multi-view approach would be one of the main aspects to focus on in our future work.

Furthermore, while we extensively analyze the impact of the proposed pipeline on a wide array of pathological data with varying sources of acquisition, we would further benefit from assessing its confidence and identifying possible prediction uncertainties under difficult settings. In the same line, extending this study on other open-source datasets, as well as to clinical settings, would allow us to further identify the necessary points of improvement, particularly when performing the evaluation on other unseen cardiac pathologies. Additionally, this involves exploring whether various elements of the suggested method can enhance the efficacy of other models in the field, as well as evaluation against other generation models. Although we focus this work on handling the variation in cardiac tissue shape and appearance in the presence of various diseases, we note that the segmentation performance on healthy patients does not show significant improvements. Thus, ensuring that the model generalizes well to both healthy and diseased tissue is another major focus of our future experiments.

To balance out the number of pathological and normal cases from the M&Ms-2 challenge, we identify outlier subjects by calculating the right-ventricle volume using the ground truth labels and taking into account the mean and standard deviation values. However, this method for outlier detection may not necessarily cover all pathological cases available in the dataset, as just the volume may not be indicative of a cardiac disease. Instead, having access to labels for each pathology, one can synthesize more subjects for a particular disease in such a way to obtain a balanced number of different diseases present in the data.

6. Conclusions

In this work, we propose a pipeline including three distinct modules to handle different challenges of multi-vendor and multi-disease cardiac MR images for the task of increasing the segmentation robustness and outlier reduction. We demonstrate the ability of the proposed approach to balance the segmentation of outlier cases, typically related to increased levels of artifacts and shape deformations induced by the presence of pathologies, with those more commonly represented in the training set. Synthesizing a diverse training dataset, carefully designed to increase the variation of cardiac shapes and appearances during training, plays a significant role in not only boosting the model performance in terms of standard quantitative metrics, but also in improving the automatically derived clinical metrics denoting the function of the heart. This in turn leads to improved stability and reliability of the predictions across both short-axis and long-axis images. Such observations are additionally confirmed on completely unseen images, extracted from other publicly available datasets, whereby we observe both outlier reduction and better adaptation to the presence of the domain shift between datasets. Future work includes more precise synthesis of pathological cases, conditioned on the pathology type, as well as utilizing the availability of LA images to inform the positioning of the basal plane for more accurate segmentation.

CRediT authorship contribution statement

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

Declaration of competing interest

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. Cristian Lorenz and Jürgen Weese are employees of Philips GmbH Innovative Technologies.

Acknowledgments

This research is a part of the openGTN project, supported by the EU Marie Curie Innovative Training Networks (ITN) fellowship 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.compbiomed.2023.106973.

References

[1] T. Leiner, D. Rueckert, A. Suinesiaputra, B. Baeßler, R. Nezafat, I. Išgum, A.A. Young, Machine learning in cardiovascular magnetic resonance: basic concepts and applications, J. Cardiovasc. Magn. Reson. 21 (1) (2019) 1–14.

[2] C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, D. Rueckert, Deep learning for cardiac image segmentation: A review, Front. Cardiovasc. Med. 7 (2020) 25.

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

[4] V.M. Bashyam, J. Doshi, G. Erus, D. Srinivasan, A. Abdulkadir, A. Singh, M. Habes, Y. Fan, C.L. Masters, P. Maruff, et al., Deep generative medical image harmonization for improving cross-site generalization in deep learning predictors, J. Magn. Reson. Imag. 55 (3) (2022) 908–916.

[5] R. Attar, M. Pereanez, A. Gooya, X. Alba, L. Zhang, M.H. de Vila, A.M. Lee, N. Aung, E. Lukaschuk, M.M. Sanghvi, et al., Quantitative CMR population imaging on 20,000 subjects of the UK Biobank imaging study: LV/RV quantification pipeline and its evaluation, Med. Image Anal. 56 (2019) 26–42.

[6] R. Volpi, H. Namkoong, O. Sener, J.C. Duchi, V. Murino, S. Savarese, Generalizing to unseen domains via adversarial data augmentation, Adv. Neural Inf. Process. Syst. 31 (2018).

[7] Q. Dou, D. Coelho de Castro, K. Kamnitsas, B. Glocker, Domain generalization via model-agnostic learning of semantic features, Adv. Neural Inf. Process. Syst. 32 (2019).

[8] C. Petitjean, M.A. Zuluaga, et al., Right ventricle segmentation from cardiac MRI: A collation study, Med. Image Anal. 19 (1) (2015) 187–202.

[9] S. Marchesseau, J.X. Ho, J.J. Totman, Influence of the short-axis cine acquisition protocol on the cardiac function evaluation: a reproducibility study, Eur. J. Radiol. Open 3 (2016) 60–66.

[10] P. Yilmaz, K. Wallecan, W. Kristanto, J.-P. Aben, A. Moelker, Evaluation of a semi-automatic right ventricle segmentation method on short-axis MR images, J. Digit. Imag. 31 (5) (2018) 670–679.

[11] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M.A.G. Ballester, et al., Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging 37 (11) (2018) 2514–2525.

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

[13] C.F. Baumgartner, L.M. Koch, M. Pollefeys, E. Konukoglu, 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, 2017, pp. 111–119.

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

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

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

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

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

[19] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2009) 1345–1359.

[20] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, J. Mach. Learn. Res. 17 (1) (2016) 2030–2096.

[21] E. Tzeng, J. Hoffman, T. Darrell, K. Saenko, Simultaneous deep transfer across domains and tasks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4068–4076.

[22] N.K. Dinsdale, M. Jenkinson, A.I. Namburete, Unlearning scanner bias for MRI harmonisation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 369–378.

[23] J.C. Acero, V. Sundaresan, N. Dinsdale, V. Grau, M. Jenkinson, 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, 2020, pp. 196–207.

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

[25] W. Li, L. Wang, F. Li, S. Qin, B. Xiao, Myocardial pathology segmentation of multi-modal cardiac MR images with a simple but efficient Siamese U-shaped network, Biomed. Signal Process. Control 71 (2022) 103174.

[26] D. Li, Y. Peng, Y. Guo, J. Sun, TAUNet: a triple-attention-based multi-modality MRI fusion U-Net for cardiac pathology segmentation, Complex Intell. Syst. 8 (3) (2022) 2489–2505.

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

[28] J.J. Jeong, A. Tariq, T. Adejumo, H. Trivedi, J.W. Gichoya, I. Banerjee, Systematic review of generative adversarial networks (gans) for medical image classification and segmentation, J. Digit. Imag. (2022) 1–16.

[29] S. Kazeminia, C. Baur, A. Kuijper, B. van Ginneken, N. Navab, S. Albarqouni, A. Mukhopadhyay, GANs for medical image analysis, Artif. Intell. Med. 109 (2020) 101938.

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

[31] G. Kwon, C. Han, D.-s. Kim, Generation of 3D brain MRI using auto-encoding generative adversarial networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 118–126.

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

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

[34] M. Rezaei, Generative adversarial network for cardiovascular imaging, in: Machine Learning in Cardiovascular Medicine, Elsevier, 2021, pp. 95–121.

[35] A. Ferreira, J. Li, K.L. Pomykala, J. Kleesiek, V. Alves, J. Egger, GAN-based generation of realistic 3D data: A systematic review and taxonomy, 2022, arXiv preprint arXiv:2207.01390.

[36] Y. Al Khalil, S. Amirrajab, J. Pluim, M. Breeuwer, Late fusion U-Net with GAN-based augmentation for generalizable cardiac MRI segmentation, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer, 2021, pp. 360–373.

[37] S. Amirrajab, S. Abbasi-Sureshjani, Y. Al Khalil, C. Lorenz, J. Weese, J. Pluim, M. Breeuwer, 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, 2020, pp. 128–137.

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

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

[40] S. Amirrajab, Y. Al Khalil, C. Lorenz, J. Weese, J. Pluim, M. Breeuwer, Label-informed cardiac magnetic resonance image synthesis through conditional generative adversarial networks, Comput. Med. Imaging Graph. 101 (2022) 102123.

[41] Y. Al Khalil, S. Amirrajab, C. Lorenz, J. Weese, J. Pluim, M. Breeuwer, On the usability of synthetic data for improving the robustness of deep learning-based segmentation of cardiac magnetic resonance images, Med. Image Anal. 84 (2023) 102688.

[42] V. Fernandez, W.H.L. Pinaya, P. Borges, P.-D. Tudosiu, M.S. Graham, T. Vercauteren, M.J. Cardoso, Can segmentation models be trained with fully synthetically generated data? in: Simulation and Synthesis in Medical Imaging: 7th International Workshop, SASHIMI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings, Springer, 2022, pp. 79–90.

[43] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.

[44] N. Painchaud, Y. Skandarani, T. Judge, O. Bernard, A. Lalande, P.-M. Jodoin, Cardiac segmentation with strong anatomical guarantees, IEEE Trans. Med. Imaging 39 (11) (2020) 3703–3713.

[45] K. Kamnitsas, C. Ledig, V.F. Newcombe, J.P. Simpson, A.D. Kane, D.K. Menon, D. Rueckert, B. Glocker, Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation, Med. Image Anal. 36 (2017) 61–78.

[46] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, D. Shen, Deep convolutional neural networks for multi-modality isointense infant brain image segmentation, NeuroImage 108 (2015) 214–224.

[47] X. Zhuang, L. Li, C. Payer, D. Štern, M. Urschler, M.P. Heinrich, J. Oster, C. Wang, Ö. Smedby, C. Bian, et al., Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge, Med. Image Anal. 58 (2019) 101537.

[48] B. Huang, F. Yang, M. Yin, X. Mo, C. Zhong, A review of multimodal medical image fusion techniques, Comput. Math. Methods Med. 2020 (2020).

[49] N. Srivastava, R.R. Salakhutdinov, Multimodal learning with deep Boltzmann machines, Adv. Neural Inf. Process. Syst. 25 (2012).

[50] J. Dolz, C. Desrosiers, I.B. Ayed, IVD-Net: Intervertebral disc localization and segmentation in MRI with a multi-modal UNet, in: Int. Workshop and Challenge on Comp. Methods and Clinical Applications for Spine Imaging, 2018, pp. 130–143.

[51] D. Nie, L. Wang, Y. Gao, D. Shen, Fully convolutional networks for multi-modality isointense infant brain image segmentation, in: 2016 IEEE 13Th International Symposium on Biomedical Imaging, ISBI, IEEE, 2016, pp. 1342–1345.

[52] G. van Tulder, M. de Bruijne, Learning cross-modality representations from multi-modal images, IEEE Trans. Med. Imaging 38 (2) (2018) 638–648.

[53] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[54] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, I.B. Ayed, HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation, IEEE Trans. Med. Imaging 38 (5) (2018) 1116–1126.

[55] T. Zhang, G.-J. Qi, B. Xiao, J. Wang, Interleaved group convolutions, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4373–4382.

[56] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[57] I. Higgins, L. Matthey, A. Pal, C.P. Burgess, X. Glorot, M.M. Botvinick, S. Mohamed, A. Lerchner, beta-VAE: Learning basic visual concepts with a constrained variational framework, in: ICLR, 2017.
[58] L.G. Nyúl, J.K. Udupa, X. Zhang, New variants of a method of MRI scale standardization, IEEE Trans. Med. Imaging 19 (2) (2000) 143–150.
[59] L.I. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms, Physica D 60 (1–4) (1992) 259–268.
[60] F. Isensee, P.F. Jaeger, S.A. Kohl, J. Petersen, K.H. Maier-Hein, nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation, Nature Methods 18 (2) (2021) 203–211.
[61] Y. Amano, M. Kitamura, H. Takano, F. Yanagisawa, M. Tachi, Y. Suzuki, S. Kumita, M. Takayama, Cardiac MR imaging of hypertrophic cardiomyopathy: techniques, findings, and clinical relevance, Magn. Reson. Med. Sci. 17 (2) (2018) 120.
[62] X. Sun, L.-H. Cheng, R.J. Geest, Right ventricle segmentation via registration and multi-input modalities in cardiac magnetic resonance imaging from multi-disease, multi-view and multi-center, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer, 2021, pp. 241–249.
[63] T.W. Arega, F. Legrand, S. Bricq, F. Meriaudeau, Using MRI-specific data augmentation to enhance the segmentation of right ventricle in multi-disease, multi-center and multi-view cardiac MRI, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer, 2021, pp. 250–258.
[64] L. Li, W. Ding, L. Huang, X. Zhuang, Right ventricular segmentation from short- and long-axis MRIs via information transition, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer, 2021, pp. 259–267.
[65] C. Galazis, H. Wu, Z. Li, C. Petri, A.A. Bharath, M. Varela, Tempera: Spatial transformer feature pyramid network for cardiac MRI segmentation, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer, 2021, pp. 268–276.
[66] S. Jabbar, S.T. Bukhari, H. Mohy-ud Din, Multi-view SA-LA Net: A framework for simultaneous segmentation of RV on multi-view cardiac MR Images, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer, 2021, pp. 277–286.
[67] S. Queirós, Right ventricular segmentation in multi-view cardiac MRI using a unified U-Net model, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer, 2021, pp. 287–295.
[68] M.J. Fulton, C.R. Heckman, M.E. Rentschler, Deformable Bayesian convolutional networks for disease-robust cardiac MRI segmentation, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer, 2021, pp. 296–305.
[69] Z. Gao, X. Zhuang, Consistency based co-segmentation for multi-view cardiac MRI using vision transformer, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer, 2021, pp. 306–314.
[70] D. Liu, Z. Yan, Q. Chang, L. Axel, D.N. Metaxas, Refined deep layer aggregation for multi-disease, multi-view & multi-center cardiac MR segmentation, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer, 2021, pp. 315–322.
[71] M. Beetz, J. Corral Acero, V. Grau, A multi-view crossover attention U-Net cascade with Fourier domain adaptation for multi-domain cardiac MRI segmentation, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer, 2021, pp. 323–334.
[72] M. Mazher, A. Qayyum, A. Benzinou, M. Abdel-Nasser, D. Puig, Multi-disease, multi-view and multi-center right ventricular segmentation in cardiac MRI using efficient late-ensemble deep learning approach, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer, 2021, pp. 335–343.
[73] K. Punithakumar, A. Carscadden, M. Noga, Automated segmentation of the right ventricle from magnetic resonance imaging using deep convolutional neural networks, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer, 2021, pp. 344–351.
[74] L. Tautz, L. Walczak, C. Manini, A. Hennemuth, M. Hüllebrand, 3D right ventricle reconstruction from 2D U-Net segmentation of sparse short-axis and 4-chamber cardiac cine MRI views, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer, 2021, pp. 352–359.
[75] F. Galati, M.A. Zuluaga, Using out-of-distribution detection for model refinement in cardiac image segmentation, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer, 2021, pp. 374–382.