Synthesising Images and Labels Between MR Sequence Types with CycleGAN

Eric Kerfoot[1(✉)], Esther Puyol-Antón[1], Bram Ruijsink[1,2], Rina Ariga[3], Ernesto Zacur[3], Pablo Lamata[1], and Julia Schnabel[1]

[1] School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK
eric.kerfoot@kcl.ac.uk
[2] St Thomas' Hospital NHS Foundation Trust, London, UK
[3] University of Oxford, Oxford, UK

Abstract. Real-time (RT) sequences for cardiac magnetic resonance imaging (CMR) have recently been proposed as alternatives to standard cine CMR sequences for subjects unable to hold the breath or suffering from arrhythmia. RT image acquisitions during free breathing produce comparatively poor quality images, a trade-off necessary to achieve the high temporal resolution needed for RT imaging and hence are less suitable in the clinical assessment of cardiac function. We demonstrate the application of a CycleGAN architecture to train autoencoder networks for synthesising cine-like images from RT images and vice versa. Applying this conversion to real-time data produces clearer images with sharper distinctions between myocardial and surrounding tissues, giving clinicians a more precise means of visually inspecting subjects. Furthermore, applying the transformation to segmented cine data to produce pseudo-real-time images allows this label information to be transferred to the real-time image domain. We demonstrate the feasibility of this approach by training a U-net based architecture using these pseudo-real-time images which can effectively segment actual real-time images.

Keywords: Cardiac MR · Cardiac quantification · Convolutional neural networks · Generative adversarial networks · Image synthesis

1 Introduction

Free breathing non-gated real-time cine (RT) is a cardiac magnetic resonance imaging (cMRI) protocol proposed as a solution to restrictions present in the standard short axis cine protocol [12,13]. The latter protocol reconstructs images from multiple cardiac cycles and thus relies on electrocardiogram (ECG) gating, consistent cardiac cycle periods, and breath holds during acquisition. For patients with arrhythmia or who cannot hold their breath, a characteristic often seen in patients with heart diseases, this protocol is not feasible for producing useful CMR images.

Q. Wang et al. (Eds.): DART 2019/MIL3ID 2019, LNCS 11795, pp. 45–53, 2019.
https://doi.org/10.1007/978-3-030-33391-1_6

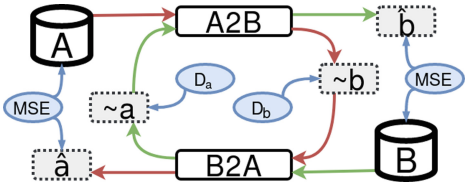

Fig. 1. The CycleGAN network setup. The path for real-time images is in red and that for cine MR images in green. Image data is shown as grey boxes and loss functions in blue. (Color figure online)

Real-time cine by contrast is neither ECG-gated nor breath-held, capturing cardiac motion over multiple cycles which is not directly reconstructed. The drawbacks to this protocol include through-plane motion during respiration impacting quantification of flow, the absence of ECG data informing where in the cardiac cycle each image is placed, and the poor quality of RT images in terms of feature resolution relative to standard acquisitions.

Previous work to address motion issues and identify cardiac cycle position [9,15] can be applied to mitigate these drawbacks, however this does not improve image quality directly. Real-time images suffer from a blurring effect due to fast acquisition, reconstructing a sharper version of these images with clearer delineation of cardiac tissues would potentially aid in visual assessment by clinicians.

In this paper we outline a method for converting between real-time images and short-axis images using a CycleGAN [22] based neural network architecture. We used trained autoencoders to enhance the quality of real-time images by converting them to a pseudo-cine image which present a better delineation between myocardial and surrounding tissues. Since we also can convert cine to pseudo-real-time images, we demonstrate the feasibility of converting segmented cine data to pseudo-real-time to use as training input. Consecutively we trained a U-net [14] based network to segment real-time images and compare its output against clinician labelled data, showing that our approach also allows transfer of training data to improve automation of analysis of this less common used technique lacking rich labelled datasets.

Related Work: Deep learning has recently shown great promise in synthesising medical images across different sequencing types within the same image modality based on conditional generative networks [8] and cycle generative adversarial networks [7,19,20,22]. Huo *et al.* [7] proposed an end-to-end synthetic segmentation network for abdominal images and for intracranial volume synthetic segmentation. A known problem with image synthesis is that of "hallucinating" data [2], where features commonly found in a target domain, but absent in the source image, are added to synthesised images. We address this problem in this work by using images representing roughly the same cardiac geometry and excluding

cases having pathological geometric variation from the datasets, thus no image feature is expected to coincide with one image domain.

Several non-gated RT imaging techniques have been proposed to overcome the limitations with ECG-gated CMR [3,12,13,16]. Despite solving an important problem for patients with severe heart disease (who often have arrhythmia and problems holding breath) feature definition and image quality remains inferior to standard imaging, even using advanced acceleration techniques.

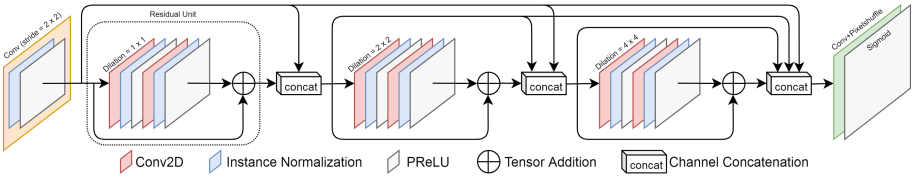

Fig. 2. Autoencoder network topology. Input tensors are first downsamples to half the original dimensions using a strided convolution (yellow). Each subsequent residual block (first one marked with dotted box) is composed of convolution layers with strides of 1 and dilation factors given above. All convolutions are 2D with 3×3 kernels. (Color figure online)

2 Method

To train our image synthesis autoencoders we employed a CycleGAN approach outlined in Fig. 1. Given two image distributions A and B (in our case real-time and cine images respectively), we train one autoencoder $A2B$ to convert an image from distribution A to appear like one from distribution B (labelled $\sim b$), and another autoencoder $B2A$ to perform the opposite translation ($\sim a$). Discriminator networks D_A and D_B are used to discriminate between real and synthetic images from each distribution.

These are then passed through the other network to produce reconstructed images $\hat{a}$ and $\hat{b}$ which are then compared against the original images using mean squared error. The discriminator networks are trained subsequently using the generated batches $\sim a$ and $\sim b$, plus images from A and B, as input.

The input data consists of ~20000 real-time CMRI images [15] acquired on a 1.5T Philips Ingenia MR scanner (Philips, Best, Netherlands) from 10 healthy participants as dataset A, plus ~5000 short-axis cine images acquired from the same participants as dataset B. Having an unbalanced ratio between the size of A and B was found to produce poor results during training, and so ~15000 cine images from the ACDC challenge dataset [1] were added to B. The selected ACDC images were acquired from Siemens 1.5T and 3T scanners (Siemens Medical Solutions, Germany) and include healthy subjects as well as those with myocardial infarction.

Binary cross entropy is used as the loss function to discriminate between real and synthetic images, given that 1 labels real images and 0 synthetic:

$$\mathcal{L}_{D_A} = \mathbb{E}_{a\sim A}[log(D_A(a))] + \mathbb{E}_{b\sim B}[log(1 - D_A(B2A(b)))]$$

$$\mathcal{L}_{D_B} = \mathbb{E}_{b\sim B}[log(D_B(b))] + \mathbb{E}_{a\sim A}[log(1 - D_B(A2B(a)))]$$

Mean squared error is used as the loss between real and synthetic images:

$$\mathcal{L}_A = \mathbb{E}_{a\sim A}[\|B2A(A2B(a)) - a\|_2], \quad \mathcal{L}_B = \mathbb{E}_{b\sim B}[\|A2B(B2A(b)) - b\|_2]$$

The final loss for training the two autoencoders together is the following, using a value of 10 for the hyperparameter λ:

$$\mathcal{L} = \lambda\mathcal{L}_A + \lambda\mathcal{L}_B + \mathcal{L}_{D_A} + \mathcal{L}_{D_B}$$

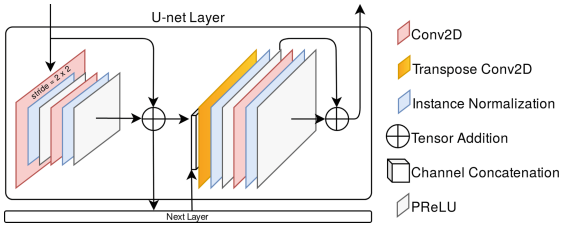

Fig. 3. The U-net segmentation network is built as a stack of these layers. The "Next Layer" is either another such layer or the bottom layer consisting of only a convolution/instance-norm/PReLU sequence. In this way each layer represents both the encoding and decoding pathway where input data flows through the left-hand residual block, through next layer, up through the right-hand residual block which also upsamples and concatenates, and then the layer above.

Figure 2 illustrates the architecture of the two autoencoders. As a memory efficiency measure the input image is downsampled by a factor of 2 using a convolution with a stride of 2 in both dimensions, followed by instance normalisation and PReLU [4] activation. The network then consists of three residual blocks [5] containing two sets of convolution-normalisation-activation layers. The final layer is a convolution with 1×1 kernels to adjust the number of channels to 4, followed by a pixel shuffle layer [17] to output an image with the same dimensions as the original input, to which sigmoid activation is applied.

To increase the perceptive field of the residual blocks, the convolutions of the second and third block are dilated by factors of 2 and 4. Input to each block is the concatenated image volume of the outputs from the original set of convolutions plus the outputs of previous blocks. This implements a dense block similar to [6] which permits data from convolutions with smaller dilations to be passed directly to those with larger dilations.

It was found in this experiment that dilated convolutions are an effective alternative to downsampling as a means of expanding the perceptive area of convolutions. The more common autoencoder architecture involves multiple downsampling layers (the encode step) following by upsampling layers (the decode step). Using dilated kernels it was possible to define an alternative autoencoder model which operated on near-full resolution data but was still feasible to train and produced good quality output. Another factor contributing to good results was the dense block architecture which allowed information from convolutions with smaller perceptive fields to be passed directly to those with larger, thus feeding information about smaller scale features down the encoder pipeline to be integrated with information about larger scale features.

3 Segmentation

Using synthetic data we have trained a segmentation network based on the residual U-net [14,21] architecture to label the left ventricle. Our U-net architecture (Fig. 3) is defined as a stack of downstream/upstream layers, and uses residual blocks throughout the encoding and decoding path. Our network consists of four such blocks with encoding paths producing output volumes with 16, 32, 64, and 128 channels respectively. Strided convolutions and strided transpose convolutions are used to downsample and upsample data, which are followed by instance normalisation and PReLU layers.

Input data was acquired from multiple datasets of cine images defined with three label segmentations (left-ventricle chamber, left ventricle myocardium, right ventricle chamber). These datasets are the ACDC challenge dataset [1] of 100 cases, 175 healthy cases from the UK Biobank segmented by a clinician, 116 cases captured on Siemens Trio 3T scanners (Siemens Medical Solutions, Germany), and 215 cases captured on 1.5T Philips Ingenia scanners. The total number of images used for training is 9095. With the images converted to pseudo-real-time with unaltered segmentations, the network was trained to transfer the amalgamated label information from one form of MR to another.

During training, random batches of image/segmentation pairs are drawn from this training dataset and a random set of operations are applied to each. This follows [10] in its use of data augmentation [11,18] as so is not suitable to be trained with the generator networks as in [7]. These augmentations include simple array transforms (flip, rotate, transpose, etc.) but more importantly also include randomised free-form deformations. The previous work has shown that this combination of network architecture and training process results in robust and accurate segmentation networks.

4 Results

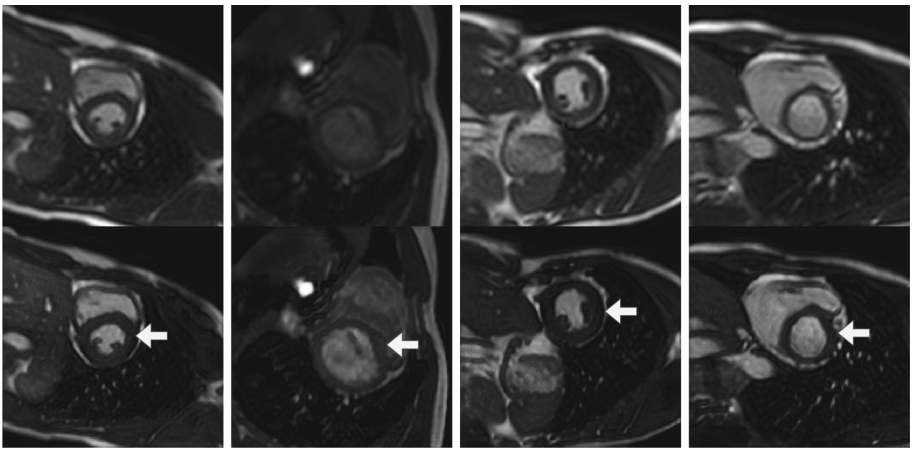

Fig. 4. Examples of image synthesis, arrows pointing to cardiac area of interest. For each image pair the original real-time image is above and the generated pseudo-cine image below.

Image Comparison. Figure 4 gives examples of the translation from real-time images to pseudo-cine images. The generated images exhibit greater contrast than the originals and the myocardium (specifically the left ventricle) is better defined with a more distinct boundary between myocardial tissue and pool or surrounding tissue. The peak signal-to-noise ratio between the generated images the ~20000 original images is -54.22 dB ($\sigma = 1.42$), and the structural similarity index between them is 0.77 ($\sigma = 0.08$).

Temporal information is not used in the transformation process so the relative motion between images is preserved in the pseudo-real-time images. This allows arrhythmia to be observed in the translated images as the cardiac cycle time is not affected. Typically arrhythmia excludes the use of cine MR thus our approach is an effective means of analysing such patients with high quality imagery.

Wall motion abnormalities are difficult to visualise in cine MR images as the wall position will vary from cycle to cycle, when multiple cycles are combined the resulting image is less distinct in this area. These abnormalities are thus more clearly identified with the sharper imaging produced by our method which does not reconstruct images based on the assumption of uniform geometry across cycles.

Segmentation Comparison. To assess the segmentation network, we use our set of manually-segmented real-time images as a ground truth comparison with generated segmentations. The manual segmentations were drawn by clinicians on images where they were confident the left ventricle was sufficiently distinct

to define a reasonable boundary for the segmentation. This data is composed of binary segmentations of the LV myocardium only, thus our comparison study will use only these labels from the generated label set.

The segmentation network was applied to the 3194 expert-segmented real-time images in our dataset. Of these we were able to predict 2747 correct annular segmentations after extracting the largest element from each segmentation image. The mean dice score between these segmentations and the clinician-defined ground truth is 0.783 with a standard deviation of 0.083. Figure 5 illustrates examples of generated segmentations as compared to their ground truths.

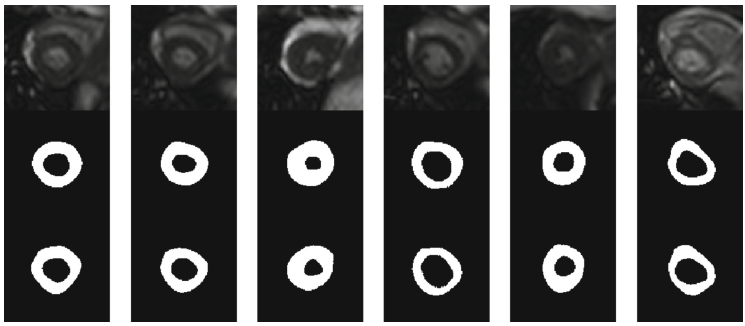

Fig. 5. Segmentation examples with original real-time image (top of each column), ground truth segmentation (middle), and predicted segmenation (bottom).

5 Conclusion

We have in this work defined a methodology for overcoming some of the deficiencies with real-time MR sequences by using a CycleGAN deep learning architecture to improve image quality. By training autoencoders to produce cine-like images from real-time images, we can produce a processed version of a real-time image sequence where the myocardial tissue is better differentiated from its surroundings and the ventricular cavities. This can serve as an aid to clinicians when assessing cardiac function by providing a sharper and more distinct image with improved contrast levels.

Using the second autoencoder to convert cine MR images to pseudo-real-time images allows a segmentation network to be trained using existing labelled cine data but which can be applied to real-time images. We have demonstrated the feasibility of this approach to reusing data between MR image types, which is especially important for real-time images as very little labelled data exists for the left ventricle and none for the right ventricle. In the future we intend to use this combined data to train networks capable of segmenting both ventricles despite current real-time datasets only having left ventricle labels.

Acknowledgements. This research was supported by the National Institute for Health Research (NIHR) Biomedical Research Centre (BRC) at Guy's and St Thomas' NHS Foundation Trust, and by the Wellcome EPSRC Centre for Medical Engineering at the School of Biomedical Engineering and Imaging Sciences, King's College London (WT 203148/Z/16/Z). This research has been conducted using the UK Biobank Resource under Application Number 17806.

References

1. Bernard, O., Lalande, A., Zotti, C., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging **37**(11), 2514–2525 (2018)
2. Cohen, J.P., Luck, M., Honari, S.: Distribution matching losses can hallucinate features in medical image translation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 529–536. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_60
3. Feng, L., Srichai, M.B., Lim, R.P., et al.: Highly accelerated real-time cardiac cine MRI using k-t sparse-sense. Magn. Reson. Med. **70**(1), 64–74 (2013)
4. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
5. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38
6. Huang, G., Liu, Z., Weinberger, K.Q.: Densely connected convolutional networks. CoRR abs/1608.06993 (2016)
7. Huo, Y., Xu, Z., Moon, H., et al.: Synseg-net: synthetic segmentation without target modality ground truth. IEEE Trans. Med. Imaging **38**(4), 1016–1025 (2018)
8. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
9. Kerfoot, E., Puyol Anton, E., Ruijsink, B., Clough, J., King, A.P., Schnabel, J.A.: Automated CNN-based reconstruction of short-axis cardiac MR sequence from real-time image data. In: Stoyanov, D., et al. (eds.) RAMBO/BIA/TIA -2018. LNCS, vol. 11040, pp. 32–41. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00946-5_4
10. Kerfoot, E., Clough, J., Oksuz, I., Lee, J., King, A.P., Schnabel, J.A.: Left-ventricle quantification using residual U-net. In: Pop, M., et al. (eds.) STACOM 2018. LNCS, vol. 11395, pp. 371–380. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-12029-0_40
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
12. La Gerche, A., Claessen, G., Van de Bruaene, A., et al.: Cardiac MRI: a new gold standard for ventricular volume quantification during high-intensity exercise. Circ. Cardiovasc. imaging **6**(2), 329–38 (2013)
13. Lurz, P., Muthurangu, V., Schievano, S., et al.: Feasibility and reproducibility of biventricular volumetric assessment of cardiac function during exercise using real-time radial k-t SENSE magnetic resonance imaging. J. Magn. Reson. Imaging **29**(5), 1062–1070 (2009)

14. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
15. Ruijsink, B., et al.: Semi-automatic cardiac and respiratory gated MRI for cardiac assessment during exercise. In: Cardoso, M.J., et al. (eds.) CMMI/SWITCH/RAMBO -2017. LNCS, vol. 10555, pp. 86–95. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67564-0_9
16. Setser, R.M., Fischer, S.E., Lorenz, C.H.: Quantification of left ventricular function with magnetic resonance images acquired in real time. J. Magn. Reson. Imaging **12**(3), 430–438 (2000)
17. Shi, W., et al.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1874–1883 (2016)
18. Simard, P.Y., Steinkraus, D., Platt, J.C., et al.: Best practices for convolutional neural networks applied to visual document analysis. In: ICDAR, vol. 3 (2003)
19. Welander, P., Karlsson, S., Eklund, A.: Generative adversarial networks for image-to-image translation on multi-contrast MR images-a comparison of cyclegan and unit. arXiv preprint arXiv:1806.07777 (2018)
20. Wolterink, J.M., Dinkla, A.M., Savenije, M.H.F., Seevinck, P.R., van den Berg, C.A.T., Išgum, I.: Deep MR to CT synthesis using unpaired data. In: Tsaftaris, S.A., Gooya, A., Frangi, A.F., Prince, J.L. (eds.) SASHIMI 2017. LNCS, vol. 10557, pp. 14–23. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68127-6_2
21. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. IEEE Geosci. Remote Sens. Lett. **15**(5), 749–753 (2018)
22. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV) (2017)