Super Resolution of Cardiac Cine MRI Sequences Using Deep Learning

Nicolas Basty(✉) and Vicente Grau

Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK
nicolas.basty@eng.ox.ac.uk

Abstract. Cardiac cine MRI facilitates structural and functional analysis of the heart through the dynamic aspect of the sequences. Clinical acquisitions consist of sparse 2D images instead of 3D volumes, taken at landmark points of the ECG to cover the whole heartbeat. A stack of short axis images and a small number of long axis views are generally acquired. Efforts have been made to accelerate acquisitions at the acquisition stage as well as at post-processing. A major part of current research in medical image processing focuses on deep learning approaches driven by large datasets. However, most of those methods leave out the dynamic aspect of temporal data and treat frames of cine MRI sequences individually. We propose a super resolution network based on the U-net and long short-term memory layers to exploit the temporal aspect of the dynamic cardiac cine MRI data. When given a sequence of low resolution long axis images, our method is able to render a high resolution sequence. Results on synthetic data simulating a stack of short axis images show quantitative and qualitative improvements over traditional interpolation methods or the equivalent machine learning method using a single frame, including the ability of the network to recover important image features such as the apex.

Keywords: Super-resolution · Cardiac cine MRI · Deep learning

1 Introduction

Cardiac cine MRI allows functional and structural analysis of the heart. Due to its exceptional soft tissue contrast, reproducibility and safety considerations it is commonly taken as the gold standard for cardiac imaging. To capture the whole heartbeat in a sequence of images, scans are produced at landmark times synchronised with ECG readings. To minimise the imaging times, clinical acquisitions consist of anisotropic 2D slices instead of a 3D volume. A stack of parallel short axis (SA) slices and a small number of orthogonal long axis (LA) slices are generally acquired for each frame of the sequence. The number of slices in the SA stack is dependent on the size of the heart but generally ranges between 8 and 12, and the number of LA slices is also variable. Standardised protocols

D. Stoyanov et al. (Eds.): RAMBO 2018/BIA 2018/TIA 2018, LNCS 11040, pp. 23–31, 2018.
https://doi.org/10.1007/978-3-030-00946-5_3

such as the UK biobank protocol include three LA views: the vertical long axis (VLA, also called 2-chamber view), the horizontal long axis (HLA, also called 4-chamber view), and the left ventricular outflow tract (LVOT) view [1,2]. Some of the main issues associated with cine MRI are the slice misalignment occurring due to patient motion and breath hold variations between acquisitions, intensity differences between slices due to flow artefacts and magnetic field inhomogeneities, and sometimes contrast agents, as well the sparsity of the data occasionally resulting in a lack of coverage of the left ventricle by the SA stack [3]. The dynamic aspect of cardiac MRI is used to evaluate cardiovascular function metrics such as the ejection fraction and the stroke volume, to quantify wall motion and thickness and identify scar tissue in follow-up scans from patients who have suffered a myocardial infarct.

The MRI pulse sequence most commonly used in cardiac cine MRI for left ventricular structural and functional analysis is the b-SFFP sequence. This is due to its excellent signal-to-noise ratio per unit time and T2/T1 contrast and the fact that it does not suffer from excessive signal loss from motion [4]. There also exists a 3D version of the b-SFFP sequence, which allows isotropic acquisitions but in turn has worse contrast between blood and the myocardium and is therefore not commonly used in clinical practice.

MRI acquisitions may be accelerated at the acquisition stage, by undersampling k-space and reconstructing images with incomplete data, which is referred to as compressed sensing. Most compressed sensing approaches work on an individual image basis. One of the few that uses temporal context is [5] where a dynamic 2D+t dictionary is learnt and used to recover missing k-space data.

The limitations caused by the relatively long time required for MRI acquisition have also led to interest in the development of super resolution methods at the post-processing end of the imaging pipeline. A large part of the literature uses non-machine learning approaches. Most of these methods involve least squares error regularisation and assume overlap between numerous slices [6]. Few approaches to super resolution of medical images, more specifically cardiac cine MRI, actually make use of the temporal aspect. In work by Odille *et al.*, a parallel SA stack and two additional stacks taken at orthogonal orientations are used to produce a 3D reconstruction of the heart using regularised least squares, after applying a motion compensating algorithm using the data from the whole cine sequence [7].

Recently, machine learning methods have dominated the research in the biomedical image analysis field. With the increasing availability of computing power, large labelled datasets and open source libraries, deep learning has quickly become the benchmark for many tasks such as image classification and segmentation. The first application of deep learning to image super-resolution consisted of a simple network with three layers, inspired by the idea behind dictionary learning applications to super resolution. The first layer has a small filter size similar to a LR dictionary extracting a small LR image patch, the third layer a larger filter size similar to a HR dictionary upsampling to a bigger higher resolution patch, and the middle layer introduces a non linear mapping between

the two [8]. A small number of training images underlines the simplicity of the approach, which shows pleasing results on natural images, and has become a benchmark in deep learning super resolution.

Some of the best results in image segmentation have been produced by the U-net architecture introduced by Ronneberger *et al.* [9]. The U-net is a convolutional neural network, similar to an autoencoder but including skip connections between input and output layers. The skip connections allow high frequency as well as low frequency information to be processed and make it suitable for super resolution, for which it has been applied to 3D microscopy in two recent studies. The first of them uses a U-net to generate a residual image containing the high frequency information to be added to the LR input [10]. The second compares a U-net to a Super-Resolution Convolutional Neural Network (SRCNN) [8] in 3D to upsample synthetically downsampled microscopy images, showing that both architectures can be used for the task at hand with the U-net consistently outperforming SRCNN [11].

Deep learning has also been applied to super resolution of cardiac MRI in [12], where a single image and a multi-image network are trained to predict residuals which are added to the LR image and give it high frequency information. The data used in that study comes from synthetically down-sampled 3D b-SFFP acquisitions that do not require realignment to account for breathing between acquisitions or patient motion. The same group recently extended the network to an anatomically constrained neural network that resembles a U-net and is able to do super resolution and segmentation aided by the addition of shape priors [13]. In contrast, our work aims to improve standard dynamic 2D data acquired in clinical practice, using the dynamic information in the time sequence to improve the reconstruction. We present a network learning a one-to-one mapping between low resolution (LR) and high resolution (HR) 2D image sequences to generate additional HR LA views from a dynamic SA stack.

Recurrent neural networks (RNN) and especially long short term memory (LSTM) are starting to be applied in medical image analysis. Recurrence can be applied in a spatial sense, by considering adjacent slices in a 3D image. In a study on prostate MRI for cancer segmentation, adjacent 2D slices were fed into a U-net fitted with recurrent layers at every convolution [14]. RNNs have been applied to cardiac cine MRI first by Poudel *et al.* [15], at the lowest resolution level of a U-net to take advantage of low frequency features in consecutive frames of the cardiac cycle. LSTMs have been applied to enhance performance of myocardium segmentations in cardiac cine MRI sequences [16]. In that study, similar to [15], recurrent layers are present in the lower resolution levels of the network architecture.

Up to our knowledge, recurrent networks have not been used for cardiac cine MRI sequence reconstruction. In this paper we propose a method using temporal recurrence to recover HR LA slices from LR acquisitions. Our results show that introducing recurrence improves the quality of the reconstruction, as compared to equivalent single-frame approaches.

2 Materials and Methods

Our method uses an architecture inspired by the U-net, with added recurrent layers, sharing some characteristics with those used in [14–16] for segmentation. While [14] used recurrence on all levels of a U-net, [16] only on the two lower resolution levels, and [15] only on the lowest resolution level, we included recurrence layers on the first two layers, corresponding to the highest resolutions. We limited the number of levels to the first two for two reasons: to save memory and because unlike with segmentation work where the lower frequency features are more important, we are particularly interested in the high frequency information which is needed to convert LR into HR images.

Figure 1 shows the network architecture. The network we propose is inspired by the U-net with a contractive part and skip connections sensitive to low and high frequency details, respectively. At the first and second levels, we introduced LSTM convolution layers. There are a total of five levels in the network each initiated by a 2×2 Max pooling layer. The input data has a size of $128 \times 128 \times 10$, the lowest level therefore operates on samples of size $8 \times 8 \times 10$ where only the very low frequency features are present. We chose to put the recurrent layers on the top levels since we want to enhance the high frequency features of the images and they are mostly present in the first and second levels. Going down to the

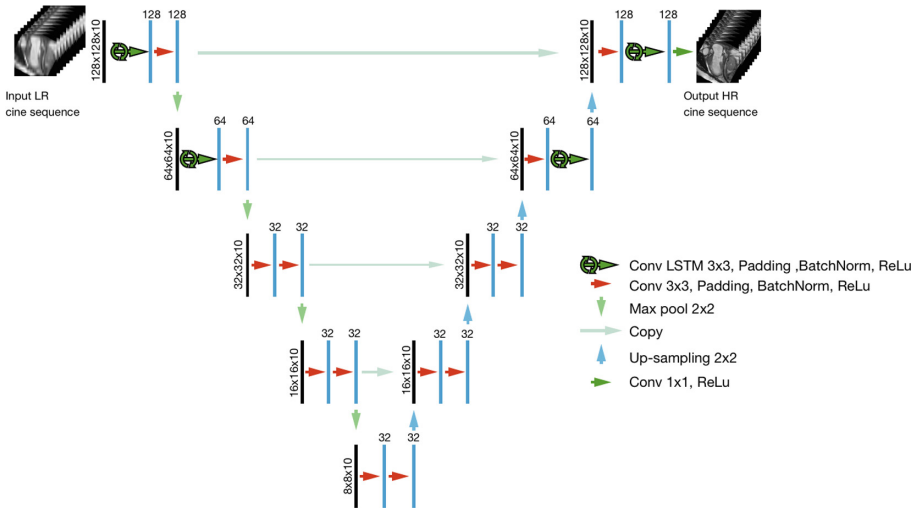

Fig. 1. Network architecture. A network inspired by the U-net with four LSTM modules at the top of the network aiming to enhance high frequency features by taking into account the dynamic aspect of the cine sequence on the original image size and after the first max pooling operation which decreases the image size by a factor of two. Filter numbers are indicated on the diagram on top of the blue bars after every convolution. Arrows represent the different operations. Height, width, and time sizes are shown for every level. (Color figure online)

third level, the data is now $32 \times 32 \times 10$ which we deemed to be less relevant to the high frequency content. We also trained a U-net with the same architecture where all the convolution layers are conventional convolutions with filters of size 3×3, to compare performance between a context sensitive and a static network. Training was performed using an Adam optimiser to minimise mean squared error over 90 epochs, with a learning rate of $4.5.10^{-5}$. The network was written in Python using the Keras library (https://keras.io) running on Tensorflow backend, and training was performed on a Nvidia GeForce GTX 1080 Ti 256 RAM GPU.

2.1 Data

LA views from the Kaggle Data Science Bowl Cardiac Challenge Data [17] were used in training. The dataset consists of cine MRI sequences of over five hundred patients. Every data set has 30 frames, however the number of SA and LA slices differ, as a standardised imaging protocol was not used. VLA and HLA acquisitions were present for most of the patients but a non negligible part of the data had only one or no LA views. The patients have a large spread of age and size which is advantageous to preserve the generalisation properties of the method.

After discarding unusable data (e.g. the ones affected by very strong artefacts or wrongly labeled as LA) by visual inspection, all remaining images were resampled to isotropic resolution of $1.4\,\text{mm} \times 1.4\,\text{mm}$, rotated to the same upright orientation where the base is towards the top of the image and the apex towards the bottom, and down-sampled in the baso-apical direction to match the slice thickness of standard SA slices of 10 mm. In this way, we generated images similar to those that would be reconstructed from the SA stack. Every sequence was also normalised such that all image intensities lie in the range between 0 and 1. We did not differentiate between HLA and VLA views, both were included together in the training, validation and testing datasets. In this way, we aimed to demonstrate the ability of the method to recover images with different appearance (e.g. in terms of the number of chambers), with the eventual goal of using the network to produce slices in any arbitrary orthogonal orientation from SA stacks.

After splitting the sequences of 30 frames into shorter sequences of 10 frames each (to reduce the time needed for training), 3342 LR-HR sequence pairs of 10 frames per sequence were available. 3000 sequences were used for training, 171 set aside for validation, and 171 for testing, ensuring that none of the split sequences were spanning over the training and the validation or testing set. For the static network, all the frames in a sequence were used, which increased the training, validation, and testing data by a factor of 30.

3 Results

Results on the first 5 frames of a HLA sequence are shown in Fig. 2. A representative result on 5 non-adjacent frames of a VLA sequence can be seen in Fig. 3,

with the cardiac contraction more easily visible due to the frames spanning a longer time. Both figures display a sequence from the unseen testing dataset using cubic interpolation in the first row, the result of using static frames only in the second row, the result of the proposed network in the third row, and the ground truth on the bottom row. Each frame has been magnified around the apex, one of the features that is most prone to being missed by the SA stack acquisition. The proposed network manages to recover the apex across the sequence, with much better definition than previously used standard U-Nets.

In addition to qualitative improvements, the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) of the dynamic network output outperform the static network and interpolation. More quantitative results are shown in Table 1, which contains the average values for PSNR and SSIM of the whole testing data set and shows that the dynamic network output is superior to the static network as well as interpolation.

Table 1. Quantitative evaluation (PSNR and SSIM) of interpolated, single frame U-net, and the proposed network results on the whole testing data set which has not been seen by the networks in training.

	Interpolated	U-net	LSTM
PSNR	23.17 dB	25.23 dB	26.57 dB
SSIM	0.72	0.77	0.81

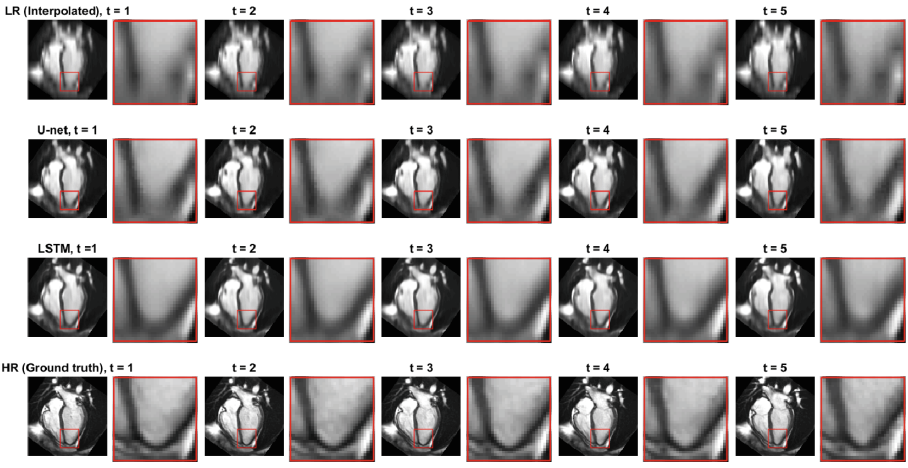

Fig. 2. Result on the first 5 frames of a HLA view cine sequence. The top row shows the LR interpolated input, the second row shows the result given by the static U-net, the third the result given by network including LSTM layers, and the bottom row shows the HR ground truth. This proposed enhanced 4-chamber sequence has a PSNR of 25.28 dB and a SSIM of 0.82 while the static U-net gives a PSNR of 23.83 dB and a SSIM of 0.79 and the interpolated sequence a PSNR of 22.15 dB and a SSIM of 0.73.

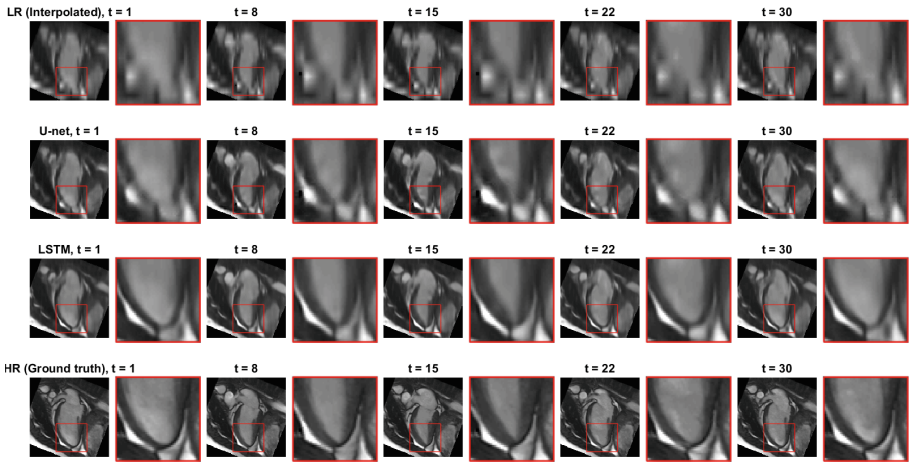

Fig. 3. Result on the 1st, 8th, 15th, 22nd, and 30th frames of a VLA view cine sequence. The top row shows the LR interpolated input, the second row shows the result given by the static U-net, the third the result given by network including LSTM layers, and the bottom row shows the HR ground truth. This proposed enhanced 2-chamber sequence has a PSNR of 27.55 dB and a SSIM of 0.83 while the static U-net gives a PSNR of 25.69 dB and a SSIM of 0.78 and the interpolated sequence a PSNR of 22.23 dB and a SSIM of 0.71.

4 Discussion and Conclusion

We have showed that there is an advantage to using the temporal context for super resolution of cardiac cine MRI sequences, in comparison to the more common approach of reconstructing individual frames. Our proposed architecture, which includes LSTM layers on the upper layers of a U-net, gives qualitatively and quantitatively superior results to an equivalent U-net architecture with no recurrence.

In order to concentrate on the effects of recurrence on reconstructions, we chose experiments that avoid common acquisition artifacts such as misalignment between slices and intensity mismatches. Future work will look into the reconstruction from clinical SA stacks suffering from these artifacts, as well as reconstruction of arbitrarily oriented slices, eventually aiming to reconstruct complete 3D datasets.

Acknowledgments. NMB acknowledges the support of the RCUK Digital Economy Programme grant number EP/G036861/1 (Oxford Centre for Doctoral Training in Healthcare Innovation).

References

1. Petersen, S.E., et al.: UK Biobank's cardiovascular magnetic resonance protocol (2015)
2. Kramer, C.M., Barkhausen, J., Flamm, S.D., Kim, R.J., Nagel, E.: Standardized cardiovascular magnetic resonance imaging (CMR) protocols, society for cardiovascular magnetic resonance: board of trustees task force on standardized protocols. J. Cardiovasc. Magn. Reson. **10**(1), 35 (2008)
3. Ferreira, P.F., Gatehouse, P.D., Mohiaddin, R.H., Firmin, D.N.: Cardiovascular magnetic resonance artefacts. J. Cardiovasc. Magn. Reson. **15**(1), 41 (2013)
4. Scheffler, K., Lehnhardt, S.: Principles and applications of balanced SSFP techniques. Eur. Radiol. **13**(11), 2409–2418 (2003)
5. Caballero, J., Price, A.N., Rueckert, D., Hajnal, J.V.: Dictionary learning and time sparsity for dynamic MR data reconstruction. IEEE Trans. Med. Imaging **33**(4), 979–994 (2014)
6. Plenge, E., et al.: Super-resolution methods in MRI: can they improve the trade-off between resolution, signal-to-noise ratio, and acquisition time? Magn. Reson. Med. **68**(6), 1983–1993 (2012)
7. Odille, F., Bustin, A., Chen, B., Vuissoz, P.-A., Felblinger, J.: Motion-corrected, super-resolution reconstruction for high-resolution 3D cardiac cine MRI. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 435–442. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_52
8. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 184–199. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_13
9. 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
10. Weigert, M., Royer, L., Jug, F., Myers, G.: Isotropic reconstruction of 3D fluorescence microscopy images using convolutional neural networks. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10434, pp. 126–134. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66185-8_15
11. Heinrich, L., Bogovic, J.A., Saalfeld, S.: Deep learning for isotropic super-resolution from non-isotropic 3D electron microscopy. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10434, pp. 135–143. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66185-8_16
12. Oktay, O., et al.: Multi-input cardiac image super-resolution using convolutional neural networks. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9902, pp. 246–254. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46726-9_29
13. Oktay, O., et al.: Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. IEEE Trans. Med. Imaging **37**(2), 384–395 (2018)

14. Zhu, Q., Du, B., Turkbey, B., Choyke, P., Yan, P.: Exploiting interslice correlation for MRI prostate image segmentation, from recursive neural networks aspect. Complexity **2018**, 10 (2018). https://doi.org/10.1155/2018/4185279. Article ID 4185279
15. Poudel, R.P.K., Lamata, P., Montana, G.: Recurrent fully convolutional neural networks for multi-slice MRI cardiac segmentation. In: Zuluaga, M.A., Bhatia, K., Kainz, B., Moghari, M.H., Pace, D.F. (eds.) RAMBO/HVSMR - 2016. LNCS, vol. 10129, pp. 83–94. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-52280-7_8
16. Zhang, D., et al.: A multi-level convolutional LSTM model for the segmentation of left ventricle myocardium in infarcted porcine cine MR images. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 470–473. IEEE (2018)
17. Kaggle data science bowl cardiac challenge data (2015). https://www.kaggle.com/c/second-annual-data-science-bowl/data