# Multi-Input Cardiac Image Super-Resolution using Convolutional Neural Networks

Ozan Oktay[1], Wenjia Bai[1], Matthew Lee[1], Ricardo Guerrero[1],
Konstantinos Kamnitsas[1], Jose Caballero[3], Antonio de Marvao[2],
Stuart Cook[2], Declan O'Regan[2], and Daniel Rueckert[1]

[1] Biomedical Image Analysis Group, Imperial College London, UK
[2] Institute of Clinical Science, Imperial College London, UK
[3] Magic Pony Technology, London, UK

**Abstract.** 3D cardiac MR imaging enables accurate analysis of cardiac morphology and physiology. However, due to the requirements for long acquisition and breath-hold, the clinical routine is still dominated by multi-slice 2D imaging, which hamper the visualization of anatomy and quantitative measurements as relatively thick slices are acquired. As a solution, we propose a novel image super-resolution (SR) approach that is based on a residual convolutional neural network (CNN) model. It reconstructs high resolution 3D volumes from 2D image stacks for more accurate image analysis. The proposed model allows the use of multiple input data acquired from different viewing planes for improved performance. Experimental results on 1233 cardiac short and long-axis MR image stacks show that the CNN model outperforms state-of-the-art SR methods in terms of image quality while being computationally efficient. Also, we show that image segmentation and motion tracking benefits more from SR-CNN when it is used as an initial upscaling method than conventional interpolation methods for the subsequent analysis.

## 1 Introduction

3D magnetic resonance (MR) imaging with near isotropic resolution provides a good visualization of cardiac morphology, and enables accurate assessment of cardiovascular physiology. However, 3D MR sequences usually require long breath-hold and repetition times, which leads to scan times that are infeasible in clinical routine, and 2D multi-slice imaging is used instead. Due to limitations on signal-to-noise ratio (SNR), the acquired slices are usually thick compared to the in-plane resolution and thus negatively affect the visualization of anatomy and hamper further analysis. Attempts to improve image resolution are typically carried out either during the acquisition stage (sparse k-space filling) or retrospectively through super resolution (SR) of single/multiple image acquisitions.

**Related work:** Most of the SR methods recover the missing information through the examples observed in training images, which are used as a prior to link low and high resolution (LR-HR) image patches. Single image SR methods, based on the way they utilize training data, fall into two categories: non-parametric

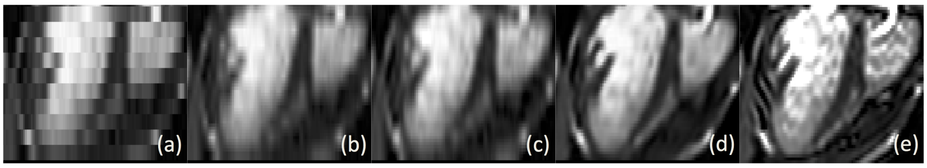

Fig. 1: The low resolution image (a) is upscaled using linear (b) and cubic spline (c) interpolations, and the proposed method (d) which shows a high correlation with the ground-truth high resolution image (e) shown on the rightmost.

and parametric. The former aims to recover HR patches from LR ones via a co-occurrence prior between the target image and external training data. Atlas-based approaches such as the patch-match method [15] and non-local means based single image SR [10] methods are two examples of this category. These approaches are computationally demanding as the candidate patches have to be searched in the training dataset to find the most suitable HR candidate. Instead, compact and generative models can be learned from the training data to define the mapping between LR and HR patches. Parametric generative models, such as coupled-dictionary learning based approaches, have been proposed to upscale MR brain [14] and cardiac [3] images. These methods benefit from sparsity constraint to express the link between LR and HR. Similarly, random forest based non-linear regressors have been proposed to predict HR patches from LR data and have been successfully applied on diffusion tensor images [1]. Recently, convolutional neural network (CNN) models [5,6] have been put forward to replace the inference step as they have enough capacity to perform complex nonlinear regression tasks. Even by using a shallow network composed of a few layers, these models [6] achieved superior results over other state-of-the-art SR methods.

**Contributions:** In the work presented here, we extend the SR-CNN proposed by [5,6] with an improved layer design and training objective function, and show its application to cardiac MR images. In particular, the proposed approach simplifies the LR-HR mapping problem through residual learning and allows training a deeper network to achieve improved performance. Additionally, the new model can be considered more data-adaptive since the initial upscaling is performed by learning a deconvolution layer instead of a fixed kernel [6]. More importantly, a multi-input image extension of the SR-CNN model is proposed and exploited to achieve a better SR image quality. By making use of multiple images acquired from different slice directions one can further improve and constrain the HR image reconstruction. Similar multi-image SR approaches have been proposed in [11,12] to synthesize HR cardiac images; however, these approaches did not make use of available large training datasets to learn the appearance of anatomical structures in HR. Compared to the state-of-the-art image SR approaches [6,15], the proposed method shows improved performance in terms of peak signal-to-noise-ratio (PSNR) and structural similarity index measure (SSIM) [18]. We show that cardiac image segmentation can benefit from SR-CNN as the seg-

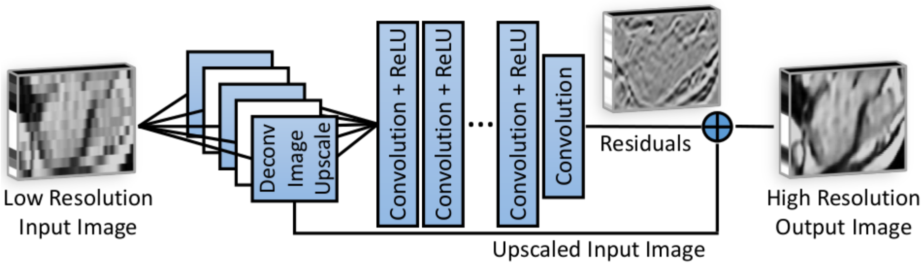

Fig. 2: The proposed single image super resolution network model

mentations generated from super-resolved images are shown to be similar to the manual segmentations on HR images in terms of volume measures and surface distances. Lastly, we show that cardiac motion tracking results can be improved using SR-CNN as it visualizes the basal and apical parts of the myocardium more clearly compared to the conventional interpolation methods (see Fig.1).

## 2 Methodology

The SR image generation is formulated as an inverse problem that recovers the high dimensional data through the MR image acquisition model [7], which has been the starting point of approaches in [3,11,15]. The model links the HR volume $\boldsymbol{y} \in \mathbb{R}^M$ to the low dimensional observation $\boldsymbol{x} \in \mathbb{R}^N$ $(N \ll M)$ through the application of a series of operators as: $\boldsymbol{x} = \boldsymbol{DBSMy} + \boldsymbol{\eta}$ where $\boldsymbol{M}$ defines the spatial displacements caused due to respiratory and cardiac motion, $\boldsymbol{S}$ is the slice selection operator, $\boldsymbol{B}$ is a point-spread function (PSF) used to blur the selected slice, $\boldsymbol{D}$ is a decimation operator, and $\boldsymbol{\eta}$ is the Rician noise model. The solution to this inverse problem estimates a conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ that minimizes the cost function $\Psi$ defined by $\boldsymbol{y}$ and its estimate $\Phi(\boldsymbol{x}, \boldsymbol{\Theta})$ obtained from LR input data. The estimate is obtained through a CNN parameterized by $\boldsymbol{\Theta}$ that models the distribution $p(\boldsymbol{y}|\boldsymbol{x})$ via a collection of hidden variables. For the smooth $\ell_1$ norm case, the loss function is defined as $\min_{\boldsymbol{\Theta}} \sum_i \Psi_{\ell_1}\left(\Phi(\boldsymbol{x}_i, \boldsymbol{\Theta}) - \boldsymbol{y}_i\right)$, where $\Psi_{\ell_1}(r) = \{0.5\, r^2 \text{ if } |r| < 1\,,\, |r| - 0.5 \text{ otherwise}\}$ and $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ denote the training samples. The next section describes the proposed CNN model.

**Single Image Network:** The proposed model, shown in Fig. 2, is formed by concatenating a series of convolutional layers (Conv) and rectified linear units (ReLU) [8] to estimate the non-linear mapping $\Phi$, as proposed in [6] to upscale natural images. The intermediate feature maps $h_j^{(n)}$ at layer $n$ are computed through Conv kernels (hidden units) $w_{kj}^n$ as $\max\left(0, \sum_{k=1}^K h_k^{(n-1)} * w_{kj}^n\right) = h_j^n$ where $*$ is the convolution operator. As suggested by [16], in order to obtain better non-linear estimations, the proposed architecture uses small Conv kernels (3x3x3) and a large number of Conv+ReLU layers. Such approach allows training of a deeper network. Different to the models proposed in [6,5], we include an

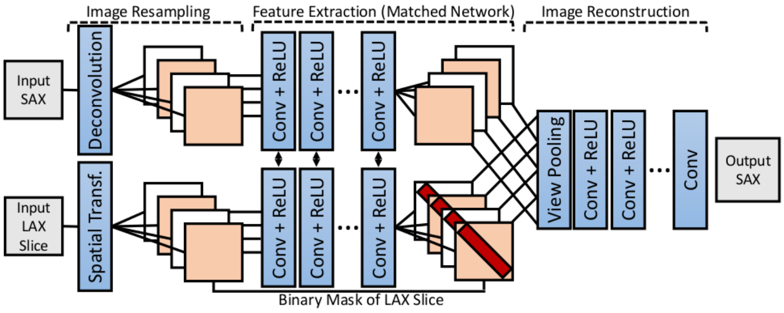

Fig. 3: The proposed Siamese multi-image super resolution network model.

initial upscaling operation within the model as a deconvolution layer (Deconv) $(\boldsymbol{x} \uparrow U) * w_j = h_j^0$ where $\uparrow$ is a zero-padding upscaling operator and $U = M/N$ is the upscaling factor. In this way, upsampling filters can be optimized for SR applications by training the network in an end-to-end manner. This improves the image signal quality in image regions closer to the boundaries. Instead of learning to synthesize a HR image, the CNN model is trained to predict the residuals between the LR input data and HR ground-truth information. These residuals are later summed up with the linearly upscaled input image (output of Deconv layer) to reconstruct the output HR image. In this way, a simplified regression function $\Phi$ is learned where mostly high frequency signal components, such as edges and texture, are predicted (see Fig 2). At training time, the correctness of reconstructed HR images is evaluated based on the $\Psi_{\ell_1}(.)$ function, and the model weights are updated by back-propagating the error defined by that function. In [19] the $\ell_1$ norm was shown to be a better metric than the $\ell_2$ norm for image restoration and SR problems. This is attributed to the fact that the weight updates are not dominated by the large prediction errors.

**Multi-Image Network:** The single image model is extended to multi-input image SR by creating multiple input channels (MC) from given images which are resampled to the same spatial grid and visualize the same anatomy. In this way, the SR performance is enhanced by merging multiple image stacks, e.g. long-axis (LAX) and short axis (SAX) stacks, acquired from different imaging planes into a single SR volume. However, when only a few slices are acquired, a mask or distance map is required as input to the network to identify the missing information. Additionally, the number of parameters is supposed to be increased so that the model can learn to extract in image regions where the masks are defined, which increases the training time accordingly. For this reason, a Siamese network [4] is proposed as a third model (see Fig. 3) for comparison purposes, which was used in similar problems such as shape recognition from multiple images [17]. The first stage of the network resamples the input images into a fixed HR spatial grid. In the second stage the same type of image features are extracted from each channel which are sharing the same filter weights. In the final stage, the features are pooled and passed to another Conv network to reconstruct

| Exp (a) | PSNR (dB) | SSIM | # Filters/Atlases |
|---|---|---|---|
| Linear | 20.83±1.10 | .70±.03 | – |
| CSpline | 22.38±1.13 | .73±.03 | – |
| MAPM | 22.75±1.22 | .73±.03 | 350 |
| sh-CNN | 23.67±1.18 | .74±.02 | 64,64,32,1 |
| CNN | 24.12±1.18 | .76±.02 | 64,64,32,16,8,4,1 |
| de-CNN | **24.45±1.20** | **.77±.02** | 64,64,32,16,8,4,1 |

Table 1: Quantitative comparison of different image upsampling methods.

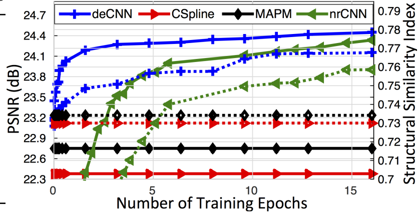

Fig. 4: Results on the testing data, PSNR (solid) and SSIM (dashed)

the output HR image. The view pooling layer averages the corresponding features from all channels over the areas where the images are overlapping. The proposed models are initially pre-trained with small number of layers to better initialize the final deeper network training, which improves the network performance [5].

## 3 Results

The models are evaluated on end-diastolic frames of cine cardiac MR images acquired from 1233 healthy adult subjects. The images are upscaled in the direction orthogonal to the SAX plane. The proposed method is compared against linear, cubic spline, and multi-atlas patchmatch (MAPM) [15] upscaling methods in four different experiments: image quality assessment for (a-b) single and multi-input cases, (c) left-ventricle (LV) segmentation, (d) LV motion tracking.

**Experimental details:** In the first experiment, an image dataset containing 1080 3D SAX cardiac volumes with voxel size 1.25x1.25x2.00 mm, is randomly split into two subsets and used for single-image model training (930) and testing (150). The images are intensity normalized and cropped around the heart. Synthetic LR images are generated using the acquisition model given in Section 2, which are resampled to a fixed resolution 1.25x1.25x10.00 mm. The PSF is set to be a Gaussian kernel with a full-width at half-maximum equal to the slice thickness [7]. For the LR/HR pairs, multiple acquisitions could be used as well, but an unbalanced bias would be introduced near sharp edges due to spatial misalignments. For the evaluation of multi-input models, a separate clinical dataset of 153 image pairs of LAX cardiac image slices and SAX image stacks are used, of which 10 pairs are split for evaluation. Spatial misalignment between SAX and LAX images are corrected using image registration [9]. For the single/multi image model, seven consecutive Conv layers are used after the upscaling layer. In the Siamese model, the channels are merged after the fourth Conv layer.

**Image Quality Assessment:** The upscaled images are compared with the ground-truth HR 3D volumes in terms of PSNR and SSIM [18]. The latter measure assesses the correlation of local structures and is less sensitive to image noise. The results in Table 1 show that learning the initial upscaling kernels (de-CNN) can improve ($p = .007$) the quality of generated HR image compared

Table 2: Image quality results obtained with three different models: single-image de-CNN, Siamese, and multi-channel (MC) that uses multiple input images.

| Exp (b) | de-CNN(SAX) | Siamese(SAX/4CH) | MC(SAX/4CH) | MC(SAX/2/4CH) |
|---|---|---|---|---|
| PSNR (dB) | 24.76±0.48 | 25.13±0.48 | 25.15±0.47 | **25.26±0.37** |
| SSIM | .807±.009 | .814±.013 | .814±.012 | **.818±.012** |
| $p$ - values | 0.005 | 0.016 | 0.017 | - |

Table 3: Segmentation results for different upsampling methods, CSpline ($p = .007$) and MAPM ($p = .009$). They are compared in terms of mean and Hausdorff distances (MYO) and LV cavity volume differences (w.r.t. manual annotations).

| | | Linear | CSpline | MAPM | de-CNN | High Res |
|---|---|---|---|---|---|---|
| Exp (c) | LV Vol Diff (ml) | 11.72±6.96 | 10.80±6.46 | 9.55±5.42 | **9.09±5.36** | 8.24±5.47 |
| | Mean Dist (mm) | 1.49±0.30 | 1.45±0.29 | 1.40±0.29 | **1.38±0.29** | 1.38±0.28 |
| | Haus Dist (mm) | 7.74±1.73 | 7.29±1.63 | 6.83±1.61 | **6.67±1.77** | 6.70±1.85 |

to convolution only network (CNN) using the same number of trainable parameters. Additionally, the performance of 7-layer network is compared against the 4-layer shallow network from [6] (sh-CNN). Addition of extra Conv layers to the 7-layer model is found to be ineffective due to increased training time and negligible performance improvement. In Fig. 4, we see that CNN based methods can learn better HR synthesis models even after a small number of training epochs. On the same figure, it can be seen that the model without the residual learning (nrCNN) underperforms and requires a large number of training iterations.

**Multi-Input Model:** In the second experiment, we show that the single image SR model can be enhanced by providing additional information from two and four chamber (2/4CH) LAX images. The results given in Table 2 show that by including LAX information in the model, a modest improvement in image visual quality can be achieved. The improvement is mostly observed in image regions closer to areas where the SAX-LAX slices overlap, as can be seen in Fig. 5 (a-d). Also, the results show that the multi-channel (MC) model performs slightly better than Siamese model as it is given more degrees-of-freedom, whereas the latter is more practical as it trains faster and requires fewer trainable parameters.

**Segmentation Evaluation:** As a subsequent image analysis, 18 SAX SR images are segmented using a state-of-the-art multi-atlas method [2]. The SR images generated from clinical 2D stack data with different upscaling methods are automatically segmented and those segmentations are compared with the manual annotations performed on ground-truth HR 3D images. Additionally, the HR images are segmented with the same method to show the lower error bound. The quality of segmentations are evaluated based on the LV cavity volume measure and surface-to-surface distances for myocardium (MYO). The results in Table 3 show that CNN upscaled images can produce segmentation results similar to the ones obtained from HR images. The main result difference between the SR meth-

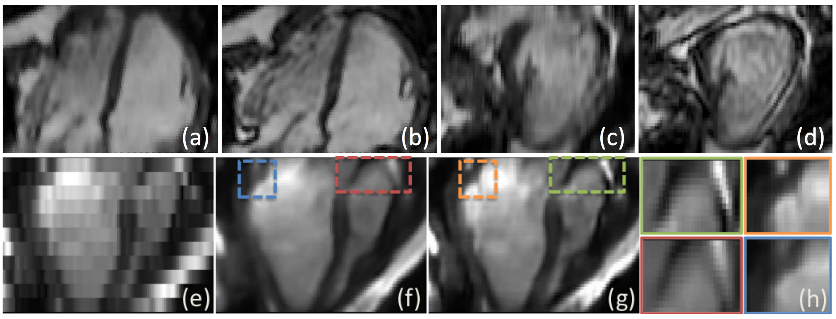

Fig. 5: The LV is better visualized by using multi-input images (b,d) compared to single image SR (a,c). Also, the proposed method (g) performs better than MAPM [15] (f) in areas where uncommon shapes are over-smoothed by atlases.

ods is observed in image areas where thin and detailed boundaries are observed (e.g. apex). As can be seen in Fig. 5 (e-h), the MAPM over-smooths areas closer to image boundaries. Inference of the proposed model is not as computationally demanding as brute-force searching (MAPM), which requires hours for a single image, whereas SR-CNN can be executed in 6.8 s on GPU or 5.8 mins CPU on average per image. The shorter runtime makes the SR methods more applicable to subsequent analysis, as they can replace the standard interpolation methods.

**Motion Tracking:** The clinical applications of SR can be extended to MYO tracking as it can benefit from SR as a preprocessing stage to better highlight the ventricle boundaries. End-diastolic MYO segmentations are propagated to end-systolic (ES) phase using B-Spline FFD registrations [13]. ES meshes generated with CNN and linear upscaling methods are compared with tracking results obtained with 10 3D-SAX HR images based on Hausdorff distance. The proposed SR method produces tracking results (4.73±1.03 mm) more accurate ($p = 0.01$) than the linear interpolation (5.50±1.08 mm). We observe that the images upscaled with the CNN model follow the apical boundaries more accurately, which is shown in the supplementary material: `www.doc.ic.ac.uk/~oo2113/M16`

## 4 Discussion and Conclusion

The results show that the proposed SR approach outperforms conventional upscaling methods both in terms of image quality metrics and subsequent image analysis accuracy. Also, it is computationally efficient and can be applied to image analysis tasks such as segmentation and tracking. The experiments show that these applications can benefit from SR images since 2D stack image analysis with SR-CNN can achieve similar quantitative results as the analysis on isotropic volumes without requiring long acquisition time. We also show that the proposed model can be easily extended to multiple image input scenarios to obtain better SR results. SR-CNN's applicability is not only limited to cardiac images but to other anatomical structures as well. In the proposed approach, inter-slice and stack spatial misalignments due to motion are handled using a registration method. However, we observe that large slice misplacements can degrade SR accuracy. Future research will focus on that aspect of the problem.

## References

1. Alexander, D.C., Zikic, D., Zhang, J., Zhang, H., Criminisi, A.: Image quality transfer via random forest regression: Applications in diffusion MRI. In: MICCAI, pp. 225–232. Springer (2014)
2. Bai, W., Shi, W., O'Regan, D.P., Tong, T., Wang, H., Jamil-Copley, S., Peters, N.S., Rueckert, D.: A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: Application to cardiac MR images. IEEE TMI 32(7), 1302–1315 (2013)
3. Bhatia, K.K., Price, A.N., Shi, W., Rueckert, D.: Super-resolution reconstruction of cardiac MRI using coupled dictionary learning. In: IEEE ISBI. pp. 947–50 (2014)
4. Bromley, J., Guyon, I., Lecun, Y., Sckinger, E., Shah, R.: Signature verification using a "Siamese" time delay neural network. NIPS pp. 737–744 (1994)
5. Dong, C., Deng, Y., Change Loy, C., Tang, X.: Compression artifacts reduction by a deep convolutional network. In: IEEE CVPR. pp. 576–584 (2015)
6. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE PAMI 38(2), 295–307 (2016)
7. Greenspan, H.: Super-resolution in medical imaging. The Computer Journal 52(1), 43–63 (2009)
8. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998)
9. Lötjönen, J., Pollari, M., Lauerma, K.: Correction of movement artifacts from 4-D cardiac short-and long-axis MR data. In: MICCAI, pp. 405–12. Springer (2004)
10. Manjón, J.V., Coupé, P., Buades, A., Fonov, V., Collins, D.L., Robles, M.: Non-local MRI upsampling. MedIA 14(6), 784–792 (2010)
11. Odille, F., Bustin, A., Chen, B., Vuissoz, P.A., Felblinger, J.: Motion-corrected, super-resolution reconstruction for high-resolution 3D cardiac cine MRI. In: MICCAI, pp. 435–442. Springer (2015)
12. Plenge, E., Poot, D., Niessen, W., Meijering, E.: Super-resolution reconstruction using cross-scale self-similarity in multi-slice MRI. In: MICCAI. pp. 123–130 (2013)
13. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J.: Non-rigid registration using free-form deformations: Application to breast MR images. IEEE TMI 18(8), 712–721 (1999)
14. Rueda, A., Malpica, N., Romero, E.: Single-image super-resolution of brain MR images using overcomplete dictionaries. MedIA 17(1), 113–132 (2013)
15. Shi, W., Caballero, J., Ledig, C., Zhuang, X., Bai, W., Bhatia, K., de Marvao, A., Dawes, T., O'Regan, D., Rueckert, D.: Cardiac image super-resolution with global correspondence using multi-atlas patchmatch. In: MICCAI. pp. 9–16 (2013)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
17. Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E.: Multi-view convolutional neural networks for 3D shape recognition. In: IEEE CVPR. pp. 945–953 (2015)
18. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE TIP 13(4), 600–612 (2004)
19. Zhao, H., Gallo, O., Frosio, I., Kautz, J.: Is L2 a good loss function for neural networks for image processing? arXiv preprint arXiv:1511.08861 (2015)