Cardiac MR Segmentation from Undersampled *k*-space Using Deep Latent Representation Learning

Jo Schlemper[1(✉)], Ozan Oktay[1], Wenjia Bai[1], Daniel C. Castro[1], Jinming Duan[1], Chen Qin[1], Jo V. Hajnal[2], and Daniel Rueckert[1]

[1] Biomedical Image Analysis Group, Imperial College London, London, UK
{js3611,o.oktay13,w.bai,dc315,j.duan,c.qin15,d.rueckert}@imperial.ac.uk
[2] Imaging and Biomedical Engineering Clinical Academic Group, King's College London, London, UK
jo.hajnal@kcl.ac.uk

Abstract. Reconstructing magnetic resonance imaging (MRI) from undersampled *k*-space enables the accelerated acquisition of MRI but is a challenging problem. However, in many diagnostic scenarios, perfect reconstructions are not necessary as long as the images allow clinical practitioners to extract clinically relevant parameters. In this work, we present a novel deep learning framework for reconstructing such clinical parameters directly from undersampled data, expanding on the idea of *application-driven* MRI. We propose two deep architectures, an end-to-end synthesis network and a latent feature interpolation network, to predict cardiac segmentation maps from extremely undersampled dynamic MRI data, bypassing the usual image reconstruction stage altogether. We perform a large-scale simulation study using UK Biobank data containing nearly 1000 test subjects and show that with the proposed approaches, an accurate estimate of clinical parameters such as ejection fraction can be obtained from fewer than 10 *k*-space lines per time-frame.

1 Introduction

Cardiovascular MR (CMR) imaging enables accurate quantification of cardiac chamber volume, ejection fraction and myocardial mass, which are crucial for diagnosing, assessing and monitoring cardiovascular diseases (CVDs), the leading cause of death globally. However, one limitation of CMR is the slow acquisition time. A routine CMR protocol can take from 20 to 60 min, which makes the tool costly and less accessible to worldwide population. In addition, CMR often requires breath-holds which can be difficult for patients; therefore, accelerating the CMR acquisition is essential. Over the last decades, numerous approaches have been proposed for accelerated MR imaging, including parallel imaging, compressed sensing [7] and, more recently, deep learning approaches [6].

Reconstructing images from accelerated and undersampled MRI is an ill-posed problem and, essentially, all approaches must exploit some type of redundancies or assumptions on underlying data to resolve the aliasing caused by

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11070, pp. 259–267, 2018.
https://doi.org/10.1007/978-3-030-00928-1_30

sub-Nyquist sampling. In the case of dynamic cardiac cine reconstruction, high spatiotemporal redundancy and sparsity can be exploited, however, the acceleration factor for a near perfect reconstruction is currently limited up to 9 [10]. We argue that one effective way of pushing the acceleration factor even higher is to move to the concept of *application-driven MRI* [2]. The key insight is that in many cases, the images are not an end in themselves, but rather means of accessing clinically relevant parameters which are obtained as post-processing steps, such as segmentation or tissue characterisation. Therefore, it is more effective to instead combine the reconstruction and post-processing steps and tailor the acquisition protocol to obtain the final results as accurately and efficiently as possible. In particular, if the end-goal is significantly more compressible than the original image, then one can expect further acceleration and still obtain satisfactory results [3,5]. This work focuses on a scenario where we obtain cardiac segmentation maps directly from heavily undersampled dynamic MR data.

Our contribution is the following: firstly, we propose two network architectures to learn such a mapping. The first model, *Syn-net*, exploits the spatiotemporal redundancy of the input to directly generate the segmentation map. However, under heavy aliasing artefact, the extracted features may not be useful for segmentation. To address the latter case, we propose the second model, *LI-net*, which first predicts the low dimensional latent code of the corresponding segmentation map, which is subsequently decoded. Secondly, we extensively evaluate the two models with large-scale simulation studies to demonstrate the effectiveness of the proposed approaches for various acceleration factors. In particular, we show that for the case where undersampled image contains sufficient geometrical information, Syn-net outperforms LI-net but in a more challenging scenario where only one line of k-space is sampled per frame, LI-net outperforms Syn-net. Finally, we study the latent space structure of these architectures to demonstrate that the models learn useful representations of the data. This work potentially enables interesting future works in which reconstruction, post-processing and analysis stages are integrated to yield smarter imaging protocols.

2 Proposed Methods

End-to-End Synthesis Network (Syn-net): Let $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$ be dataset of fully-sampled complex valued (dynamic) images $x = \{x_i \in \mathbb{C} | i \in S\}$, where S denotes indices on a pixel grid, and the corresponding segmentation labels $y = \{y_i | i \in S\}$ representing different tissue types with $y_i \in \{1, 2, \dots C\}$. Let $u = \{u_i \in \mathbb{C} \mid u = F_u^H F_u x\}$ denote an undersampled image, where F_u is the undersampling Fourier encoding matrix. Let $p(y_i \mid x)$ be the true distribution of i-th pixel label given an image, and $r(u \mid x, \mathcal{M})$ represent the sampling distribution of the undersampled images given an image x and a (pseudo) random undersampling mask generator $\mathcal{M}$. We aim to learn a synthesis network $q(y_i \mid u, \theta)$, termed *Syn-net*, which uses a convolutional neural network (CNN) to model the probability distribution of segmentation maps given the undersampled image parameterised by θ. We train the network by the following modified cross-entropy (CE) loss:

$$\mathcal{L}(\theta) = \sum_{(x,y)\in\mathcal{D}} \mathbb{E}_{u\sim r}\left[\sum_{i\in S} p(y_i \mid x_i)\log q(y_i \mid u_i, \theta)\right], \tag{1}$$

where we take the expectation over differently undersampled images. In practice, we generate one different undersampling pattern on-the-fly for each mini-batch training as an approximation to the expectation. For the network architecture, we use an architecture inspired by the state-of-the-art segmentation network, *U-net* [9], shown in Fig. 1.

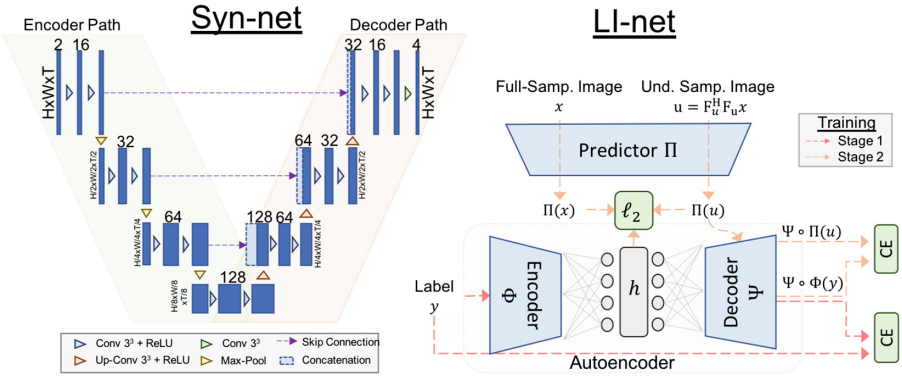

Fig. 1. *(Left)* The detailed architecture of Syn-net: the changes in the number of features are shown above the tensor. *(Right)* For LI-net, the two-stage training strategy is outlined. The same encoder and decoder as Syn-net can be used for LI-Net

Latent Feature Interpolation Network (LI-net): Syn-net assumes that the input data contains sufficient geometrical information to generate the target segmentation. For heavily undersampled (and therefore aliased) images, this assumption may not be valid as the aliasing could mislead the network from identifying the correct boundaries. In the latter case, synthesis is still possible as long as the target domain has a compact, discriminative latent representation $h \in \mathcal{H}$ that can be predicted, an approach motivated by *TL-network* [4]. Such a network can be trained in following steps. In stage 1, one trains an auto-encoder (AE) in the target domain $y = \Psi(\Phi(y;\theta_{enc});\theta_{dec})$, $y \in \mathcal{Y}$, which is a composition of encoder $\Phi : \mathcal{Y} \to \mathcal{H}$ and decoder $\Psi : \mathcal{H} \to \mathcal{Y}$, parameterised by θ_{enc} and θ_{dec} respectively and $\mathcal{H}$ is a low-dimensional latent space. The AE can be trained using the ℓ_2 norm or CE loss. In stage 2, one trains a predictor network $\Pi : \mathcal{X} \to \mathcal{H}$, parameterised by θ_{pred}. For a given input-target pair (x, y), the predictor attempts to predict the latent code $h = \Phi(y;\theta_{enc})$ from x. This is trained using the ℓ_2 norm in the latent space: $d_{\mathcal{H}}(y, x) = \|\Phi(y;\theta_{enc}) - \Pi(x;\theta_{pred})\|_2$. Once the predictor is trained, one can obtain an input-output mapping by the composition $\hat{y} = \Psi(\Pi(x;\theta_{dec});\theta_{pred})$.

In our work, the AE is trained to learn the compact representation of segmentations and the predictor is trained to interpolate these from dynamic undersampled images, hence termed a *latent feature interpolation network* (*LI-net*).

In stage 1, we train the AE using CE loss. In stage 2, we modify our objective to further encourage the network to produce a *consistent* prediction for differently undersampled versions of the same reference image. This constraint is implemented by forcing the network to produce the same latent code for undersampled images as for the fully-sampled image. Furthermore, we add a CE term $d_{\mathrm{CE}}(y, \Psi \circ \Pi(u))$ to ensure that an accurate segmentation can be obtained from the code. Therefore, our objective term is as follows (here λ_i's are hyper-parameters to be adjusted based on the preferred end-goal):

$$\mathcal{L}(\theta_{pred}) = \sum_{(x,y)\in\mathcal{D}} \mathbb{E}_{u\sim r}\Big[d_{\mathcal{H}}(y,u) + \lambda_1 d_{\mathcal{H}}(y,x) + \lambda_2 \|\Pi(x) - \Pi(u)\|_2 + \lambda_3 d_{\mathrm{CE}}(y, \Psi \circ \Pi(u)) \Big]. \quad (2)$$

3 Experiments and Results

Dataset and Undersampling: Experiments were performed using 5000 short-axis cardiac cine MR images from the UK Biobank study [8], which is acquired using bSSFP sequence, matrix size $N_x \times N_y \times T = 208 \times 187 \times 50$, a pixel resolution of $1.8 \times 1.8 \times 10.0\,\mathrm{mm}^3$ and a temporal resolution of 31.56 ms. Since the manual annotations are only available at end-systolic (ES) and end-diastolic (ED) frames but we are interested in segmenting the entire time sequence, we use [1], which well agrees with the manual segmentations, to generate the labels for the left-ventricular (LV) cavity, the myocardium and the right-ventricular (RV) cavity for all time-frames including apical, mid and basal slices, which were then treated as the ground truth labels for this work. We split the data into 4000 training subjects and 1000 test subjects, and we simulated random undersampling using variable density 1D undersampling masks. These masks were generated on-the-fly. As only the magnitude images were available we synthetically generated the phase maps (smoothly varying 2D sinusoid waves) on-the-fly to make the simulation more realistic by removing the conjugate symmetry in k-space. Different levels of acceleration factors ($1/n_l$) were considered, $n_l \in [1, 100]$ where n_l is the number of lines per time-frame. Note for fully-sampled image, $n_l = 168$.

Model and Parameters: The input to the network is 2D+t undersampled data and the output is a sequence of segmentation map. Note that z-slices were processed separately due to large slice thickness. The detail of the Syn-net is shown in Fig. 1. To make a fair comparison between the two architectures, we used the encoding path of Syn-net as both encoder Φ and predictor Π, and the decoding path as decoder Ψ. The size of the latent code was set to be $|h| = 1024$. Note that fully-connected layers are used to join the encoder, the latent code and the decoder. They were trained with mini-batch size 8 using Adam with initial learning rate 10^{-4}, which was reduced by a factor of 0.8 every 2 epochs. The AE in LI-net was trained for 30 epochs to ensure that the Dice scores for each class reached 0.95. For both models, we first trained the network to perform segmentation from fully-sampled data as a warm start. The number of

lines was gradually reduced and by 10^{th} epoch, we uniformly sampled n_l from $[0, 168]$. The training error for both models plateaued within 50 epochs. For LI-net, the hyper-parameters for the loss function were empirically chosen to be $\lambda_1 = 1$, $\lambda_2 = 10^{-4}$, $\lambda_3 = 10$, which we found to work sufficiently. For data augmentation, we generated affine transformations on-the-fly. We used PyTorch for implementation.

Evaluation: We first took the trained models and evaluated their Dice scores for LV, myocardium (Myo) and RV for $n_l \in [1, 100]$. For each subject, we only included ES and ED frames but aggregated the results across all short-axis slices. The Dice scores versus the number of acquired k-space lines are shown in Fig. 2. The networks maintained the performance up to about 20 lines per frame, demonstrating the capability of the models to directly interpolate the anatomical boundary even in the presence of the aliasing artefact. In general, Syn-net showed superior performance, indicating that the extracted spatiotemporal features are directly useful for segmentation. In particular, we report that the LI-net underperformed as it does not employ the skip-connection as Syn-net does, which limits how accurately it can delineate the boundaries. We speculate, however, increasing the capacity of network is likely to improve the results. Interestingly, LI-net outperformed Syn-net for the case of segmentation from 1 line, suggesting that in more challenging domains the approach of LI-net to interpolate the latent code is still a viable option.

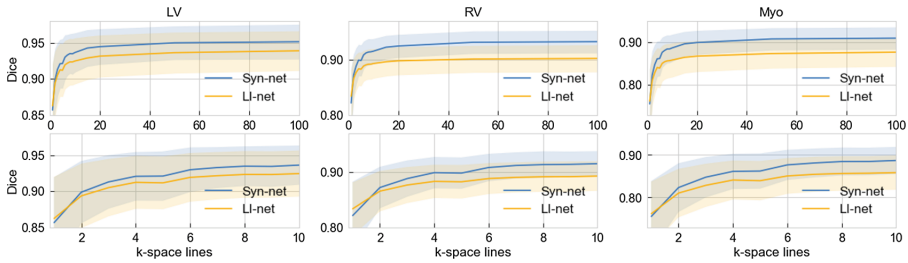

Fig. 2. Dice scores of Syn-net vs LI-net. The second row expands $n_l \in [1, 10]$, the solid lines and the shaded areas show the mean and the standard deviation respectively.

In the second experiment, the models were further fine-tuned for a fixed number of lines for $n_l \in \{1, 10, 20\}$ separately. From the obtained segmentation maps, we computed LV ES/ED volumes (ESV/EDV) RV ESV/EDV, LV mass (LVM) and ejection fraction (EF). The mean percentage errors across all test subjects were reported in Table 1. Syn-net consistently performs better than LI-net and has relatively small errors ($<$7.7%) for all values for $n_l \in \{10, 20\}$. Both models showed low error for EF, where the correlation coefficient was 0.81 for both models for $n_l = 20$. The examples of the segmentation maps are shown in Fig. 3. Note that due to heavy aliasing artefact of the input image, we instead

visualised the temporally averaged image for x-y plane, which was obtained by combining all k-space lines across the temporal axis into a single k_x-k_y grid.

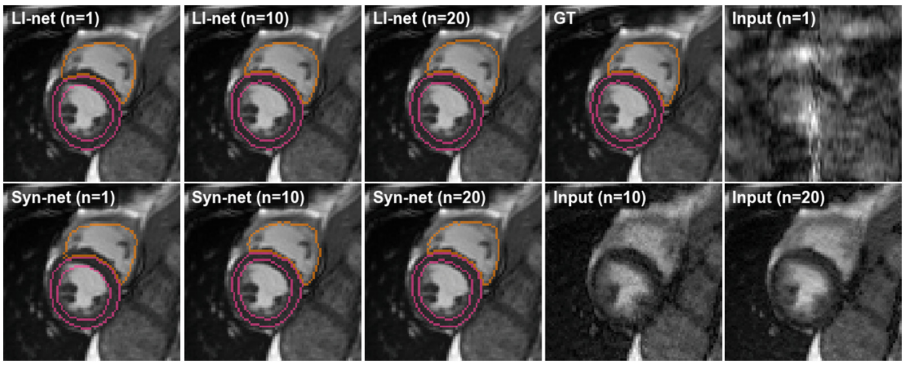

Fig. 3. Visualisation of the ground truth image overlaid with the obtained segmentations. LI-net produced more anatomically regularised, consistent segmentations. Syn-net produced segmentations that are occasionally anatomically implausible but more faithful to the boundary.

Although in theory we expect the reconstructed segmentation maps to be independent of the aliasing artefact present in the input, this is not always the case (Fig. 3). To measure such variability, we define *within subject distances*: given a fully-sampled image, we undersample it differently for $n_{\text{trial}} = 100$ times. From the predicted segmentation maps given by a model, we computed the mean shape, to which we then calculated mean contour distance (MD) and Hausdorff distance (HD) of individual predictions. Small distances indicate that the segmentation is consistent. However, if the network simply produces a *population mean shape* independent of the input, then the above distances can be very low even without producing useful segmentations. To get a better picture, we also measured the *between subject distances*, which computes MD and HD between the population mean shape (a mean predicted shape across *all* subjects) and the individual subject mean shapes. For both experiments, $n_{\text{subject}} = 100$ subject were used and the averaged distances are shown in Table 2. Indeed, we see that LI-net shows lower values for *within subject* distances, indicating that it produces more consistent segmentations than Syn-net ($p \ll 0.01$, Wilcoxon

Table 1. Average percentage errors (%) for each clinical parameter

n_l	LV ESV			LV EDV			RV ESV			RV EDV			LVM			EF		
	1	10	20	1	10	20	1	10	20	1	10	20	1	10	20	1	10	20
LI-net	7.9	3.6	3.2	15.8	7.9	7.0	11.7	6.5	6.6	18.4	10.5	10.6	25.0	13.7	12.9	8.8	5.5	4.9
Syn-net	9.0	4.2	3.4	14.6	7.2	6.1	9.9	4.9	4.1	13.4	7.7	6.5	11.4	6.8	5.8	8.2	5.5	4.6

rank-sum). High *between subject* distances indicate that both models are generating segmentation maps closer to subject-specific means than to the population mean.

Table 2. The *within-subject* and *between-subject* distances of the segmentations

n_l	HD (Within)				MD (Within)				HD (Between)				MD (Between)			
	Myo		RV		Myo		RV		Myo		RV		Myo		RV	
	1	10	1	10	1	10	1	10	1	10	1	10	1	10	1	10
LI-net	3.63	2.58	4.68	3.67	1.47	0.92	1.71	1.14	9.69	10.30	12.55	14.31	4.40	4.94	5.02	5.91
Syn-net	6.47	3.23	6.96	5.05	1.83	0.99	2.15	1.34	10.62	10.34	11.56	15.22	4.16	4.81	4.78	6.03

Finally, we investigate the latent space of the models. For 5 subjects, we generated 50 undersampled images for each number of lines $n_l \in \{1, 5, 10, 15, 20\}$. Here all undersampled images have the same target segmentation per subject. For LI-net, we plotted the predicted latent code $h \in \mathcal{H}$ for these images. For Syn-net, we plotted the activation map before the first upsampling layer to see whether the network exploits any latent space structure for generating the segmentations. We visualised them using Principal Component Analysis (PCA) and t-distributed stochastic neighbour embedding (t-SNE) with $d = 2$, as shown in Fig. 4, where subjects are colour-coded and brighter means higher acceleration factor.

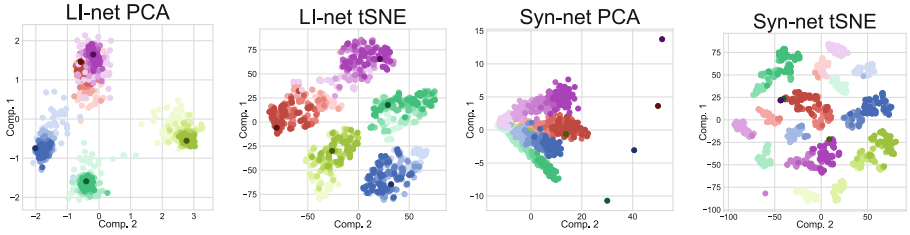

Fig. 4. Visualising the distribution of the latent representations of LI-net and Syn-net. *(Left to right)* LI-net PCA, LI-net t-SNE, Syn-net PCA, Syn-net t-SNE. The darkest points are the latent representations of the fully sampled images, for reference.

For LI-net, both for PCA and t-SNE, the latent space is clearly clustered by individual subjects, indicating that the predictions are indeed consistent for different undersampling patterns. In addition, as the latent code is discriminative for each subject, it enables fitting a classifier for subject-based prediction tasks. On the other hand, for Syn-net, although there are per-subject clusters, there is also a clear tendency to favour clustering the points by the acceleration factors, as seen in the t-SNE plot. Note that since Syn-net also exploits skip connections, one can conclude that the network exploits different reconstruction strategies

for different acceleration factors. Another interesting observation is that in Syn-net PCA, the distances between all the points are reduced as the acceleration factor is increased. This means that the latent features for Syn-net are less discriminative when images are heavily aliased. However, the extracted features become gradually more discriminative as more lines are sampled.

4 Conclusion and Discussion

In this work we explored an application-driven MRI, where our end-goal was to extract segmentation maps directly from extremely undersampled data, bypassing image reconstruction. Remarkably, when at least 10 lines per frame are acquired, we showed that we could already compute clinical parameters within 10% error. Even though Syn-net provided better performance overall, and LI-net exhibited more well-behaved latent-space structure. In future work, the latent code of LI-net could be used as a feature for classification tasks, where we may be able to classify whether a patient is abnormal, directly from a few lines of k-space. This work opens a huge avenue for future research where joint pipelines can be exploited for smarter MR imaging that is both fast and accurate.

Acknowledgements. JS is partially funded by EPSRC Grant (EP/P001009/1).

References

1. Bai, W., et al.: Human-level CMR image analysis with deep fully convolutional networks (2017). arXiv preprint: arXiv:1710.09289
2. Caballero, J., Bai, W., Price, A.N., Rueckert, D., Hajnal, J.V.: Application-driven MRI: joint reconstruction and segmentation from undersampled MRI data. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014, Part I. LNCS, vol. 8673, pp. 106–113. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10404-1_14
3. Gaur, P., Grissom, W.A.: Accelerated MRI thermometry by direct estimation of temperature from undersampled k-space data. Magn. Reson. Med. **73**(5), 1914–1925 (2015)
4. Girdhar, R., Fouhey, D.F., Rodriguez, M., Gupta, A.: Learning a predictable and generative vector representation for objects. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016, Part VI. LNCS, vol. 9910, pp. 484–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_29
5. Guo, Y., Lingala, S.G., Zhu, Y., Lebel, R.M., Nayak, K.S.: Direct estimation of tracer-kinetic parameter maps from highly undersampled brain dynamic contrast enhanced MRI. Magn. Reson. Med. **78**(4), 1566–1578 (2017)
6. Hammernik, K., et al.: Learning a variational network for reconstruction of accelerated MRI data. Magn. Reson. Med. (2017)
7. Lustig, M., Donoho, D., Pauly, J.M.: Sparse MRI: the application of compressed sensing for rapid MR imaging. Magn. Reson. Med. **58**(6), 1182–1195 (2007)
8. Petersen, S.E., et al.: UK Biobank's cardiovascular magnetic resonance protocol. J. Cardiovasc. Magn. Reson. **18**(1), 8 (2016)

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, Part III. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
10. Schlemper, J., Caballero, J., Hajnal, J.V., Price, A., Rueckert, D.: A deep cascade of convolutional neural networks for dynamic MR image reconstruction. IEEE Trans. Med. Imaging **37** (2017)