XmoNet: A Fully Convolutional Network for Cross-Modality MR Image Inference

Sophia Bano[1(✉)], Muhammad Asad[2], Ahmed E. Fetit[3], and Islem Rekik[4]

[1] Wellcome/EPSRC Centre for Interventional and Surgical Sciences and Department of Computer Science, University College London, London, UK
sophia.bano@ucl.ac.uk
[2] Imagination Technologies, Kings Langley, UK
[3] Biomedical Image Analysis Group, Imperial College London, London, UK
[4] BASIRA Lab, CVIP, School of Science and Engineering (Computing), University of Dundee, Dundee, UK

Abstract. Magnetic resonance imaging (MRI) can generate multi-modal scans with complementary contrast information, capturing various anatomical or functional properties of organs of interest. But whilst the acquisition of multiple modalities is favourable in clinical and research settings, it is hindered by a range of practical factors that include cost and imaging artefacts. We propose XmoNet, a deep-learning architecture based on fully convolutional networks (FCNs) that enables cross-modality MR image inference. This multiple branch architecture operates on various levels of image spatial resolutions, encoding rich feature hierarchies suited for this image generation task. We illustrate the utility of XmoNet in learning the mapping between heterogeneous T1- and T2-weighted MRI scans for accurate and realistic image synthesis in a preliminary analysis. Our findings support scaling the work to include larger samples and additional modalities.

Keywords: Fully convolutional networks · MRI · Multimodal Image generation

1 Introduction

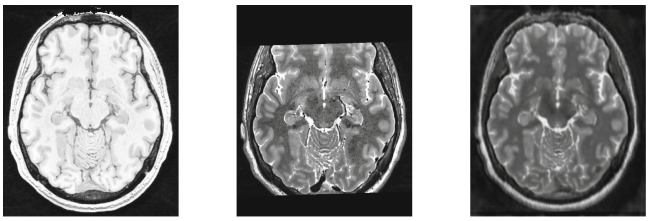
Magnetic resonance imaging (MRI) is the key imaging technology used to aid the diagnosis and management of a wide range of diseases. Visual characteristics of tissues of interest can be acquired via a variety of MR modalities (e.g. T1-weighted, T2-weighted, FLAIR, diffusion-weighted and diffusion-tensor imaging), each offering complementary contrast mechanisms. For instance in neuro-oncology, T1-weighted scans are favourable for observing brain structures whereas T2-weighted scans can provide rich information for tumour localisation. However, a number of factors impede acquisition of multimodal scans in clinical settings; particularly cost, limited availability of scanning time and patient discomfort [7]. In research settings and imaging clinical trials, it is common to face

I. Rekik et al. (Eds.): PRIME 2018, LNCS 11121, pp. 129–137, 2018.
https://doi.org/10.1007/978-3-030-00320-3_16

heterogeneous or incomplete datasets due to similar reasons, as well as acquisition artefacts and data corruption. This has motivated various efforts in the MR literature that can broadly be divided into two categories: (i) improving image acquisition and reconstruction strategies, and (ii) synthesising a target modality given a separate source modality; also known as cross-modality generation.

Cross-modality generation has attracted the attention of the medical image computing community in recent years. Work by D. H. Ye *et al.* [14] investigated a modality propagation approach, where for each point in the target image a patch-based search is carried out across a database of images, utilising nearest neighbours' information for estimating target modality values. The work was motivated by the observation that local and contextual similarities observed in one modality can often extend to other modalities. Evaluation of the approach illustrated effectiveness in synthesising T2-weighted and DTI signals given a source T1-weighted input, including successful application on brain tumour scans. Y. Lu *et al.* [10] proposed a novel distance measure that used patch based intensity histogram and Weber Local Descriptor features to search the most similar patch from the database for modality synthesis.

Recently, Y. Huang *et al.* [7] proposed a weakly supervised technique that requires only a few registered multi-modal image pairs for effective cross-modality generation. The technique works through mapping different image features of the underlying tissues, preserving global statistical image properties across modalities, and subsequently refining the features to ensure local geometrical structures are preserved within each modality. Additionally, manifold matching is used to select target-modality features from the most similar source-modality subjects; thus complementing unpaired data with the original training pairs. Effectiveness of the technique was illustrated in cross modality generation between T1- and T2-weighted scans, as well as T2- and PD-weighted scans.



(a) (b) (c)

Fig. 1. The proposed XmoNet enables cross-modality MR image inference, as demonstrated here with an example. The architecture takes as input a T1-weighted slice in (a) and predicts the corresponding T2-weighted slice in (c). Ground-truth T2-weighted slice is shown in (b) for reference. Visual inspection of (b) and (c) illustrates practical utility of XmoNet in achieving cross-modality mapping, along with generation of areas which have missing ground-truth data; a high-value application in clinical and research settings.

Deep learning algorithms, particularly Convolutional Neural Networks (CNNs), have rapidly gained widespread adoption within the medical image computing community. Work by Bahrami *et al.* [2] studied the utility of CNNs for mapping cross-domain scans, albeit for a resolution mapping problem (3T to 7T MRI) as opposed to generation of missing modalities. In their earlier work, Bahrami *et al.* [3] made use of high- and low-frequency visual features, thus capturing variations among 3 T scans with various levels of quality. Evaluation was carried out on various paired MR datasets of healthy subjects, as well as patients with epilepsy and MCI. A. Ben-Cohen *et al.* [4] combined a fully convolutional network (FCN) with a conditional generative adversarial network (GAN) to generate PET data from CT for improving automated lesion detection. Y. Hiasa *et al.* [5] proposed CycleGAN-based MR to CT orthopedic image synthesis method in which the accuracy at the bone boundaries was improved by adding the gradient consistency loss.

We contribute XmoNet, a deep learning architecture for rapid and accurate cross(X)-MOdality learning; and carry out a preliminary analysis to examine its effectiveness on heterogeneous MR data. The architecture is based on fully convolutional networks (FCN) and utilises parallel pathways to encode low- and high-frequency visual features, allowing mapping of rich feature hierarchies. Preliminary analysis demonstrated accurate and realistic synthesis of target T2-weighted images from source T1-weighted data (see Fig. 1); our findings support scaling the work to include larger samples and additional modalities.

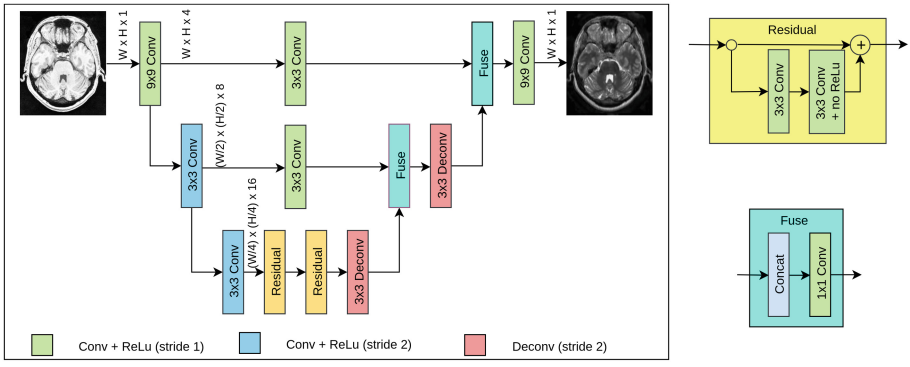


Fig. 2. Flowchart of the proposed XmoNet. The input T1-weighted slice is convolved using multiple pathways at different resolutions. The output from each pathway is upsampled with a deconvolution operation and then fed into a fusion layer. The multiple higher resolution pathways allows high-frequency patterns to be preserved. Multiple residual layers are added to the lowest resolution path, which ensures mapping of low-frequency visual patterns from the source data.

2 Proposed Method

Inspired by recent successes of fully convolutional networks (FCNs) [1,9,11] the XmoNet utilises a FCN architecture that learns the cross-modality mapping from T1- onto T2-weighted MR data. Figure 2 shows a flowchart of the proposed architecture. Given an input slice, the network utilises several strided-convolutional layers to reduce spatial dimensionality whilst increasing the number of activation channels at every branch, following intuition from the well-established VGG architecture [12]. Through the use of multiple pathways, we map different frequency levels of visual features from the input scan. The use of multiple pathways is inspired by FCN methods proposed for semantic segmentation [1,9], and ensures capturing of high-frequency visual patterns. Merging with deconvolution layers is carried out in order to spatially upsample the activations whilst reducing the number of channels. These are followed by fuse layers for pathway concatenation. Residual layers are also used for cross-modality mapping of low-frequency visual patterns. The network uses 4, 8 and 16 filters in the first, second and third convolutional pathways respectively. The two residual blocks use 16 filters each and the filters in the upsampling layers are reduced to 8 and 4 in the first and second branch, respectively. L2 loss is used for the network training.

3 Experimental Analysis

3.1 Dataset

In this preliminary analysis we used the public MNI-HISUB25 dataset by Kulaga-Yoskovitz *et al.* [8] which includes submillimetric, high-resolution T1- and T2-weighted brain scans of 25 healthy subjects. The dataset is available in NIfTI format and is labelled for hippocampal subfields. Resolutions are $0.6 \times 0.6 \times 0.6\,\text{mm}^2$ and $0.45 \times 0.45 \times 2.0\,\text{mm}^2$ for T1- and T2-weighted scans, respectively. Kulaga-Yoskovitz *et al.* [8] pre-processed the captured scans for spatial normalisation to MNI152-space as well as registration of the two modalities. The final, pre-processed T1- and T2-weighted scans have a $0.4 \times 0.4 \times 0.4\,\text{mm}^3$ resolution in MNI152-space which are used in our experiments.

3.2 Experimental Setup

We used the open-source med2image[1] tool for MRI axial slice extraction. This was then followed by extracting only those slices that contained hippocampi since region around hippocampi is of high relevance to the diagnosis of brain disorders such as Alzheimers' disease. In total, 2431 slices (452×542 pixels) contained hippocampi regions; these formed the data for our experiments. We performed two experiments: (i) input to XmoNet was the whole T1-weighted image (452×542 pixels), and (ii) input to XmoNet was a cropped region selected around right hippocampus of the T1-weighted image (128×128 pixels).

[1] https://github.com/FNNDSC/med2image. last access: 20072018.

T2-weighted images in the dataset failed to capture complete brain structures; most of them had zero-pixel regions in place of lower or/and upper parts of the images (Fig. 1). Incorporating corrupted regions into the learning process would obscure network training; we alleviated this through generating exclusion masks obtained by detecting regions in the T2-weighted images where no signal was present. A blob size threshold was used to ensure zero-pixel brain structures were not included within the masks. Such masks were subsequently used during model training, ensuring the loss is computed only for pixels within which an anatomical signal was present. Similarly, the masks were used during the validation stage when computing evaluation metrics.

3.3 Validation Protocol

(a) 80% of the data was selected (first 20 subjects; 1961 slices in total) for model training. The remaining data (5 subjects; 470 slices) were completely unseen during the training process but held out for evaluation. (b) Furthermore, we performed k-fold cross-validation (k=5) to provide additional reassurance; each fold contained an average of 485 slices representing the scans of 5 subjects. The cross-validation loop consisted of model training over 4 folds and subsequent testing on the remaining fold. An i7-CPU workstation with NVIDIA 1080 GTX card installed was used for the analysis. The training process took place over 20 hours (approx. 5 hours per fold) for 5-fold validation. Observed testing rate was 48 slices per second.

3.4 Evaluation Metrics

Peak signal-to-noise ratio (PSNR) and structural similarity (SIMM) [13] metrics are used in existing method [2,3,6] for the quantitative evaluation of reconstructed images/patches, hence we used the same evaluation metrics. Given a ground-truth X and a generated image Y both of height H and width W; mean square error (MSE) is first obtained:

$$MSE = \frac{1}{HW} \sum_{i=0}^{H-1} \sum_{i=0}^{W-1} [X(i,j) - Y(i,j)]^2 \tag{1}$$

PSNR (in dB) is then computed as follows (MAX_X is the maximum possible pixel intensity; 255 here):

$$PSNR = 10 \log_{10} \left(\frac{MAX_X^2}{MSE} \right) \tag{2}$$

SIMM measures the perceived change in Y relative to X and is computed as:

$$SIMM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{3}$$

where μ_x and μ_y are the mean, σ_x and σ_y are the variance and σ_{xy} are the covariances of X and Y. c_1 and c_2 depend on the dynamic range of pixel intensities; needed to stabilise division on weak denominators [13]. Increase in PSNR suggests an improvement in signal to noise ratio i.e. lower noise and/or better image generation. SIMM, on the other hand, captures the structural similarity between a synthesised and a ground-truth image. PSNR and SIMM were computed for only those pixels that lie outside the defined exclusion masks. Visual inspection was further carried out to assess realism of synthesised images, particularly regions where no T2 ground-truth is available.

Table 1. Mean and standard deviation (Std) for PSNR and SIMM obtained via 5-fold cross validation for synthesis of T2-weighted (i) complete images and (ii) right hippocampus subregions.

fold#		1		2		3		4		5	
		Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Complete slice	PSNR	30.48	0.58	30.74	0.56	30.98	0.67	30.96	0.53	31.11	1.21
	SIMM	0.77	0.09	0.79	0.10	0.80	0.10	0.80	0.10	0.78	0.11
Hippocampi region	PSNR	28.45	0.78	27.75	0.14	27.83	0.34	27.76	0.22	29.24	0.72
	SIMM	0.60	0.12	0.61	0.13	0.61	0.12	0.60	0.13	0.63	0.14

4 Results and Discussion

Table 1 shows the result for the 5-fold validation for the complete and right-hippocampus T2-weighted sub-region generation. Both PSNR and SIMM measures are higher for the complete T2-weighted image synthesis compared to the T2-weighted sub-region synthesis as complete image synthesis managed to better capture high resolution details resulting in relatively accurate and sharp image generation. This is because the variance of each pixel in complete T2-weighted image is low during training compared to the sub-region image.

Figure 3 shows a set of original images (T1-weighted network input and noisy T2-weighted ground-truth) as well as synthesised T2-weighted images for 8 different subjects. The proposed XmoNet is capable of achieving cross-modality mapping from T1 onto T2. Visual inspection of these figures suggests that synthesised images better capture overall brain structures (with respect to source T1-weighted images) than the original T2 scans; successful synthesis of regions with heavily missing T2 signal is achieved (Fig. 3(d)–(f)).

A number of limitations exist in this study. Firstly, the generated brain regions for which no T2 baseline exists require thorough validation and assessment by medical experts. Additionally, network input-output is currently a T1-T2 generation route; exploring the opposite scenario of T2-T1 generation was not carried out. Furthermore, testing data used in the study was obtained from the same source as the training/fine-tuning data; studying network's generalisability to different acquisition settings was not carried out.

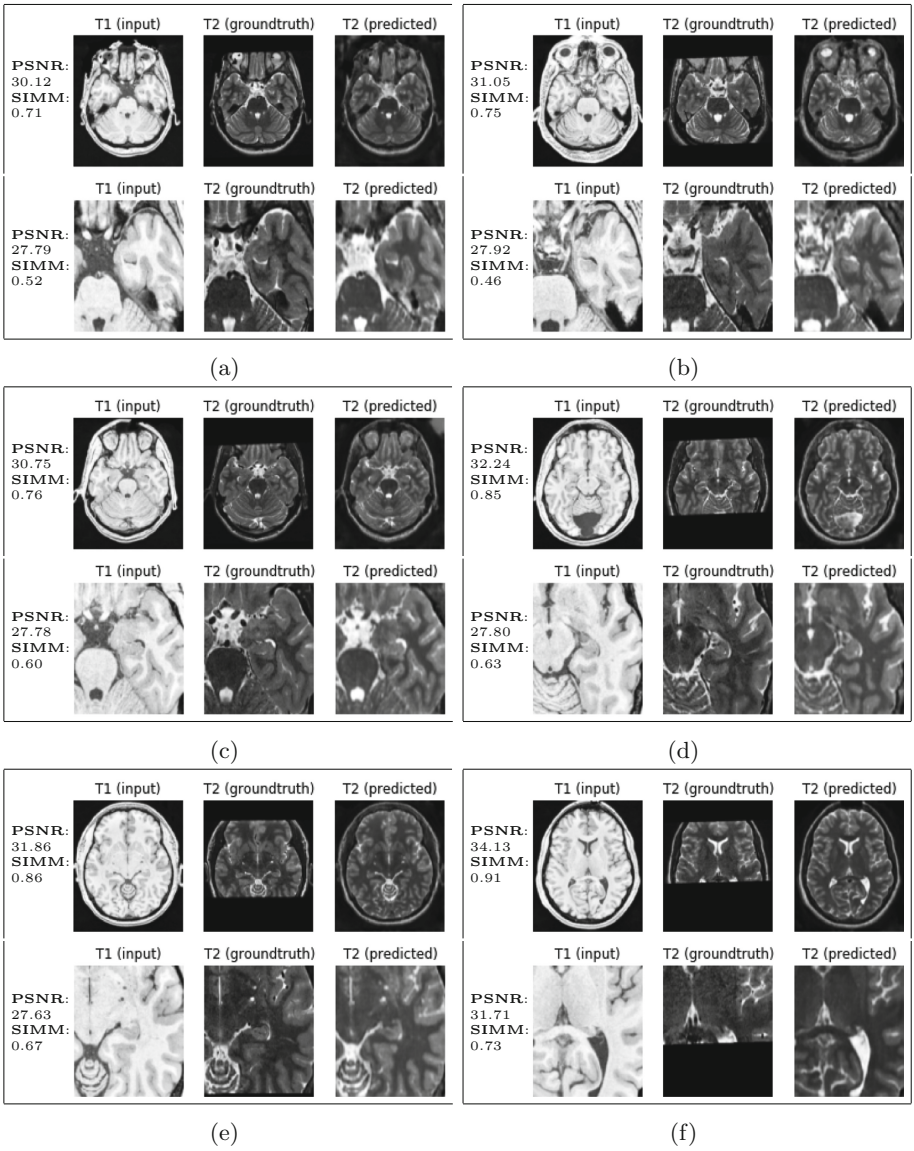


Fig. 3. Representative images of axial slices from six subjects (a)–(f); (a)–(c) sections at the level of Pons showing missing frontal lobe parts in the T2 (ground-truth) scans, (d)–(f) showing missing frontal and parietal lobe parts in the T2 (ground-truth) scans. XmoNet automatically generated the missing parts as shown in T2 (predicted). For each subject, upper row shows complete image synthesis while lower row shows results on the hippocampus sub-region images.

In addition to the above, validating XmoNet on larger datasets will drive our future efforts. Additionally, rigorous comparison against performance of state-of-the-art methods is crucial. An interesting application of the work is synthesis of images of non-healthy regions e.g. brain tumours. Although the model was designed for MR image generation, it can be adopted to incorporate non-MR based modalities (e.g. CT). Moreover, cross modality inference in 3D images is also of interest [6], hence adopting our model to 3D images can also be considered.

5 Conclusions

We proposed XmoNet, a CNN designed for the problem of cross-modality MR image generation. The network utilises a fully convolutional architecture, where multiple pathways are used to capture a hierarchy of low- and high-frequency visual patterns. A preliminary analysis was carried out on brain MR scans of 25 healthy subjects. Quantitative evaluation and qualitative visual inspection illustrated the utility of XmoNet for accurate and realistic synthesis of T2-weighted images from source T1-weighted data. Our findings support extending the analysis to incorporate larger datasets and additional modalities.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
2. Bahrami, K., Rekik, I., Shi, F., Shen, D.: Joint Reconstruction and segmentation of 7T-like MR images from 3T MRI based on cascaded convolutional neural networks. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10433, pp. 764–772. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66182-7_87
3. Bahrami, K., Shi, F., Zong, X., Shin, H.W., An, H., Shen, D.: Reconstruction of 7T-like images from 3T MRI. IEEE Trans. Med. Imaging **35**(9), 2085 (2016)
4. Ben-Cohen, A., et al.: Cross-modality synthesis from CT to PET using FCN and GAN networks for improved automated lesion detection. arXiv preprint arXiv:1802.07846 (2018)
5. Hiasa, Y., et al.: Cross-modality image synthesis from unpaired data using CycleGAN: Effects of gradient consistency loss and training data size. arXiv preprint arXiv:1803.06629 (2018)
6. Huang, Y., Shao, L., Frangi, A.F.: Simultaneous super-resolution and cross-modality synthesis of 3D medical images using weakly-supervised joint convolutional sparse coding. arXiv preprint arXiv:1705.02596 (2017)
7. Huang, Y., Shao, L., Frangi, A.F.: Cross-modality image synthesis via weakly coupled and geometry co-regularized joint dictionary learning. IEEE Trans. Med. Imaging **37**(3), 815–827 (2018)
8. Kulaga-Yoskovitz, J., et al.: Multi-contrast submillimetric 3 tesla hippocampal subfield segmentation protocol and dataset. Sci. Data **2**, 150059 (2015)
9. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)

10. Lu, Y., Sun, Y., Liao, R., Ong, S.H.: A modality synthesis framework: using patch based intensity histogram and weber local descriptor features. In: International Symposium on Biomedical Imaging (ISBI), pp. 1126–1129. IEEE (2015)
11. 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
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
13. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, vol. 2, pp. 1398–1402. IEEE (2003)
14. Ye, D.H., Zikic, D., Glocker, B., Criminisi, A., Konukoglu, E.: Modality propagation: coherent synthesis of subject-specific scans with data-driven regularization. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. LNCS, vol. 8149, pp. 606–613. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40811-3_76