Robust Multi-Modal MR Image Synthesis

Thomas Joyce*, Agisilaos Chartsias*, and Sotirios A. Tsaftaris

The University of Edinburgh
t.joyce@ed.ac.uk

Abstract. We present a multi-input encoder-decoder neural network model able to perform MR image synthesis from any subset of its inputs, outperforming prior methods in both single and multi-input settings. This is achieved by encouraging the network to learn a modality invariant latent embedding during training. We demonstrate that a spatial transformer module [7] can be included in our model to automatically correct misalignment in the input data. Thus, our model is robust both to missing and misaligned data at test time. Finally, we show that the model's modular nature allows transfer learning to different datasets.

Keywords: MRI, synthesis, neural network, brain

1 Introduction

Automatic generation of synthetic images has a wide range of applications within medical imaging research and technology. For example, synthetic images can replace corrupt or missing images, extend a dataset to additional modalities [16], be used for attenuation correction [2, 13], pathology detection [1, 18] and for improving the performance of other imaging algorithms [6, 16]. However, in order for a synthesis system to be widely applicable, in addition to producing high quality synthetic images it should also be robust to the suboptimal conditions that are often present in real use situations, such as missing and unaligned data.

Here we focus on Magnetic Resonance Imaging (MRI), a non-invasive imaging technique that can obtain images of different contrasts, referred to here as modalities, by applying different sets of pulse sequence parameters. In this context, we demonstrate that a multi-input neural network can be trained end-to-end to synthesise images in a target modality. Most previous synthesis work has focused on single input synthesis, and multi-input synthesis introduces additional potential problems, such as missing inputs in some modalities, and misaligned inputs. Here, due to the composition of the network and the costs imposed during training, the resulting model is naturally robust to missing inputs, and so any subset of the possible inputs can be used at test time to generate an output. Additionally, through the inclusion of a spatial transformer module [7], the network can overcome misalignment between inputs.

Another key problem in MRI synthesis is that many different MR scanners are used, and the different images produced (of the non parametric type) have

* These authors contributed equally.

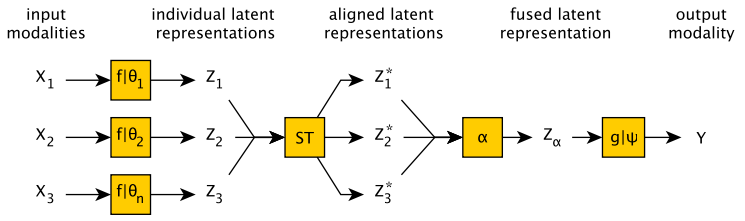

Fig. 1. A high-level schematic of our model.

non-identical statistical properties, which typically require several pre-processing steps to alleviate. Thus, an algorithm trained on images from a particular scanner may degrade significantly in performance when applied to images from other sources. To address this, we demonstrate transfer learning by fine-tuning a trained decoder with a very small number of volumes from a different source. This improves transfer performance on images from the second source by $\sim 70\%$.

In summary, we propose a model that outperforms recent approaches in both single and multi-input image synthesis by $\sim 14-30\%$, is robust to missing and misaligned inputs, and is able to achieve good transfer learning results with a very small number of volumes.

2 Previous Work

Because of its broad applicability, there has been significant previous work on MR image synthesis. The majority of this work has focused on synthesising an output modality from a single input modality, and has been mainly concerned with producing increasingly accurate synthesis results [18, 4, 12]. In addition, there has been work on creating pseudo-healthy data [1], synthesising CT from MRI [5], and estimating pulse sequence parameters to achieve accurate synthesis [8].

Improvements can be made by incorporating information into the input, in addition to raw pixel intensities [17, 9]. In [17] the Location Sensitive Deep Network (LSDN) is proposed, which improves results by conditioning the synthesis on the position in the volume from which the patch comes. Another approach [9], uses random forests to solve the patch based regression problem, and incorporates both multi-scale information and context description features in order to improve the final synthesis. This approach has also been demonstrated in a multi-input task, and multi-input synthesis has also been explored in an atlas based setting [14]. However, although both approaches can fuse information from multiple sources to produce accurate synthetic results, they are not designed to robustly handle missing input modalities.

3 Proposed Approach

Our model, illustrated in Fig 1, is comprised of encoder modules, alignment and fusion modules and a decoder module. The model takes as input full 2D image

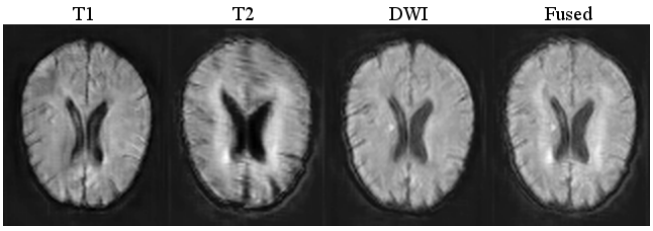

Fig. 2. Example channel from a latent embedding. The first 3 images are the T1, T2 and DWI embeddings respectively, the final image is the fused embedding. Informative aspects from the individual embeddings are combined in the fused embedding.

slices of each input modality and synthesises an image of the same size in the output modality. The network has one encoder for each input modality, which maps the input images into a shared latent representation space. The latent representations are then aligned by the spatial transformer module [7], and combined to produce a single fused latent representation. Finally, this fused latent representation is decoded to produce an output image in the target modality.

The encoded representation of a MR brain image from one modality should be semantically equivalent to the encoded representation from any other modality. For example, although the cerebrospinal fluid in the lateral ventricles produces high intensity voxels in T2 modality images, and low intensity voxels in T1 modality images, in the shared latent representation, these voxels should have the same representation (see Fig. 2).

3.1 Details

Encoders: Let $\mathbf{X}_1, ..., \mathbf{X}_n$ be image slices in the n input modalities. Each of the n inputs has its own encoder f, parametrised by θ_i. The encoded latent representations are defined as $\mathbf{Z}_i = f(\mathbf{X}_i|\theta_i)$, $i \in [0, n]$, where $\mathbf{Z}_i$ is a tensor, and there is one $\mathbf{Z}_i$ for each input modality. The encoders are fully convolutional networks, as shown in Fig. 3, with an architecture inspired by the U-Net [11], due to its proven representational power. Each latent representation $\mathbf{Z}_i$ is a 16 channel tensor with the same width and height as the input image.
Alignment: Next, the n latent representations are aligned using the spatial transformer module, yielding aligned latent representations $\mathbf{Z}_1^*, ..., \mathbf{Z}_n^*$. The spatial transformer is a neural network able to learn to apply affine transformations to its inputs. Here we use it to align all latent representations to the first. To achieve this, we input $\mathbf{Z}_1$ and $\mathbf{Z}_i$ into the spatial transformer for each $i \in [2, n]$, and get $\mathbf{Z}_i^*$ output, which is a geometrically transformed $\mathbf{Z}_i$. As all other representations are transformed to match $\mathbf{Z}_1$, $\mathbf{Z}_1$ is left unchanged, and so $\mathbf{Z}_1^* = \mathbf{Z}_1$. The spatial transformer is parametrised by weights μ, which are learnt implicitly by the overall cost function, see Sec 4.
Fusion: The n aligned latent representations $\mathbf{Z}_1^*, ..., \mathbf{Z}_n^*$ are then combined via a fusion operation α to produce a single fused latent representation $\mathbf{Z}_\alpha$. Here we

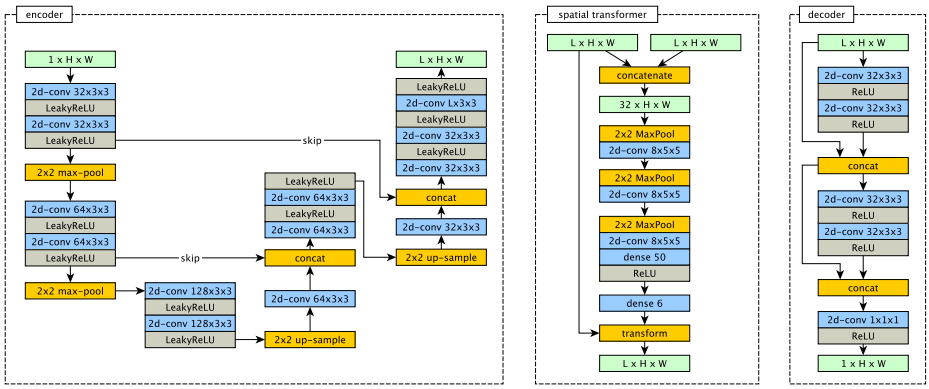

Fig. 3. Architectures of encoder (left), spatial transformer (center), and decoder (right) modules. The encoders are inspired by U-Net [11] and each one, f, has its own set of weights. In our experiments we use L=16.

use the voxel-wise max as the fusion operation, which allows features present in only one input modality to be preserved in the fused representation.

Decoder: Finally, $\mathbf{Z}_\alpha$ is decoded by the decoder g, parametrised by ψ, giving the final prediction $g(\mathbf{Z}_\alpha|\psi)$. The exact architecture of g is shown in Fig 3 and is chosen to be shallow to encourage an informative latent representation. It also contains skip connections as they help gradient flow [3].

4 Cost Function

One central goal for the model is to exploit all inputs in order to produce an accurate final synthetic image. To encourage this directly, we train the model to minimise the mean absolute error (MAE) between the synthesised image and the target. This is the first term in our cost function (below). However, simply training the network end-to-end to minimise final synthesis accuracy does not produce a model that is robust to missing inputs. This is because the model is not encouraged to learn a shared latent representation. In order to encourage robustness, two additional costs are used during training. Firstly, the latent representations are encouraged to be similar, by minimising their pixel-wise variance, and secondly, the latent representations are individually decoded by g, and these results are also encouraged to be accurate by minimising the MAE. Useful unique information is therefore maintained by minimising the synthesis error of both the fused and individual latent representations. The representations are also implicitly made similar by sharing the same decoder g. Our final cost is a sum of these three components, where $\mathbf{Y}$ is the target image, MAE is the mean absolute error, and MVV is the mean voxel-wise variance:

$$c(\theta, \psi, \mu) = MAE(g(\mathbf{Z}_\alpha|\psi), \mathbf{Y}) + MVV(\mathbf{Z}_1^*, ..., \mathbf{Z}_n^*) + \sum_{i=1}^{n} MAE(g(\mathbf{Z}_i^*|\psi), \mathbf{Y}).$$

Table 1. The test MSE (± s.d.) for unimodal models on ISLES and BRATS datasets.

	T1 → *T2*		*T1*→ *FLAIR*	
	ISLES	BRATS	ISLES	BRATS
DEDIS [15]	2.265 ±0.88	7.289 ±7.80	1.956 ±0.78	7.589 ±8.08
LSDN [17]	1.282 ±0.68	3.259 ±0.85	0.961 ±0.39	2.446 ±1.21
RF [9]	1.153 ±0.56	3.407 ±1.15	0.986 ±0.41	2.382 ±1.32
Proposed	**1.018** ±0.47	**2.213** ±0.70	**0.822** ±0.34	**1.689** ±1.02

5 Experiments

Here we detail experiments that examine the model's accuracy, robustness to missing and misaligned data, and propensity for transfer learning. We train all models using Adam [10] with a batch-size of 32, and all results are from 5 fold cross validation. Models were implemented in Python with Keras.

Datasets: We use the SISS dataset from the *Ischemic Stroke Lesion Segmentation* (ISLES) 2015 challenge[1]. This contains 28 volumes in T1w, T2w, Flair and DWI contrast sequences that have been skull-stripped and resampled to an isotropic spacing of $1mm^3$ and co-registered to the FLAIR contrast sequence. We also use the low grade glioma cases from the multimodal *Brain Tumor Segmentation* (BRATS) 2015 challenge[2]. This data contains 54 volumes imaged in T1w, T1c, T2w and Flair and are skull striped, co-aligned, and interpolated to $1mm^3$ resolution. Our architecture uses 2D axial-plane slices of the volumes.

Preprocessing: We trim and downsample the data to create volumes of 112×80 pixel images for the ISLES dataset and 128×128 for the BRATS dataset. We thus remove uninformative background areas of the image and reduce the image size such that it is divisible by 4, so that the 2×2 max-pooling and upsampling operations of the encoders do not change the final image size. Finally, we normalise by dividing each volume by its mean in order to shift its mean to 1 and also maintain positive values for all pixel values. This also maintains small variance differences between volumes and keeps background pixels at 0.

Error Metric: We use the mean squared error (MSE) calculated over all non-background voxels in the volume.

5.1 Synthesis Accuracy

We compare the accuracy of our model to the accuracy of a number of state of the art models. We compare against LSDN [17], Deep Encoder-Decoder Image Synthesis [15] (DEDIS), and a recent random forest approach [9] (RF). The results are shown in Table 1, and show that when used as a single input model our approach achieves a $\sim 14\%$ increase in performance on average for ISLES tasks and $\sim 30\%$ for BRATS tasks, compared to the second best method.

[1] http://www.isles-challenge.org/ISLES2015/

[2] https://sites.google.com/site/braintumorsegmentation/home/brats2015

Table 2. Comparison of multi-input results in synthesising FLAIR.

	T1,T2,DWI:	✓,×,×	×,✓,×	×,×,✓	✓,✓,×	✓,×,✓	×,✓,✓	✓,✓,✓
(A)	RF [9]	0.986	1.295	0.901	0.917	0.712	0.789	0.670
	Proposed (different models)	0.822	1.051	**0.821**	0.717	0.661	**0.672**	**0.576**
(B)	Proposed (missing inputs)	**0.799**	**1.048**	0.842	**0.699**	**0.641**	0.686	**0.576**

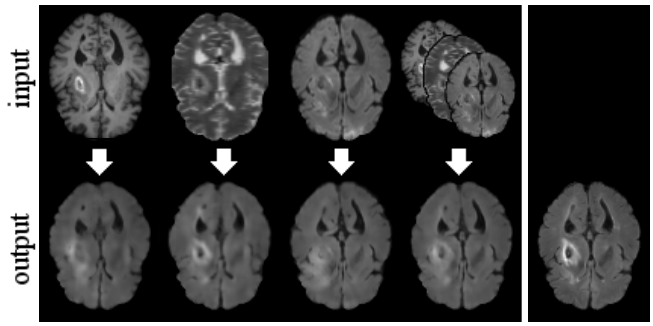

Fig. 4. Example of our model's multi-modal synthesis (test case). The first column shows T1 alone synthesising FLAIR, similarly, the second and third columns show the results of using just T2 and DWI respectively. The fourth column shows the synthesis output of all inputs. The final column shows the ground truth FLAIR. Observe that, despite the presence of a lesion, our model is still able to achieve accurate synthesis.

5.2 Multi-Input Synthesis, and Robustness to Missing Inputs

To test robustness to missing inputs, we first trained 7 separate models, one for each combination of T1, T2 and DWI inputs for both our model and the RF (see Table 2 (A)). Next we took our model trained with all three inputs and tested it with all 7 possible combinations of missing inputs (see Table 2 (B)). It can be seen that our proposed method outperforms RF on all input combinations, improving on average by $\sim 16\%$. Further, our full model performs just as well with missing inputs as a model trained specifically for the limited input case, showing robustness to missing data. Example images are shown in Fig. 4.

5.3 Robustness to Data Misalignment

To examine the performance of our model on unaligned data we trained and tested a model for synthesising FLAIR from T1 and DWI on data in which each T1 volume was randomly rotated about all axes by a number of degrees sampled uniformly at random from [-8,8] and was shifted randomly on each axis by a (not necessarily integer) number of pixels from [-2,2]. This produced data with misalignment between modalities of the sort that remains after a simple alignment procedure had been performed. When trained on aligned data, our model and RF achieve MSE of 0.661 and 0.712 respectively, which increases

to 0.793 and 0.885 respectively on the unaligned task. However, compared to the unimodal case where only DWI is given as input, which achieves MSEs of 0.821 and 0.901, we observe an improvement of 6% and 2% for our model and RF respectively. Although seemingly a small improvement, rotating and shifting across the z-axis changes the anatomy present in the image, necessarily resulting in performance degradation, and additionally information is lost by blurring during rotation. However, our model is still able to capture the limited information of the distorted T1 input to improve on the unimodal result.

5.4 Transfer Learning

Here we examine the model's ability to generalise to MRI data with different intensity characteristics, not seen during training. We use a model synthesising T2 from T1 trained on BRATS, and test it on ISLES volumes. We first use the model as-is without any fine tuning and get a MSE of 3.990 which we use as a baseline. Then, based on the assumption that the deeper layers of the network are task specific [19], we fine-tune just the decoder using 1, 2 and 3 volumes and get a MSE of 1.439, 1.356 and 1.227 respectively. The MSE decreases significantly, improving 70% compared to the baseline and approaches the MSE obtained from the ISLES trained model of Table 2, which is 1.018. In addition, fine tuning the decoder is extremely fast, taking ~ 4 minutes on one Titan X GPU.

6 Conclusion

We have demonstrated that a multi-input neural network model is able to outperform other recently proposed synthesis methods in both the single input and multi-input settings. In the multi-input case the final model is robust to missing inputs, as it is able to create accurate synthetic images from any subset of input images. In fact, when given only a subset of inputs it performs as well as a model trained specifically for that subset. In addition to being robust to missing data, the model is also robust to misaligned inputs. In particular, it is able to benefit from multiple inputs, even when they are not well aligned, such that it still outperforms the single input case, even though misalignment means that the slice may well contain different anatomy. Finally, we demonstrated that fine-tuning only the network's decoder on a very small number of volumes allowed the model to synthesise volumes from an otherwise unseen source with high accuracy.

Acknowledgements This work was supported in part by the US National Institutes of Health (2R01HL091989-05) and UK EPSRC (EP/P022928/1). We thank NVIDIA for donating a Titan X GPU.

1. Bowles, C., Qin, C., Ledig, C., Guerrero, R., Gunn, R., Hammers, A., Sakka, E., Dickie, D.A., Hernández, M.V., Royle, N.: Pseudo-healthy Image Synthesis for White Matter Lesion Segmentation. In: SASHIMI, pp. 87–96 (2016)

2. Burgos, N., Cardoso, M.J., Thielemans, K., Modat, M., Pedemonte, S., Dickson, J., Barnes, A., Ahmed, R., Mahoney, C.J., Schott, J.M.: Attenuation correction synthesis for hybrid PET-MR scanners: application to brain studies. IEEE TMI 33(12), 2332–2341 (2014)
3. Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., and Pal, C.: The importance of skip connections in biomedical image segmentation. In: LABELS, pp. 179–187 (2016)
4. Huang, Y., Beltrachini, L., Shao, L., and Frangi, A.F.: Geometry Regularized Joint Dictionary Learning for Cross-Modality Image Synthesis in Magnetic Resonance Imaging. In: SASHIMI, pp. 118–126 (2016)
5. Huynh, T., Gao, Y., Kang, J., Wang, L., Zhang, P., Lian, J., and Shen, D.: Estimating CT image from MRI data using structured random forest and auto-context model. IEEE TMI 35(1), 174–183 (2016)
6. Iglesias, J.E., Konukoglu, E., Zikic, D., Glocker, B., Van Leemput, K., and Fischl, B.: Is synthesizing MRI contrast useful for inter-modality analysis? In: MICCAI, pp. 631–638 (2013)
7. Jaderberg, M., Simonyan, K., Zisserman, A.: Spatial transformer networks. In: NIPS, pp. 2017–2025 (2015)
8. Jog, A., Carass, A., Roy, S., Pham, D.L., and Prince, J.L.: MR image synthesis by contrast learning on neighborhood ensembles. MIA 24(1), 63–76 (2015)
9. Jog, A., Carass, A., Roy, S., Pham, D.L., and Prince, J.L.: Random forest regression for magnetic resonance image synthesis. MIA 35, 475–488 (2017)
10. Kingma, D., and Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv: 1412.6980 (2014)
11. Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. MICCAI (2015)
12. Roy, S., Chou, Y.-Y., Jog, A., Butman, J.A., and Pham, D.L.: Patch Based Synthesis of Whole Head MR Images: Application To EPI Distortion Correction. In: SASHIMI, pp. 146–156 (2016)
13. Roy, S., Wang, W.-T., Carass, A., Prince, J.L., Butman, J.A., and Pham, D.L.: PET attenuation correction using synthetic CT from ultrashort echo-time MR imaging. Journal of Nuclear Medicine 55(12), 2071–2077 (2014)
14. Roy, S., Carass, A., and Prince, J.L.: Magnetic resonance image example-based contrast synthesis. IEEE TMI 32(12), 2348–2363 (2013)
15. Sevetlidis, V., Giuffrida, M.V., and Tsaftaris, S.A.: Whole Image Synthesis Using a Deep Encoder-Decoder Network. In: SASHIMI, pp. 97–107 (2016)
16. Tulder, G. van, and Bruijne, M. de: Why does synthesized data improve multi-sequence classification? In: MICCAI, pp. 531–538 (2015)
17. Van Nguyen, H., Zhou, K., and Vemulapalli, R.: Cross-domain synthesis of medical images using efficient location-sensitive deep network. In: MICCAI, pp. 677–684 (2015)
18. Ye, D.H., Zikic, D., Glocker, B., Criminisi, A., and Konukoglu, E.: Modality propagation: coherent synthesis of subject-specific scans with data-driven regularization. In: MICCAI, pp. 606–613 (2013)
19. Yosinski, J., Clune, J., Bengio, Y., and Lipson, H.: How transferable are features in deep neural networks? In: NIPS, pp. 3320–3328 (2014)