# Uncertainty-aware multi-resolution whole-body MR to CT synthesis

Kerstin Kläser[1,2], Pedro Borges[1,2], Richard Shaw[1,2], Marta Ranzini[1,2], Marc Modat[2], David Atkinson[4], Kris Thielemans[3], Brian Hutton[3], Vicky Goh[2], Gary Cook[2], M Jorge Cardoso[2], and Sébastien Ourselin[2]

[1]Dept. Medical Physics & Biomedical Engineering, University College London, UK
[2]School of Biomedical Engineering & Imaging Sciences, King's College London, UK
[3]Institute of Nuclear Medicine, University College London, UK
[4]Centre for Medical Imaging, University College London, UK

**Abstract.** Synthesising computed tomography (CT) images from magnetic resonance images (MRI) plays an important role in the field of medical image analysis, both for quantification and diagnostic purposes. Especially for brain applications, convolutional neural networks (CNNs) have proven to be a valuable tool in this image translation task, achieving state-of-the-art results. Full body image synthesis, however, remains largely uncharted territory, bearing many challenges including a limited field of view and large image size, complex spatial context and anatomical differences between time-elapsing image acquisitions. We propose a novel multi-resolution cascade 3D network for end-to-end full-body MR to CT synthesis. We show that our method outperforms popular CNNs like U-Net in 2D and 3D. We further propose to include uncertainty in our network as a measure of safety and to account for intrinsic noise and misalignment in the data.

**Keywords:** MR to CT synthesis · Multi-resolution CNN · Uncertainty.

## 1 Introduction

Simultaneous positron emission tomography and magnetic resonance imaging (PET/MRI) is an important tool in both clinical and research applications that allows for a multiparametric evaluation of an individual. It combines the high soft-tissue contrast from MRI with radiotracer uptake distribution information obtained from PET imaging. To accurately reconstruct PET images, it is essential to correct for photon attenuation throughout the patient. A multi-center study on brain images has shown that obtaining tissue attenuation coefficients from synthesised computed tomography (CT) images leads to state-of-the-art results for PET/MRI attenuation correction [12]. In recent years, the field of MR to CT synthesis has shifted towards the use of convolutional neural networks (CNNs) that have proved to be a powerful tool in the MR to CT image translation task, outperforming existing multi-atlas-based methods [11, 17]. However, the problem full-body MR to CT synthesis has largely remained untackled. In

2019, Ge et al. [5] attempted to translate full-body MR images to CT images by introducing a multi-view adversarial learning scheme that predicts 2D pseudo-CT (pCT) images along three axes (i.e. axial, coronal, sagittal). 3D volumes are obtained for each axis by stacking 2D slices together before an average fusion is performed to obtain one final 3D volume. The synthesis performance is then evaluated on sub-regions of the body (lungs, femur bones, spine etc). They do not, however, provide results on the full volume. We propose a novel learning scheme for uncertainty aware multi-resolution MR to CT synthesis of the full body (MultiRes). Multi-resolution learning has been used for many computer vision tasks such as dynamic scene deblurring [15], optical flow prediction [3] and depth map estimation [4]. In the field of medical imaging, multi-resolution learning is a popular method for image classification [9] and segmentation [8]. These methods learn strong features at multiple levels of scale and abstraction, therefore finding the input/output voxel correspondence based on these features. Due to the large image size of full-body acquisitions and physical GPU memory constraints, high-resolution 3D image synthesis networks can only be trained in a patch-wise manner, thus capturing a limited amount of spatial context. We show that incorporating feature maps learned at multiple resolutions results in significantly better pCT images than using high-resolution images alone. As a means of providing a measure of algorithm safety, and to account for the limited number of training samples, we also model uncertainty [16]. It is important to distinguish between two types of uncertainty: *aleatoric* and *epistemic* uncertainty. Aleatoric uncertainty captures the irreducible variance that exists in the data, whereas epistemic uncertainty accounts for the uncertainty in the model [10]. Aleatoric uncertainty can be further subcategorized into *homoscedastic* and *heteroscedastic*. Homoscedastic uncertainty is constant across all input data, while heteroscedastic uncertainty varies across the input data. It is evident that in our setting the aleatoric uncertainty should be modelled as heteroscedastic, as task performance is expected to vary spatially due to the presence of artefacts, tissue boundaries, small structures etc. By training our network with channel dropout we can stochastically sample from the approximate posterior over the network weights to obtain epistemic uncertainty measures. By explicitly modelling for the intrinsic noise in the data via modifications to our network architecture and loss function we can observe the heteroscedastic uncertainty. The network is encouraged to assign high levels of uncertainty to high error regions, providing a means of understanding what aspects of the data pose the greatest challenges.

## 2 Methods

The main challenge with whole body data is its size, and the fact that a large field of view is necessary to make accurate predictions. Common networks, such as a U-Net [2], can only store patches of size $160^3$ due to GPU memory limitations. This small field of view causes significant issues as it will be demonstrated later in the experiments section. To tackle this issue we propose an end-to-end multi-scale convolutional neural network that takes input patches from full-body

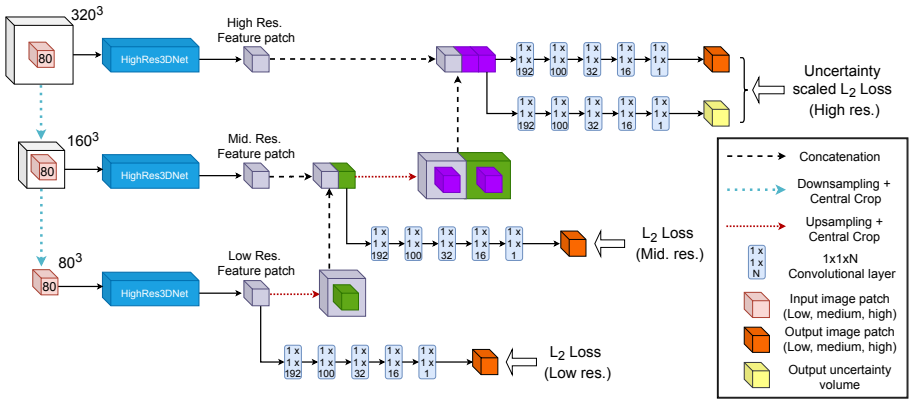


**Fig. 1.** Proposed MultiRes network architecture. An initial $T_1$ MR patch of size $320^3$ is fed into each instance of the HighRes3DNet architecture [13] at various levels of resolution and field of view. Lower level feature maps are concatenated to those at the next level until the full resolution level, where these concatenated feature maps are passed through two branches consisting of a series of $1 \times 1 \times N$ convolutional layers: one resulting in a synthesised CT patch and the other to the corresponding voxel-wise heteroscedastic uncertainty.

MR images at three resolution levels to synthesise high resolution, realistic CT patches. The network also incorporates explicit heteroscedastic uncertainty modelling by casting our task likelihood probabilistically, and epistemic uncertainty estimation via traditional Monte Carlo dropout. We employ a patch-based training approach whereby at each resolution level of the framework a combination of downsampling and cropping operations results in patches of similar size but at different resolutions, spanning varied fields of view. Three independent instances of HighRes3DNet are trained simultaneously, thus not sharing weights, taking patches of each resolution as input each resulting in a feature map with different resolution. Lower level feature maps are concatenated to those at the next level of resolution until the full resolution level, where these concatenated feature maps are passed through two branches of $1 \times 1 \times N$ convolutional layers resulting in a synthesised CT patch and the the corresponding voxel-wise heteroscedastic uncertainty. This is illustrated in Fig. 1. We posit, similarly to [8], that such a design allows the network to simultaneously benefit from the fine details afforded by the highest resolution patch and the increased spatial context provided by the higher field of view patches. However, we incorporate an additional level of deep supervision that [8] misses.

### 2.1 Modelling heteroscedastic uncertainty

Previous works on MR to CT synthesis have shown that residual errors are not homogeneously spread throughout the image, rather, they are largely concentrated around organ/tissue boundaries. As such, a heteroscedastic uncertainty

model is most suitable for this task, where data-dependent, or intrinsic, uncertainty is assumed to be variable. We begin by modelling our task likelihood as a normal distribution with mean $\boldsymbol{f^W(x)}$, the model output corresponding to the input $\mathbf{x}$, parameterised by weights $\mathbf{W}$, and voxel-wise standard deviation $\boldsymbol{\sigma^W}(\boldsymbol{x})$, the data intrinsic noise:

$$p(\boldsymbol{y}|\boldsymbol{f^W}(\boldsymbol{x})) = \mathcal{N}(\boldsymbol{f^W}(\boldsymbol{x}), \boldsymbol{\sigma^W}(\boldsymbol{x})) \tag{1}$$

Our loss function is derived by calculating the negative log of the likelihood:

$$\begin{aligned} \mathcal{L}(\mathbf{y}, \boldsymbol{x}; \mathbf{W}) &= -\log p(\boldsymbol{y}|\boldsymbol{f^W}(\boldsymbol{x})) \\ &\approx \frac{1}{2\boldsymbol{\sigma^W}(\boldsymbol{x})^2}\Big(\boldsymbol{y} - \boldsymbol{f^W}(\boldsymbol{x})\Big)^2 + \log \boldsymbol{\sigma^W}(\boldsymbol{x}) \\ &= \frac{1}{2\boldsymbol{\sigma^W}(\boldsymbol{x})}\mathcal{L}_2(\boldsymbol{y}, \boldsymbol{f^W}(\boldsymbol{x})) + \log \boldsymbol{\sigma^W}(\boldsymbol{x}) \end{aligned} \tag{2}$$

In those regions where the observed $\mathcal{L}_2$ error remains high, the uncertainty should compensate and also increase. The second term in the loss prevents the collapse to the trivial solution of assigning a large uncertainty everywhere.

### 2.2 Modelling epistemic uncertainty

Test-time dropout has been established as the go-to method for estimating model uncertainty, a Bayesian approximation at inference. By employing dropout during training and testing we can sample from a distribution of sub-nets that in the regime of data scarcity will provide varying predictions. This variability captures the uncertainty present in the network's parameters, allowing for a voxel-wise estimation by quantifying the variance across these samples. In this work, channel dropout was chosen over the traditional neuron dropout. Channel dropout has indeed been shown to be better for convolutional layers where channels fully encode image features while neurons do not encode individually such meaningful information [7]. Dropout samples at inference time are acquired by performing $N$ stochastic forward passes over the network, equivalent to sampling from the posterior over the weights. A measure of uncertainty can be obtained by calculating the variance over these samples on a voxel-wise basis.

### 2.3 Implementation details

Experiments were implemented and carried out using NiftyNet, a TensorFlow based deep learning framework tailored for medical imaging [6], and code will be made available on publication. The multi-scale network consists of three residual networks, each taking in an $80\times80\times80$ MR image patch with different resolutions and fields of view. In order of high, medium, and low resolution, the MR patches are obtained by taking an initial high resolution $320 \times 320 \times 320$ patch and

cropping the central $80 \times 80 \times 80$ region (high), downsampling the initial patch by a factor of two and taking the central $80 \times 80 \times 80$ patch (medium), and finally downsampling the initial patch by a factor of four to obtain a $80 \times 80 \times 80$ patch (low).

Starting from the lowest resolution sub-net, the output of size $80 \times 80 \times 80$ is upsampled by a factor of two and centrally cropped. This patch is concatenated with the output of the medium resolution sub-net. This concatenated patch of size $80 \times 80 \times 80 \times 2$ is then upsampled by a factor of two and centrally cropped, before being concatenated to the output of the high-resolution sub-net. These series of upsamplings and crops ensure that the final outputs contain patches with the same field of view prior to the final set of four 3D convolutions of kernel size $1 \times 1 \times 1$, which produces the CT patch.

Heteroscedastic variance is modelled by the addition of a series of four $1 \times 1 \times 1$ convolutional layers following the concatenation of the combined low-medium scale output to the high scale output, architecturally identical to the convolutional layers for the synthesis branch. Channel dropout probability (i.e.: The probability to keep any one channel in a kernel) was set to 0.5, both during training and testing, and $N$=20 forward passes were carried out for each experiment. The batch size was set to one, ADAM was used as the optimiser and networks were trained until convergence, where this was defined as a sub 5% loss change over a period of 5000 iterations.

## 3 Experiments and Results

### 3.1 Data

The dataset used for training and cross-validation consisted of 32 pairs of whole-body MR (voxel size $0.67 \times 0.67 \times 5$ mm$^3$) and CT images (voxel size $1.37 \times 1.37 \times 3.27$ mm$^3$). Whole-body MR images were acquired in four stages. MR pre-processing included bias field correction followed by fusion between stages using a percentile-based intensity harmonisation. All images were resampled to CT resolution. MR and CT images were aligned using first a rigid registration algorithm followed by a very-low-degree-of-freedom non-rigid deformation. A second non-linear registration was performed, using a cubic B-spline with normalised mutual information to correct for soft tissue shift [1][14]. Both CT and MR images were rescaled to be between 0 and 1 for increased training stability.

### 3.2 Experiments

In addition to our proposed method, we compare results quantitatively and qualitatively against two baselines: U-Net trained with 2D, $224 \times 224 \times 1$, patches with batch size one, and U-Net trained with 3D, $160 \times 160 \times 160$, patches with batch size one. An additional four convolutional layers with kernel size three were added prior to the final $1 \times 1 \times 1$ convolutional layer in the standard architecture as this was found to increase stability during training. All models were trained on the same 22 images while the remaining 10 images were equally split into validation and testing data.

**Table 1.** MAE and MSE across all experiments including number of trainable variables. Bolded entries denote best model (p-value $< 0.05$).

| Experiments | Model parameters | MAE (HU) | MSE ($HU^2$) |
|---|---|---|---|
| 2D U-Net (No Unc) | 4.84M | 112.94 ± 16.04 | 32081.18 ± 5667.11 |
| 3D U-Net (No Unc) | 14.49M | 99.87 ± 14.17 | 23217.57 ± 3515.50 |
| MultiRes | 2.54M | **62.42 ± 6.8** | **11347.16 ± 3089.12** |
| $\text{MultiRes}_{unc}$ | 2.61M | 80.14 ± 15.81 | 14113.83 ± 3668.79 |

### 3.3 Results

**Quantitative evaluation** The quantitative evaluation consists of a Mean Squared Error ($MSE = \frac{\sum(pCT-CT)^2}{V}$, with $V$ being the total number of non-zero voxels) and Mean Absolute Error ($MAE = \frac{\sum|pCT-CT|}{V}$) analysis between the network outputs and the ground truth CT. We observe that the proposed method without uncertainty performs the best, exhibiting the lowest MAE and MSE averaged across all inference subjects. A paired t-test was performed to show that the results are significantly better (p-value $< 0.05$). Furthermore, the proposed MultiRes networks show a better performance while decreasing the model size making the networks much more efficient than the U-Net models.

**Qualitative evaluation** Fig. 2 shows the ground truth CT and the pCT predictions generated with 2D U-Net, 3D U-Net, proposed MultiRes and proposed $\text{MultiRes}_{unc}$ with uncertainty and the subject's MR image as well as the models' corresponding residuals. 2D U-Net clearly cannot capture bone; likely because it lacks the spatial context necessary to construct small (relatively) cohesive structures such as vertebrae. The lungs, having a significantly larger cross-sectional area, are visible, but lack internal consistency. 3D U-Net's bone synthesis is more faithful than its 2D counterpart but is characterised by a large degree of blurriness most evident in the femurs. The proposed MultiRes model exhibits the greatest bone fidelity; the individual vertebrae are more clear, with intensities more in line with what would be expected for such tissues, and the femurs boast more well-defined borders. The proposed $\text{MultiRes}_{unc}$ model leads to similar results than the simpler proposed MultiRes model without uncertainty. However, bones are slightly blurrier, likely due to the inclusion of uncertainty term and limited network capacity, but still demonstrates superior bone reconstruction than both U-Net models.

The benefits afforded to $\text{MultiRes}_{unc}$ for being uncertainty-aware are showcased in Fig. 3. The joint histograms (a) and (b) are constructed by calculating the error rate, taken as the difference between the ground truth CT and pCT averaged across $N$=20 dropout samples, at different levels of both epistemic and heteroscedastic uncertainty (standard deviations per voxel) and taking the base 10 log. The red line describes the average error rate at each level of uncertainty. We observe a significant correlation between uncertainty and error rate, suggesting that the model appropriately assigns a higher uncertainty to those

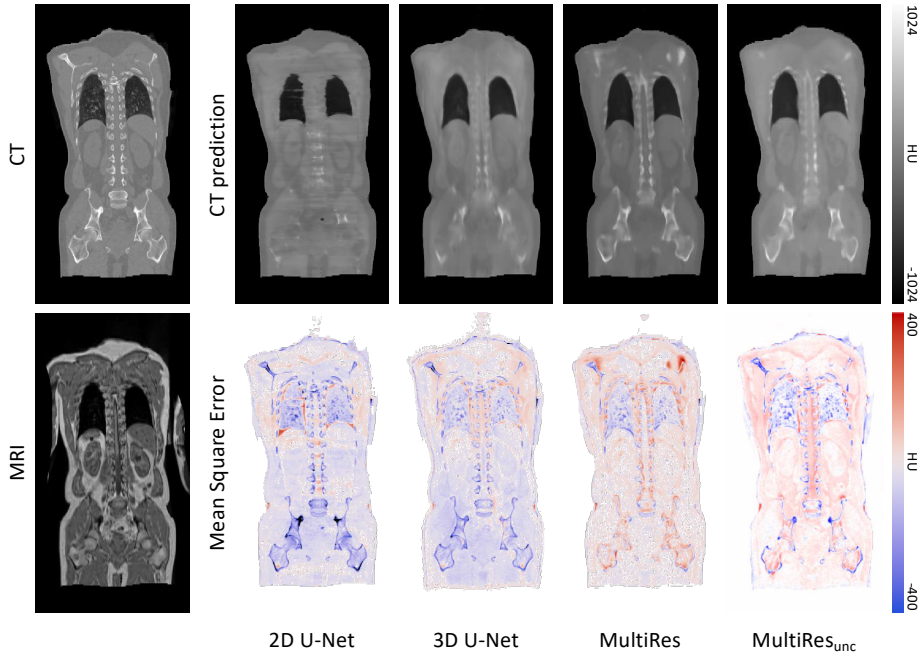


**Fig. 2.** Top: CT and pCT prediction of 2D U-Net, 3D U-Net, proposed MultiRes and proposed uncertainty aware MultiRes$_{unc}$. Bottom row: MR image and model residuals.

regions that are difficult to predict. This correlation is likewise observed when comparing the maps of uncertainty (epistemic: (c), heteroscedastic: (e)) with the corresponding absolute error map (d). Both epistemic and heteroscedastic uncertainties exhibit large values around structure borders, as expected. The borders between tissues are not sharp and there is, therefore, some ambiguity in these regions, which is mirrored by the corresponding overlapping error in the residuals. An increased amount of data should diminish the epistemic uncertainty by providing the network with a greater number of samples from which to learn the correspondence between MR and CT in these areas. The aforementioned blurriness, however, could result in some inconsistency in the synthesis process, which would be captured by the heteroscedastic uncertainty.

Of note is the high degree of uncertainty we observed in the vicinity of air pockets. Unlike corporeal structures, it is expected that these pockets are subject to deformation between the MR and CT scanning sessions, resulting in a lack of correspondence between the acquisitions in these regions. This results in the network attempting to synthesise a morphologically different pocket to what is observed in the MR, resulting in a high degree of uncertainty.

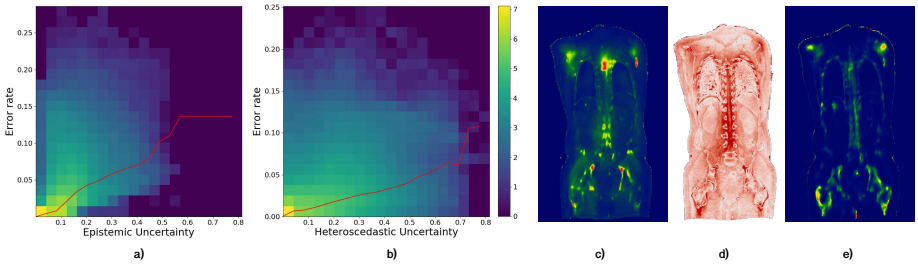


**Fig. 3.** Joint histogram of prediction uncertainty and error rate for proposed MultiRes$_{unc}$ network: a) Epistemic b) Heteroscedastic. The average error rate at different uncertainty levels is shown by the red line. Error rate tends to increase with increasing uncertainty, showing that the network correlates uncertainty to regions of error. c) Epistemic uncertainty and e) heteroscedastic uncertainty correlate with d) the MAE of the prediction error [0HU, 800HU], solidifying this point.

## 4 Discussion and Conclusions

Our contributions in this work are two-fold: MultiRes, a novel learning scheme for uncertainty aware multi-resolution MR to CT synthesis of the full body, and MultiRes$_{unc}$, a version of this model that incorporates uncertainty as a safety measure and to account for intrinsic data noise. We demonstrate the significantly superior performance (p-value $< 0.05$) of MultiRes and MultiRes$_{unc}$ by comparing it to single-resolution CNNs, 2D and 3D U-Net, and the importance of modelling uncertainty, showing that MultiRes$_{unc}$ is able to identify regions where the MR to CT translation is most difficult.

In a data-scarce environment, it becomes especially important to quantify uncertainty as networks are unlikely to have have sufficient evidence for full convergence.

After all, accurately aligning CT and MR images is inevitable to validate the voxel-wise performance of any image synthesis algorithm until other appropriate methods have been developed that allow validating on non-registered data.

Despite the slightly decreased performance of MultiRes$_{unc}$ compared to MultiRes, both from a quantitative and qualitative standpoint, we posit that the additional insight introduced by modelling uncertainty can compensate for this. Furthermore, while the model does not reconstruct bone-based structures as well as its uncertainty agnostic counterpart, it still outperforms both U-Net models qualitatively and quantitatively.

To summarise, we design a multi-scale/resolution network for MR to CT synthesis, showing that it outperforms single-resolution 2D and 3D alternatives. Furthermore, by incorporating epistemic uncertainty via test time dropout, and heteroscedastic uncertainty by casting the model probabilistically, we can showcase those regions that exhibit the greatest variability, providing a measure of safety from an algorithmic standpoint. We demonstrate that these regions correlate well with the residuals obtained by comparing the outputs with the ground

truth, lending further credence to the usefulness of uncertainty's inclusion. We argue that the slight decrease in performance of the uncertainty aware model is insignificant compared to the important additional information provided by the uncertainty.

## References


1. Burgos, N., Cardoso, M.J., Thielemans, K., Modat, M., Pedemonte, S., Dickson, J., Barnes, A., Ahmed, R., Mahoney, C.J., Schott, J.M., et al.: Attenuation correction synthesis for hybrid pet-mr scanners: application to brain studies. IEEE transactions on medical imaging **33**(12), 2332–2341 (2014)
2. Çiçek, Ö., , et al.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016)
3. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2758–2766 (2015)
4. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in neural information processing systems. pp. 2366–2374 (2014)
5. Ge, Y., Xue, Z., Cao, T., Liao, S.: Unpaired whole-body mr to ct synthesis with correlation coefficient constrained adversarial learning. In: Medical Imaging 2019: Image Processing. vol. 10949, p. 1094905. International Society for Optics and Photonics (2019)
6. Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D.I., Wang, G., Eaton-Rosen, Z., Gray, R., Doel, T., Hu, Y., et al.: Niftynet: a deep-learning platform for medical imaging. Computer methods and programs in biomedicine **158**, 113–122 (2018)
7. Hou, S., Wang, Z.: Weighted channel dropout for regularization of deep convolutional neural network. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8425–8432 (2019)
8. Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B.: Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Medical image analysis **36**, 61–78 (2017)
9. Kawahara, J., Hamarneh, G.: Multi-resolution-tract cnn with hybrid pretrained and skin-lesion trained layers. In: International workshop on machine learning in medical imaging. pp. 164–171. Springer (2016)
10. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Advances in neural information processing systems. pp. 5574–5584 (2017)
11. Kläser, K., Markiewicz, P., Ranzini, M., Li, W., Modat, M., Hutton, B.F., Atkinson, D., Thielemans, K., Cardoso, M.J., Ourselin, S.: Deep boosted regression for mr to ct synthesis. In: International Workshop on Simulation and Synthesis in Medical Imaging. pp. 61–70. Springer (2018)
12. Ladefoged, C.N., Law, I., Anazodo, U., Lawrence, K.S., Izquierdo-Garcia, D., Catana, C., Burgos, N., Cardoso, M.J., Ourselin, S., Hutton, B., et al.: A multi-centre evaluation of eleven clinically feasible brain pet/mri attenuation correction techniques using a large cohort of patients. Neuroimage **147**, 346–359 (2017)

13. Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M.J., Vercauteren, T.: On the compactness, efficiency, and representation of 3d convolutional networks: brain parcellation as a pretext task. In: International conference on information processing in medical imaging. pp. 348–360. Springer (2017)
14. Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S.: Fast free-form deformation using graphics processing units. Computer methods and programs in biomedicine **98**(3), 278–284 (2010)
15. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3883–3891 (2017)
16. Reinhold, J.C., He, Y., Han, S., Chen, Y., Gao, D., Lee, J., Prince, J.L., Carass, A.: Validating uncertainty in medical image translation. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). pp. 95–98. IEEE (2020)
17. Wolterink, J.M., Dinkla, A.M., Savenije, M.H., Seevinck, P.R., van den Berg, C.A., Išgum, I.: Deep mr to ct synthesis using unpaired data. In: International workshop on simulation and synthesis in medical imaging. pp. 14–23. Springer (2017)