Deep learning for MRI-based CT synthesis: a comparison of MRI sequences and neural network architectures

Andrés Larroza, Laura Moliner, Juan M. Álvarez-Gómez, Sandra Oliver, Héctor Espinós-Morató,

Marina Vergara-Díaz, María J. Rodríguez-Álvarez

***Abstract*— Synthetic computed tomography (CT) images derived from magnetic resonance images (MRI) are of interest for radiotherapy planning and positron emission tomography (PET) attenuation correction. In recent years, deep learning implementations have demonstrated improvement over atlas-based and segmentation-based methods. Nevertheless, several open questions remain to be addressed, such as which are the best MRI sequence and neural network architecture. In this work, we compared the performance of different combinations of two common MRI sequences (T1- and T2-weighted), and three state-of-the-art neural networks designed for medical image processing (*Vnet*, *HighRes3dNet* and *ScaleNet*). The experiments were conducted on brain datasets from a public database. Our results suggest that T1 images performs better than T2, but the results further improve when combining both sequences. The lowest mean average error over the entire head (MAE = 95.37 ± 11.70 HU) was achieved combining T1 and T2 scans with *ScaleNet*. All tested deep learning models achieved significantly lower MAE ($p < 0.05$) than a well-known atlas-based method.**

I. Introduction

Computed tomography (CT) images provide Hounsfield units (HU) as a measure of tissue attenuation, which is essential for dose calculation in radiotherapy planning and for positron emission tomography (PET) attenuation correction in popular hybrid PET-CT scanners [1]. Synthetic CT derived from magnetic resonance images (MRI) gained special interest in the past years due to unavailability of CT images as a result of the introduction of hybrid PET-MRI scanners [2]. The interest in MRI-only planning for radiotherapy treatment resides in the excellent soft tissue contrast and the lower exposure to imaging dose for patients [3].

Two clinical PET-MRI systems are commercially available: Biograph mMR (Siemens Healthcare GmbH, Erlangen, Germnay) and SIGNA (GE Healthcare, Waukesha WI, USA). The vendor-implemented CT synthesis methods for these scanners are segmentation- and/or atlas-based. Segmentation-based methods assign a predefined attenuation coefficient to different MRI tissues. However, the bone is especially difficult to visualize in standard MRI sequences. Most segmentation methods usually require specific MRI acquisitions such as Dixon and ultrashort echo time (UTE) to enhance fat/water and bone tissues respectively. Atlas-based approaches do not require specific MRI sequences. They usually rely on a previously acquired atlas of MRI and CT pairs and use that information to estimate the attenuation by registering one or more images from the atlas to the new MRI [4].

In recent years, several studies demonstrated the potential of deep learning approaches for CT synthesis [5]–[8]. Deep learning approaches are basically a regression implementation where the neural network aims to find a mapping from the domain of MRI input images to the domain of CT images. Some open questions remain to be addressed: which MRI sequence provide the best synthetic CT, and what neural network architecture is more suitable? To answer the above questions, we compared the performance of different combinations of two standard MRI sequences (T1- and T2-weighted), and three state-of-the-art neural networks designed for medical image processing (*Vnet*, *Highres3dNet* and *ScaleNet*).

II. Materials and Methods

A. Imaging Data

One of the major challenges in deep learning for medical imaging is the scarce availability of training data. This is usually solved using transfer learning, where models that were trained for different tasks are fine-tuned for the application of interest. In our experiments, we decided to train all networks from scratch, using the public dataset RIRE (Retrospective Image Registration Evaluation Project, www.insight-journal.org/rire). Whilst this dataset is very small (17 brain subjects), it is useful for the comparisons performed in this study and can be used as a benchmark for future research. The following imaging sequences were used in this study:

- **CT:** The CT volumes were acquired with a Siemens DR-H scanner, with a voxel size of 0.45 × 0.45 × 3 mm.
- **mrT1:** T1-weighted spin-echo sequences were acquired with a Siemens SP 1.5 T scanner. Echo time (TE) of 15 ms and repetition time (TR) of 800 ms, with a voxel size of 0.86 × 0.86 × 3 mm.

This work was supported by

the Spanish Government grants TEC2016-79884-C2 and RTC-2016-5186-1, and by the European Union through the European Regional Development Fund (ERDF).

A. Larroza (anlarro@gmail.com) and co-authors are with Instituto de Instrumentación para Imagen Molecular (I3M). Universitat Politècnica de València (UPV)-Consejo Superior de Investigaciones Científicas (CSIC),Camino de Vera s/n, 46022 Valencia, Spain.

- **mrT2:** T2-weighted spin-echo sequences acquired with a Siemens SP 1.5 T scanner. TE of 90 ms and TR of 3000 ms, with a voxel size of 0.86 × 0.86 × 3 mm.

For each subject, MRIs and CTs were rigidly aligned using the Fast Automatic Step Size Estimation for Gradient Descent Optimization (FASGD) method [9] in *elastix* [10].

After alignment, a head label was obtained for each patient by thresholding the mrT1 image, followed by a morphological closing operator to fill the gaps in the nasal cavities and ear canals. CT and MRI volumes were then masked with the head label to remove not overlapping areas and the stereotactic frame present in CT. Three other label masks were generated using the following thresholds: greater than 300 HU for bone, less than -500 HU for air, and otherwise soft tissue. The labels were used to evaluate the synthetic CT errors for these specific regions. Fig. 1 shows the different image modalities after applying the preprocessing steps described above.

B. Neural Network Architectures

Convolutional neural networks (CNNs) are a type of artificial neural network that has become dominant in various computer vision and medical image analysis tasks. In general, they consist of three types of layers: convolution, pooling, and fully connected layers. The convolution layers play a key role in CNNs as they perform convolution operations to learn spatial learn spatial hierarchies of features. Pooling layers are used to reduce the dimensionality of feature maps whereas fully connected layers map the extracted features into a final output such as classification. [11]. Fully convolutional networks (FCNs) replace the fully connected layers by convolutional layers to allow multiple pixels to be predicted efficiently and simultaneously. CNNs were originally designed for 2D applications but effort has been made in recent years to implement 3D CNNs concerning the nature of medical images. Three state-of-the-art CNNs designed for medical image analysis were evaluated in this study: *VNet*, *HighRes3dNet*, and *ScaleNet*.

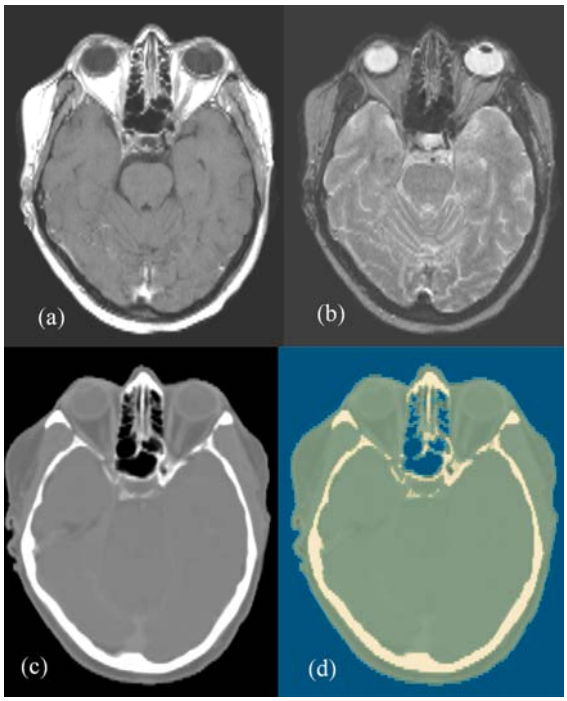

Figure 1. Axial slices of the different imaging modalities used to train the neural networks for CT synthesis: mrT1 (a), mrT2 (b), CT (c) and labels (d) for bone (yellow), soft tissue (green), and air (blue).

Vnet was one of the first neural networks architectures to implement volumetric convolutions instead of processing input volumes slice-wise. It is a FCN originally proposed to solve segmentation tasks. The network consists of two main parts to compress and decompress the signal until its original size is reached. Different stages operate at different resolutions comprising one to three convolutional layers. It also incorporates residual blocks that make use of special additive skip connections to combat vanishing gradients. Residual blocks allow the gradient to flow through the network more easily [12].

HighRes3dNet is a high-resolution fully 3D CNN that incorporates large volumetric context using dilated convolutions and residual connections. It is claimed to be conceptually simpler and more compact than other volumetric CNNs. It was originally designed for MRI brain parcellation [13].

ScaleNet is a multimodal deep learning architecture that uses nested structures to explicitly leverage features across modalities. It was developed to cope with the poorly generalization of CNNs to different image modalities for which they have been designed. For example, *HighRes3dNet* was designed for monomodal mrT1 and *ScaleNet* appropriately translates it into a network that jointly utilizes several image modalities, such as mrT1 and mrT2 [14].

C. Implementation Details

Our deep learning approach aims to transform the MRI inputs into synthetic CT outputs in a supervised regression learning setting, where corresponding data pairs are available. Therefore, the dataset was split into 70% training and 30% testing. Implementations of the neural network architectures used to perform the experiments are available in *NiftyNet*, a *TensorFlow*-based open-source CNNs platform for research in medical image analysis and image-guided therapy [15].

MRI volumes were whitened using volume level normalization calculated within the head foreground. All volumes were resampled to an isotropic voxel-size of 1.5 mm and were randomly sampled into patches of 48 × 48 × 48 voxels. These patches were sampled mainly within the head mask and were fed to the network using a batch of size 5. Data augmentation was implemented on the fly by randomly rotating each of the three orthogonal planes by an angle between -10º and 10º, and randomly scaled by a factor between 0.9 and 1.1.

The root mean square error (RMSE) was chosen as the loss function, Adam as the optimization method, Prelu as the activation function, and L2 regularization with weight decay 5 × 10-8. The learning rate was set to 0.001 for

HighRes3dNet and *ScaleNet*, and 0.0001 for *Vnet*. All models were trained from scratch on a single NVIDIA 1080 Ti GPU until convergence that occurred at approximately 20K iterations. Training took about 3 hours for *Vnet* and 5 hours for the other two networks, whereas inference time was approximately 1 minute per patient in all cases.

The multimodal *ScaleNet* network was implemented by feeding it with the auto-context model [16], which uses previous predictions as context input to the network. This was implemented at every 5K iterations and the head mask was used as the initial context input. The combinations of MRI inputs and networks were as follows:

- **T1-VN:** mrT1 + *Vnet*
- **T1-HR:** mrT1 + *HighRes3dNet*
- **T1-SN:** mrT1 + *ScaleNet*
- **T2-VN:** mrT2 + *Vnet*
- **T2-HR:** mrT2 + *HighRes3dNet*
- **T2-SN:** mrT2 + *ScaleNet*
- **T1/T2-VN:** mrT1/mrT2 + *Vnet*
- **T1/T2-HR:** mrT1/mrT2 + *HighRes3dNet*
- **T1/T2-SN:** mrT1/mrT2 + *ScaleNet*

III. RESULTS

A. Comparison of MRI sequences

Fig. 2 shows comparative box plots between MRI sequences for each neural network. The differences between mrT1 and mrT2 are not very clear but according to the values provided in Table I, mrT1 seems to achieve lower errors.

Even though mrT1/mrT2 achieved lower MAE than using only one MRI sequence, this difference was not statistically significant compared to mrT1 ($p = 0.48$, *ScaleNet*). Therefore, we can presume that by using only mrT1 sequences is possible to achieve similar performance than using mrT1 and mrT2 in a multimodal scheme. The latter is an important benchmark to focus further research in developing specific neural network architectures that provide higher performance on the most conventional and standard mrT1 sequences. The latter approach would have relevant impact on the imaging acquisition times and availability at the moment of generating synthetic CT images.

B. Comparison of neural network architectures

Fig. 3 shows the comparative plot between networks for each MRI sequence. *HighRes3dNet* achieved lower MAE than *Vnet* in all cases. *ScaleNet* outperformed the others for

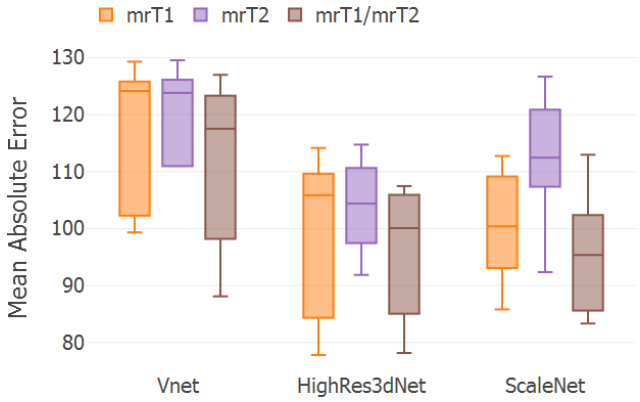

Figure 2. Box plots comparing the mean absolute error (MAE) for neural networks architectures and MRI sequences.

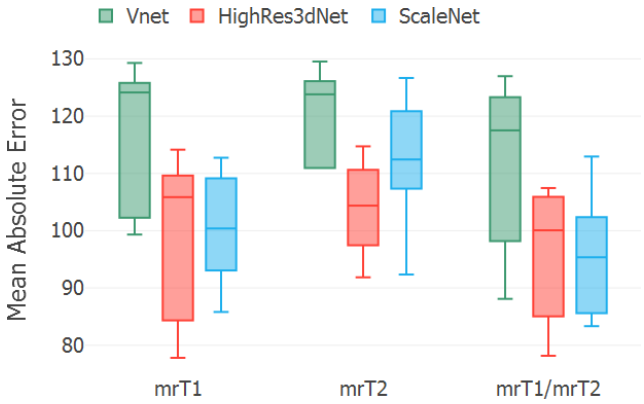

Figure 3. Box plots comparing the mean absolute error (MAE) for MRI sequences and neural networks.

mrT1 and mrT1/mrT2. It is worth mentioning that the learning parameters of the tested networks were not tuned, which, if performed could yield different results. Yet, this comparison supports the potential of deep learning for CT synthesis. Table I presents the corresponding mean absolute errors (MAE) on test set for different label masks. The background air was excluded for all calculations. The lowest errors were: 95.37 ± 11.7 HU for the whole head and 228.18 ± 31.0 HU for bone with model T1/T2-SN, 57.83 ± 3.8 HU for soft tissue using model T1/T2-HR, and 313.32 ± 48.5 HU for air with model T1-HR. Error maps for the two sample slices obtained with model T1/T2-SN is shown in Fig. 4. As expected, the highest errors were found at the contour of the head and air, especially in the nasal cavity due to the difficulty in predicting the air/bone interfaces in this region. The latter would not be a problem for dedicated PET scanners [18], as the field of view only includes the brain above the nasal region. Dedicated PET systems take advantage of MRI-based attenuation corrections as patients

TABLE I. THE MEAN ABSOLUTE ERROR (MAE) BETWEEN THE GROUND TRUTH AND SYNTHETIC CT (MEAN ± STANDARD DEVIATION)

	T1-VN	T1-HR	T1-SN	T2-VN	T2-HR	T2-SN	T1/T2-VN	T1/T2-HR	T1/T2-SN
Head	116.12 ± 13.7	98.49 ± 15.5	100.47 ± 10.5	120.04 ± 8.5	103.90 ± 8.8	112.54 ± 12.7	111.24 ± 16.0	95.68 ± 12.5	**95.37 ± 11.7**
Bone	287.50 ± 55.6	244.99 ± 40.7	262.48 ± 43.4	299.85 ± 24.5	266.46 ± 11.3	288.88 ± 75.0	267.18 ± 38.8	245.02 ± 42.5	**228.18 ± 31.0**
Tissue	73.25 ± 1.6	61.48 ± 7.4	59.78 ± 3.4	74.41 ± 7.3	61.71 ± 6.5	68.75 ± 10.6	70.52 ± 6.0	**57.83 ± 3.8**	60.97 ± 6.8
Air	342.50 ± 58.7	**313.32 ± 48.5**	354.67 ± 98.7	397.78 ± 85.6	398.76 ± 83.6	325.83 ± 67.0	392.28 ± 91.0	320.44 ± 67.6	334.51 ± 62.4

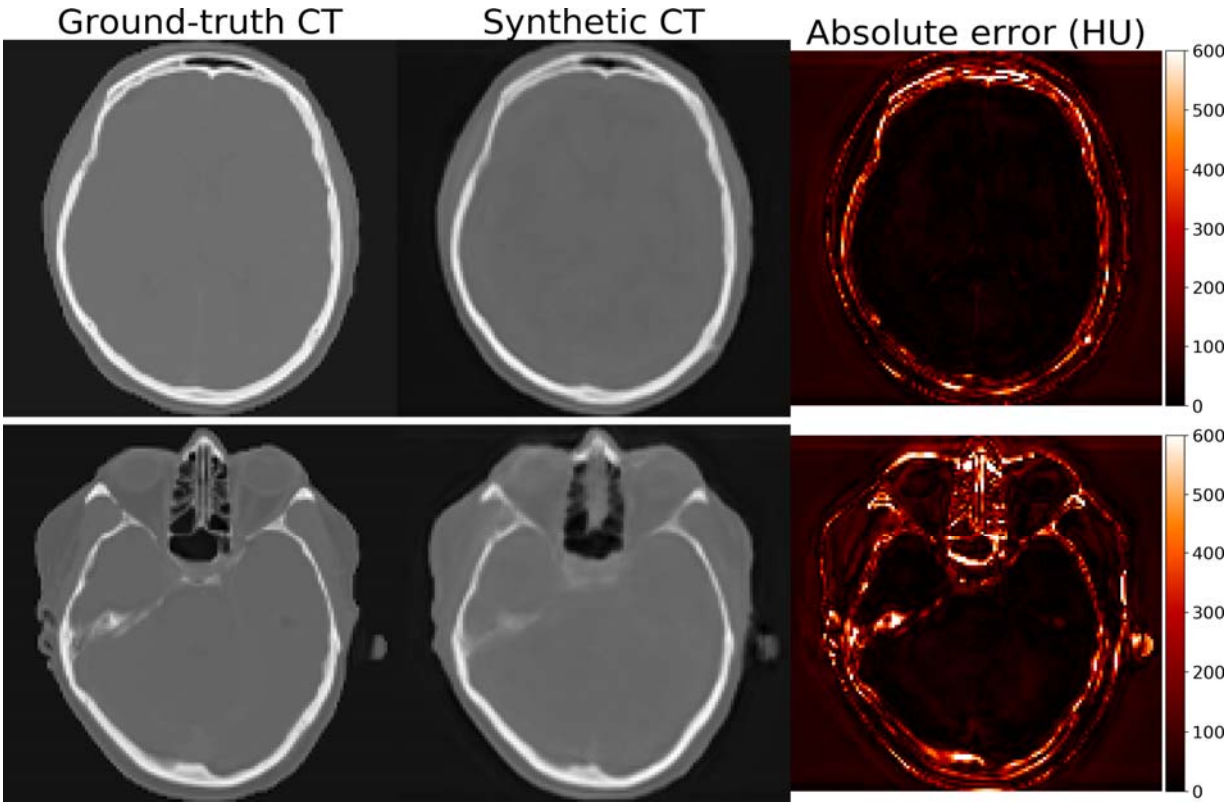

Figure 4. Error maps (right column) for two representative slices on test set obtained with model T1/T2-SN. Ground truth CT (left column) and synthetic CT (middle column) are al shown.

normally undergo routine MRI scans before the PET study.

All models presented in this study were compared to a well-known atlas-based method [17], which is accessible through *Niftyweb* (http://niftyweb.cs.ucl.ac.uk/). We submitted the subjects in our test set using the parameter "Optimize for Accuracy". The resulting MAE values within the head were: 175.81 ± 39.53 and 194.21 ± 44.63 HU for mrT1 and mrT2 respectively. These were significantly higher than ($p < 0.05$) than the deep learning implementations.

IV. Conclusion

This study presented a comparison of MRI inputs to different neural networks architectures for synthetic CT synthesis. We focused on the two most conventional and standard MRI sequences (T1- and T2- weighted) as we believe that specific MRI sequences such as Dixon or UTE are not always available, especially for radiotherapy planning and attenuation correction in dedicated brain PETs. Our results show that the three state-of-the-art neural networks (*Vnet*, *HighRes3dNet*, and *ScaleNet*) performed similarly. We plan to perform similar comparisons on larger datasets in a multicenter scheme to evaluate the generalization to MRIs acquired with different protocols and scanners.

References

[1] [1] K. P. E., T. D. W., B. T., and S. D., "Attenuation correction for a combined 3D PET/CT scanner," Med. Phys., vol. 25, no. 10, pp. 2046–2053, 1998.

[2] J. G. Mannheim, A. M. Schmid, J. Schwenck, P. Katiyar, K. Herfert, B. J. Pichler, and J. A. Disselhorst, "PET/MRI Hybrid Systems," Semin. Nucl. Med., vol. 48, no. 4, pp. 332–347, 2018.

[3] A. M. Owrangi, P. B. Greer, and C. K. Glide-Hurst, "MRI-only treatment planning: benefits and challenges," Phys. Med. Biol., vol. 63, no. 5, p. 05TR01, 2018.

[4] C. N. Ladefoged, I. Law, U. Anazodo, K. St. Lawrence, D. Izquierdo-Garcia, C. Catana, N. Burgos, M. J. Cardoso, S. Ourselin, B. Hutton, I. Mérida, N. Costes, A. Hammers, D. Benoit, S. Holm, M. Juttukonda, H. An, J. Cabello, M. Lukas, S. Nekolla, S. Ziegler, M. Fenchel, B. Jakoby, M. E. Casey, T. Benzinger, L. Højgaard, A. E. Hansen, and F. L. Andersen, "A multi-centre evaluation of eleven clinically feasible brain PET/MRI attenuation correction techniques using a large cohort of patients," Neuroimage, vol. 147, no. June 2016, pp. 346–359, 2017.

[5] X. Han, "MR-based synthetic CT generation using a deep convolutional neural network method," Med Phys, vol. 44, no. 4, pp. 1408–1419, 2017.

[6] F. Liu, H. Jang, R. Kijowski, T. Bradshaw, and A. McMillan, "Deep Learning MR Imaging – based Attenuation Correction for PET/MR Imaging," Radiology, vol. 286, no. 2, pp. 676–684, 2018.

[7] K. Kläser, P. Markiewicz, M. Ranzini, W. Li, M. Modat, B. F. Hutton, D. Atkinson, K. Thielemans, M. J. Cardoso, and S. Ourselin, "Deep Boosted Regression for MR to CT Synthesis," pp. 1–10, 2018.

[8] D. Nie, R. Trullo, C. Petitjean, S. Ruan, and D. Shen, "Medical Image Synthesis with Context-Aware Generative Adversarial Networks," 2016.

[9] Y. Qiao, B. Lew, B. Lelieveldt, and M. Staring, "Fast Automatic Step Size Estimation for Gradient Descent Optimization of Image Registration.," IEEE Trans. Med. Imaging, vol. 35, no. 2, pp. 391–403, 2015.

[10] “elastix: A Toolbox for Intensity-Based Medical Image Registration.”
[11] R. Yamashita, M. Nishio, R. Kinh, G. Do, and K. Togashi, “Convolutional neural networks : an overview and application in radiology,” 2018.
[12] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation,” pp. 1–11, 2016.
[13] W. Li, G. Wang, L. Fidon, S. Ourselin, M. J. Cardoso, T. Vercauteren, W. L. B, G. Wang, L. Fidon, S. Ourselin, M. J. Cardoso, and T. Vercauteren, “On the compactness, efficiency, and representation of 3D convolutional networks: Brain parcellation as a pretext task,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10265 LNCS, pp. 348–360, 2017.
[14] L. Fidon, W. Li, L. C. Garcia-peraza-herrera, J. Ekanayake, N. Kitchen, S. Ourselin, and T. Vercauteren, “Scalable multimodal convolutional networks for brain tumour segmentation.”
[15] E. Gibson, W. Li, C. Sudre, L. Fidon, D. I. Shakir, G. Wang, Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu, T. Whyntie, P. Nachev, M. Modat, D. C. Barratt, S. Ourselin, M. J. Cardoso, and T. Vercauteren, “NiftyNet: a deep-learning platform for medical imaging,” Comput. Methods Programs Biomed., vol. 158, pp. 113–122, 2018.
[16] Z. Tu and X. Bai, “Auto-context and its application to high-level vision tasks and 3D brain image segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 10, pp. 1744–1757, 2010.
[17] N. Burgos, M. J. Cardoso, K. Thielemans, M. Modat, S. Pedemonte, J. Dickson, A. Barnes, R. Ahmed, C. J. Mahoney, J. M. Schott, J. S. Duncan, D. Atkinson, S. R. Arridge, B. F. Hutton, and S. Ourselin, “Attenuation correction synthesis for hybrid PET-MR scanners: Application to brain studies,” IEEE Trans. Med. Imaging, vol. 33, no. 12, pp. 2332–2341, 2014.
[18] C. M. Atienza and J. L. Peris, “Development of a new device for the early diagnosis of Alzheimer ’ s disease,” no. June, 2018.