A k-Space Model of Movement Artefacts: Application to Segmentation Augmentation and Artefact Removal

Richard Shaw, Carole H. Sudre, Thomas Varsavsky, Sébastien Ourselin, and M. Jorge Cardoso

***Abstract*—Patient movement during the acquisition of magnetic resonance images (MRI) can cause unwanted image artefacts. These artefacts may affect the quality of clinical diagnosis and cause errors in automated image analysis. In this work, we present a method for generating realistic motion artefacts from artefact-free magnitude MRI data to be used in deep learning frameworks, increasing training appearance variability and ultimately making machine learning algorithms such as convolutional neural networks (CNNs) more robust to the presence of motion artefacts. By modelling patient movement as a sequence of randomly-generated, 'demeaned', rigid 3D affine transforms, we resample artefact-free volumes and combine these in k-space to generate motion artefact data. We show that by augmenting the training of semantic segmentation CNNs with artefacts, we can train models that generalise better and perform more reliably in the presence of artefact data, with negligible cost to their performance on clean data. We show that the performance of models trained using artefact data on segmentation tasks on real-world test-retest image pairs is more robust. We also demonstrate that our augmentation model can be used to learn to retrospectively remove certain types of motion artefacts from real MRI scans. Finally, we show that measures of uncertainty obtained from motion augmented CNN models reflect the presence of artefacts and can thus provide relevant information to ensure the safe usage of deep learning extracted biomarkers in a clinical pipeline.**

***Index Terms*—MRI, motion artefacts, deep learning, segmentation, data augmentation, artefact correction, uncertainty.**

I. Introduction

PATIENT movement during the acquisition of magnetic resonance images (MRI) can result in unwanted image artefacts, which manifest as blurring, ringing or ghosting effects, depending on both timing and spatial changes during a scan [1]. Motion artefacts can affect the interpretability of images, potentially affecting the quality of a patient's diagnosis, and/or leading to increased cost if the images are judged unusable and the acquisition has to be repeated. Artefacts can also affect the performance of post-processing algorithms, and it has been shown that motion artefacts consistently affect segmentation measurements on structural MR images [2]. Furthermore, in the context of research cohorts, artefacts may lead to inclusion bias in statistical analysis as more impaired subjects tend to have difficulties staying still, resulting in poorer quality scans more likely to be excluded [3]. Even if included, biomarker measures may be biased by artefacts leading to spurious findings [2].

The type of motion artefacts that appear in MR images depends on the amount and timing of patient movement with regards to the k-space acquisition trajectory. Movements that occur close to the k-space centre correspond to low image frequencies and tend to result in ghosting artefacts, where the image is repeated, as does quasi-periodic motion e.g. respiration [4]. Movements toward the edges of the k-space corresponding to the acquisition of high image frequencies, often lead to ringing artefacts. Most commonly observed MRI motion artefacts introduce minor blurring due to small movements spanning a range of frequencies during k-space acquisition. Additionally, motion artefact appearance depends on the k-space scanning strategy and notably whether the acquisition is performed in 2D or 3D.

Prior work on motion artefacts in MRI has mainly focused on designing ways of correcting for them, for example [5], [6], [7], [8], [9] and [10]. This work, however, addresses the problem of motion artefacts under a different perspective – attempting to make automated systems of image analysis more robust to their presence. In recent years, deep learning frameworks have demonstrated high performance when applied to segmentation and classification tasks. In a deep learning setup, data augmentation is a classical way to artificially increase data variability and thus increase the network's potential for generalisation [11]. While classical data augmentation usually involves random geometric transformations and/or intensity changes, biologically and physically plausible augmentation models would be beneficial to better sample the space of possible variations.

II. Motion Artefacts in Deep Learning Frameworks

Deep learning frameworks dealing with motion artefacts have so far proposed to either recognise corrupted images

R. Shaw (richard.shaw.17@ucl.ac.uk) and T. Varsavsky (thomas.varsavsky.15@ucl.ac.uk) are PhD candidates within the Department of Medical Physics and Biomedical Engineering, University College London, UK, and co-affiliated with the School of Biomedical Engineering and Imaging Sciences, King's College London, UK. C. H. Sudre (carole.sudre@kcl.ac.uk), S. Ourselin (sebastien.ourselin@kcl.ac.uk) and M. J. Cardoso (m.jorge.cardoso@kcl.ac.uk) are with the School of Biomedical Engineering and Imaging Sciences, King's College London, UK.

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

or attempted to correct for the presence of artefacts. Meding et al. [13] used convolutional neural networks to classify MR magnitude images as artefacted or not. Going beyond the binary classification task, Duffy et al. [14], also using CNNs, attempted to learn how to retrospectively remove artefacts from MR images. Their network, however, trained on synthetic data proposes an unrealistic motion model that is limited to axial translation. Using a Generative Adversarial Network, Armanious et al. proposed MedGAN [15] with the objective of 'translating' motion-corrupted MR images to their corresponding motion-free images, but restricted their work to 2D slices. Pawar et al. [16], with the objective of learning to remove artefacts, modelled 3D motion in the image domain and reconstructed the k-space from multiple resampled images using however only 2D axial slices. In contrast to these approaches, we argue that it is ultimately more useful to optimise frameworks in an end-to-end manner, rather than generating intermediate motion-corrected images, thus enforcing robustness to artefacts at the level of the internal representation of the data. Such strategy inherently avoids caveats of GANs, that may wrongly introduce non-existing information (hallucination), or of artefact removal strategies that may only account for part of the existing artefacts thus resulting in data that is unusable for further processing. Moreover, end-to-end learning allows for model artefact-induced task uncertainty to be learned directly from raw artefact inputs.

III. Motion Artefact Model

We propose a k-space augmentation method to generate motion artefacts from artefact-free magnitude MR image volumes. Our proposed method is illustrated in Fig. 1 [17]. The procedure is summarised by the five following steps: (1) Generate a random movement model by sampling movements from different probability distribution functions (PDFs). (2) Demean the generated movement transforms. (3) Resample the artefact-free volume according to the demeaned movement model. (4) Reconstruct a composite k-space from the k-spaces of multiple resampled volumes. (5) Transform the combined k-space back to the image domain to produce the final artefact sample.

Taking each of these steps in turn, we first sample movements from different probability distribution functions, modelling a patient's head motion throughout the scan as a sequence of independently occurring small and large motions (e.g. twitches/nodding). Each movement is modelled by a 3D affine A matrix comprising of a rigid 3D rotation and translation in the image domain, where the angles of rotation $\boldsymbol{\theta}$ are sampled between $(-30^{\circ}, 30^{\circ})$ and the translation $\boldsymbol{\delta}$ between (-10mm, 10mm) in all three axes. Poisson distributions are used to sample the magnitude of rotation and translation of each of the N movements – small movements are assumed to occur more often and large movements less frequently – while a uniform distribution is used to sample the time t in k-space at which each movement occurs (assuming k-space scans in the phase encoding direction). This means that a movement occurring at time t corresponds to a specific location in the k-space volume $\boldsymbol{k}_t = (k_x, k_y, k_z)$ depending on the scan trajectory, such that the brain remains in position i between k-space elements $\boldsymbol{k}_{t_i}$ and $\boldsymbol{k}_{t_{i+1}}$. The sequence of movement transforms $\{A\}_{i=1}^{N}$ is composed incrementally in log-Euclidean space [18], using the matrix exponential $\exp(A) = \sum_{k=0}^{\infty} \frac{A^k}{k!}$ and corresponding matrix logarithm. By transferring to the log-Euclidean domain this provides us with the ability to create weighted combinations of transformations and to linearly interpolate between them.

With the motion model defined, the second step is to 'demean' the movements. When applying our augmentation model to a clean magnitude image I_0, we expect the barycenter of the imaged object to remain in approximately the same position within the 3D volume as this is the position of the tissue segmentation. This is achieved by 'demeaning' each affine transform A_i by pre-multiplying by the inverse of the average transform A_{avg}, computed as the weighted sum of the

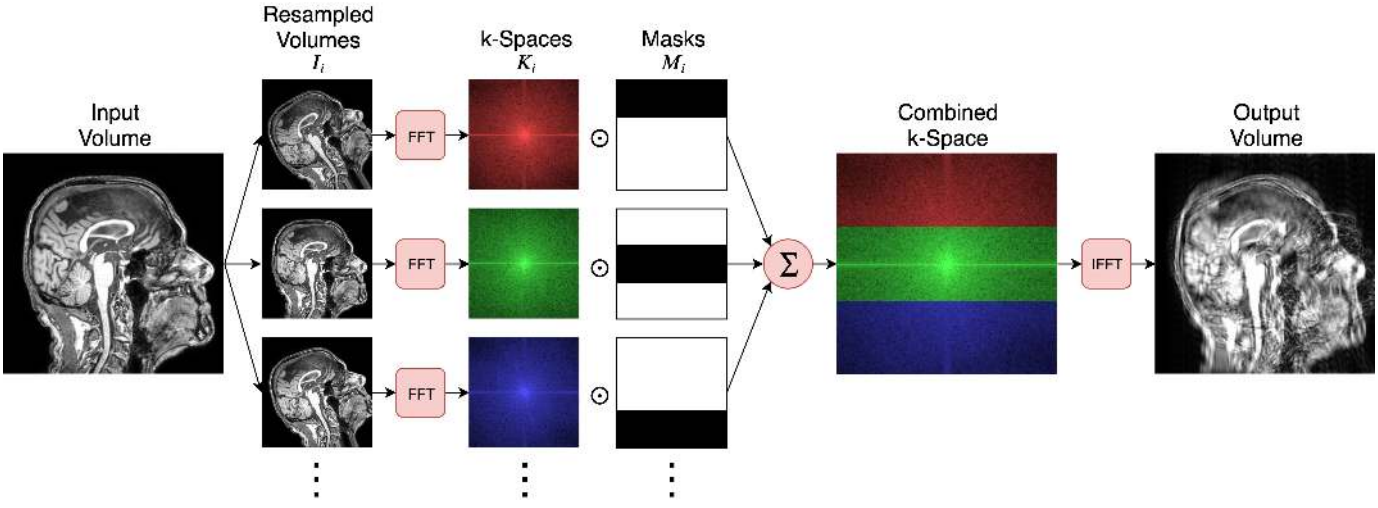


Fig. 1. Motion artefact augmentation framework: The artefact-free input volume is resampled according to a randomly sampled movement model, defined by a sequence of 'demeaned' 3D affine transforms. Their 3D Fourier transforms are combined to form a composite k-space, which is transformed back to the image domain producing the final artefact volume.

sequence of N affine transformations in log-Euclidean space, given by Equation 1,

$$A_{avg} = \exp\left(\sum_{i=1}^{N} \hat{w}_i \log(A_i)\right), \quad (1)$$

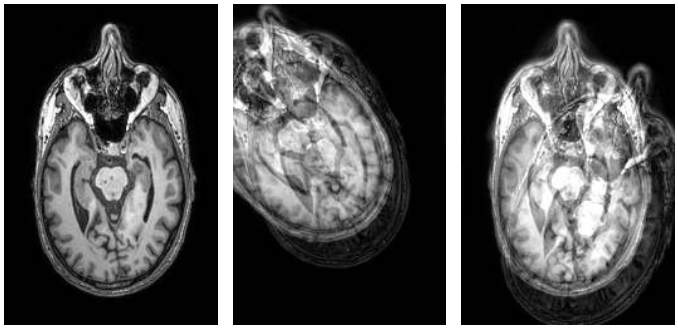
where $\hat{w}_i$ is the weighting given to the i-th movement. Since movements at different parts of the k-space contribute different spatial frequencies, we weight each A_i by its signal contribution to the final image. This means that movements at the k-space centre (low frequencies) have a higher weight since their impact on the final 3D position of the brain, and the overall Fourier power spectrum, is much greater. Each weight is estimated by masking the 3D k-space of I_0 with a binary mask M_i corresponding to the k-space elements acquired while at position i, transforming back to the image domain, and summing the resulting voxel intensities, as given by $w_i = \sum_{voxels} FFT^{-1}(M_i \odot FFT(I_0))$, with the weights then normalised to sum to 1.

The third step is to apply each demeaned affine transform A_i^d to the original artefact-free image volume I_0 and resample using b-spline interpolation. Note that we always resample the original image volume to reduce propagating interpolation errors throughout the sequence. The i-th demeaned affine transformation is therefore given as the sum in log-Euclidean space of all demeaned transforms up to this point, as given by Equation 2,

$$A_i^d = \exp\left(\log(A_{i-1}^d) + \log(A_{avg}^{-1}) + \log(A_i)\right) \quad (2)$$

where the initial transform A_0^d is set to the demeaned identity transform, i.e. $\log(A_{avg}^{-1})$. Following each transformation, we compute the k-space K_i as the 3D Fourier Transform of each resampled image I_i.

The fourth step is to combine the 3D Fourier transforms corresponding to each position of the brain in the sequence, joined together at sampled times t, forming a complete k-space of the scan containing movement, i.e. $K_c = \sum_i^N M_i \odot K_i$. Finally, the inverse 3D Fourier Transform of the composite k-space is derived, and the magnitude image provides the final artefact sample. The steps of the augmentation method are more formally outlined in Algorithm 1 and examples of our artefact augmentation are shown in Fig. 2. The effect of the demeaning process on the brain position is shown in Fig. 3.



a) Input volume b) No demeaning c) With demeaning

Fig. 3. Effect of demeaning on the position of the brain: a) the input image volume, b) the final position of the brain without demeaning, c) the demeaned position. Demeaning keeps the artefacted brain in roughly the same position as the input, while without demeaning the brain moves out of field of view.

Implementation Details

Although the proposed movement model is a simplified approximation to patient motion within the scanner, in practice the augmentation procedure is quite computationally expensive, but not prohibitively so. This is due to resampling the input image volume to generate different head positions, especially when each volume in our dataset is around 256^3 voxels in size. A significant time component is also a result

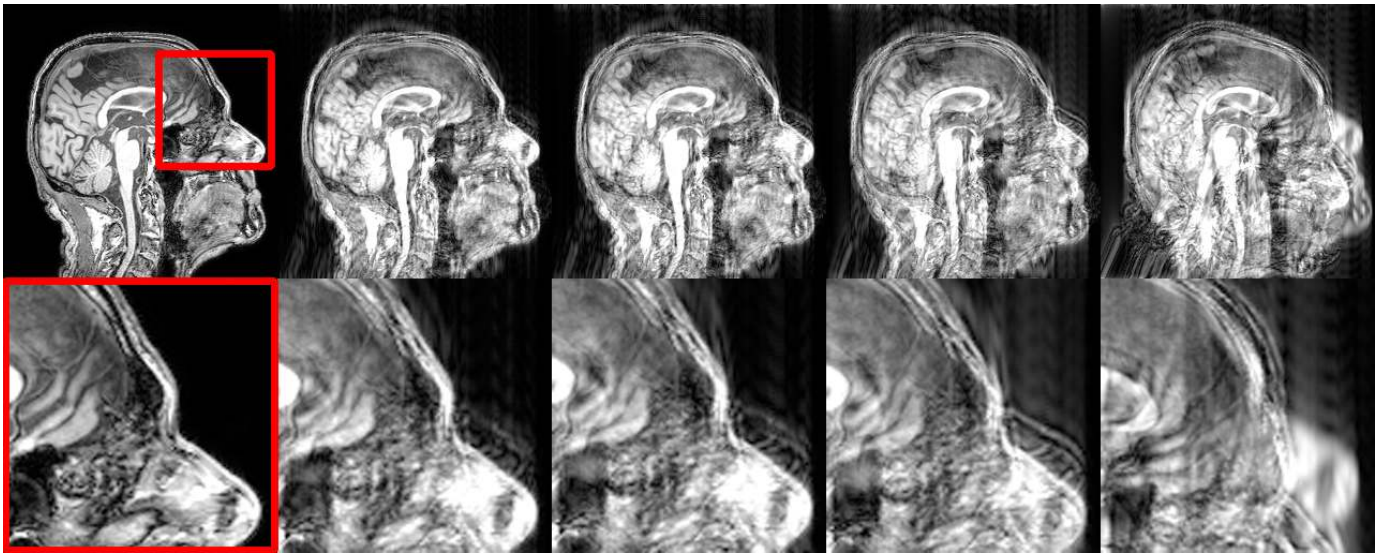


Fig. 2. Motion artefacts generated by our augmentation model as a result of rotation of the head forwards and backwards around the coronal axis, simulating patient nodding motion. Changing artefact appearance due to changing the time during acquisition at which the movement occurs, later in the k-space scan trajectory from left to right, therefore retaining more lower spatial frequencies. Best viewed zoomed in on digital copy.

of the 3D FFT of resampled images and the inverse 3D FFT of the combined k-space. As the number of times an image is resampled is randomised, the time taken to generate motion artefacts varies between samples. However, on average, it takes roughly 30 seconds to generate an artefact sample from a clean input volume on the CPU. This will slow down the training of neural networks if this is done on-the-fly, but artefacts can be pre-computed before training. Modern deep learning frameworks are also starting to allow for Fourier domain operations such as FFTs on the GPU, meaning that the proposed augmentation is likely to see significantly speed-ups in future.

Algorithm 1: Motion artefact augmentation algorithm.

Input: Artefact-free image volume I_0

Result: Artefacted image volume I_a

▷ Sample N movements

$\{\boldsymbol{\theta}, \boldsymbol{\delta}, t\}_{i=1}^{N}$

▷ Construct 3D affine matrices

$\{A\}_{i=1}^{N},\ A_i = [R(\boldsymbol{\theta}_i)|\boldsymbol{\delta}_i]$

▷ Construct k-space masks

$$\{M\}_{i=0}^{N}, M[k_x, k_y, k_z] = \begin{cases} 1 & \text{if } \boldsymbol{k}_{t_i} < \boldsymbol{k} < \boldsymbol{k}_{t_{i+1}} \\ 0 & \text{otherwise} \end{cases}$$

▷ Compute weights

$w_i = \sum_{voxels} FFT^{-1}(M_i \odot FFT(I_0))$

$\hat{w}_i = w_i / \sum_i^N w_i$

▷ Average affine transform

$A_{avg} = \exp\left(\sum_{i=1}^{N} \hat{w}_i \log(A_i)\right)$

▷ Demean and resample sequence

Init: $A_0^d = A_{avg}^{-1}$

for $i = 1, ..., N$ **do**

  $A_i^d = \exp\left(\log(A_{i-1}^d) + \log(A_{avg}^{-1}) + \log(A_i)\right)$

  $I_i \leftarrow Resample(I_0, A_i^d)$

  $K_i = FFT(I_i)$

end

▷ Combine k-spaces

$K_c = \sum_{i=0}^{N} M_i \odot K_i$

$I_a = FFT^{-1}(K_c)$

IV. Experiments

We evaluate our motion artefact augmentation model on both simulated and real-world data containing artefacts in the context of three segmentation tasks: cortical gray matter (CGM), hippocampus and total intracranial volume (TIV). Data used in this work was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). Launched in 2003, ADNI attempts to assess whether medical imaging and biological markers and clinical assessment can be combined to measure progression of Alzheimer's Disease. More information can be found at www.adni-info.org.

A. Network Architecture and Implementation Details

We used the HighRes3DNet [19] architecture implemented in NiftyNet [20], with Dice loss [21], a patch size of 80^3 and batch size of 1, trained on a single GPU with Adam optimiser [22] and a learning rate of 10^{-4}. In the context of segmentation, due to imbalance between foreground and background elements, the sampling strategy is essential to the training performance. Therefore we employed a weighted patch sampling with higher weight given to regions defined by the Gaussian blurred ground-truth segmentation labels, such that the foreground/background weight ratio is roughly equal to the ratio of foreground/background voxels. Each model was trained until overfitting was observed or when reaching 100,000 iterations.

B. Simulated Dataset

For experiments on simulated data, we use 272 MPRAGE scans from ADNI and generate 15 artefacted volumes per scan. The data was split into 80% training, 10% validation and 10% testing and separate CNNs were trained to segment CGM, hippocampus and TIV. For each segmentation task, five models were trained with varying levels and types of augmentation. One was trained only on 'clean' data, i.e. the original artefact-free scans. Another was trained with 'classical' augmentation, consisting of random rotation, translation and scaling. The remaining three models were trained with increasing amounts of motion artefact augmentation: where 25%, 40% and 50% of images seen in the training set contain movement artefacts, in addition to classical augmentations. Each model includes bias field augmentation by default to account for variability in image intensity across samples. All models are tested on the same hold-out test set containing both clean and artificially artefacted data.

Segmentation performance for the three tasks across all models is evaluated with Dice score, positive predictive value (PPV), sensitivity and average distance (avgDist) metrics and presented in the first row of Fig. 4. Results of Bonferroni corrected matched pair Wilcoxon tests between models are presented on the bottom row. Generally, across the metrics, models trained with artefacts show improved performance on the test set, with lower variance. In the case of CGM, motion augmented models show a statistically significant improvement over the clean model (particularly 40% and 50% artefact models), and similar improvement over classical augmentation, except for average distance for which classical augmentation performs better. For hippocampus, motion augmentation improves upon the clean model in all metrics but is outperformed by classical augmentation, suggesting artefacts have less impact on the hippocampal region. For TIV, the 50% augmented model consistently performs the best and is statistically significant in terms of Dice and average distance. Since the distribution and severity of artefacts in the data is randomised, the impact on model performance varies across

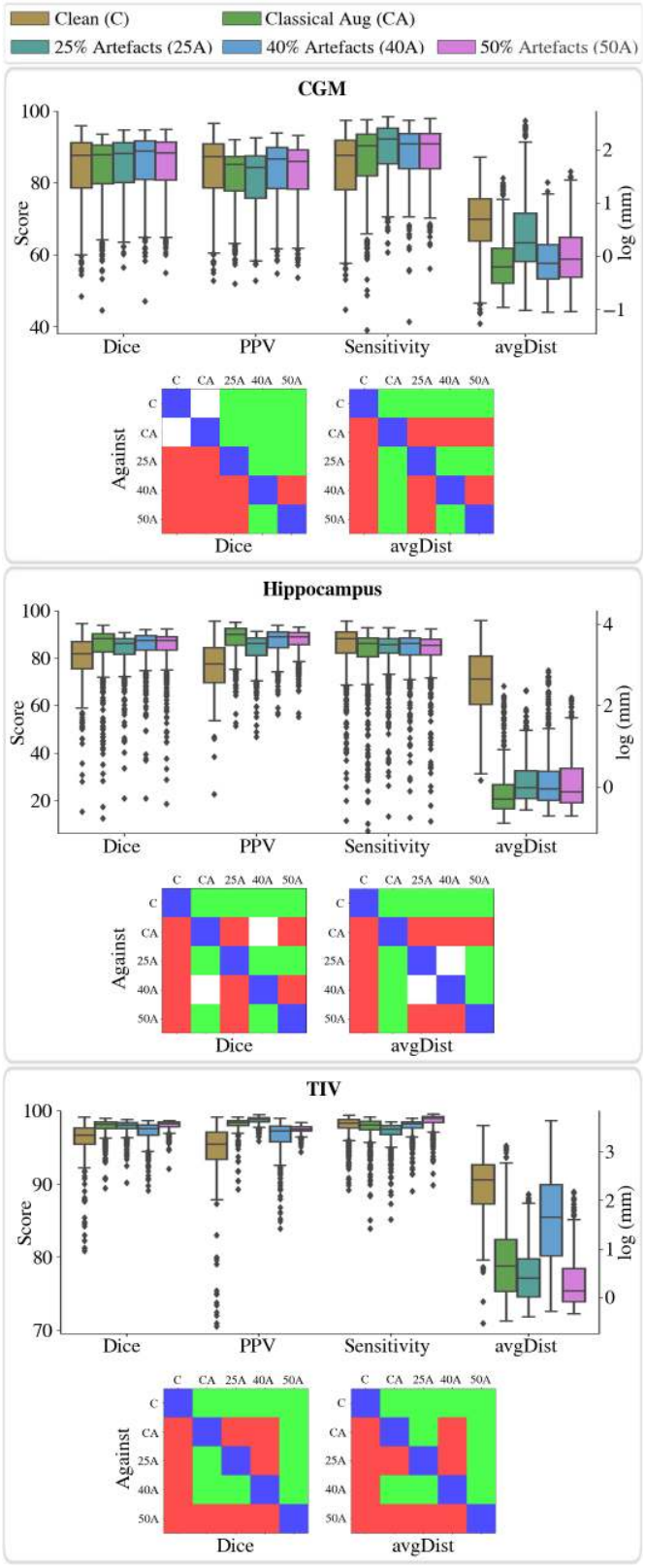


Fig. 4. Segmentation results on simulated data for CGM, hippocampus and TIV across models. Top: Boxplots of different error metrics for 5 models with different augmentation: clean, classically augmented, 25%, 40%, 50% artefacts. Bottom: Bonferroni corrected pairwise Wilcoxon tests for Dice and average distance between column and row models - Green: significantly better, White: no statistically significant difference; Red: significantly worse.

the metrics, but overall increased artefact variability translates to better performance on the test set.

C. Real-world Dataset

Robustness of CNNs trained with the proposed motion augmentation to real-world movement artefacts was then evaluated in a test-retest setting. 106 quality-controlled (QC) pairs of MPRAGE test-retest images from ADNI on which only one of the images was considered artefacted by an expert human rater were used for this purpose.

The criteria for the QC is described in detail on the ADNI website. T1 images are manually graded subjectively by trained analysts into categories: 1-3 is acceptable and 4 is failure (unusable). Images graded as "excellent" contain no artefacts and these are used as ground-truth, while images graded as "good," "poor" or "unusable" may contain artefacts. Images indicated as "containing artefacts" are used if they have a corresponding artefact-free retest scan. We specifically chose images for which the rater had commented: "contains artefacts due to motion", "blurring due to motion", "ringing" or "ghosting". Image pairs were chosen if it was mentioned that one scan was significantly better quality than the other, amounting to 106 test-retest pairs, a selection of which are shown in the Appendix.

Each test-retest image pair was rigidly registered together in a group-wise space to avoid interpolation bias. For comparison purposes, a benchmark label fusion algorithm [23] was used to perform the segmentation tasks on each pair of images. For each trained CNN model and the benchmark method, Dice score, PPV, sensitivity and average distance were used as evaluation measures between test and retest images, with the results obtained on the clean image being used as reference. Fig. 5 presents in the top row the corresponding boxplots for each segmentation task, while the second row displays the Bonferroni corrected matched-pair Wilcoxon tests across models. On the real-world data, the boxplots show generally improved robustness (higher score and smaller variance) for increasing amounts of simulated movement augmentation during training. This suggests our augmentation model translates well to a real-world setting. In terms of Dice score, the 50% artefact model performs the best across all tasks and shows a statistically significant improvement. The model also shows improvements for PPV and sensitivity but not average distance. The 40% artefact model is similar but performs poorly on TIV, perhaps due to the distribution of artefacts in the data. The clean model consistently performs the worst for all tasks and is even outperformed by the benchmark method. This is because the clean model has seen only "perfect" clean data and is therefore unable to generalise to poor-quality artefact images. The largest performance increase is between the clean and classically augmented models since spatial changes are the dominating cause of appearance variability in the data. Increasing amounts of artefact augmentation on top of classical augmentation generally show further improvements.

V. Task-specific Uncertainty Estimation

Deep learning models for segmentation tasks classically provide for each voxel a point-estimate probability of belong-

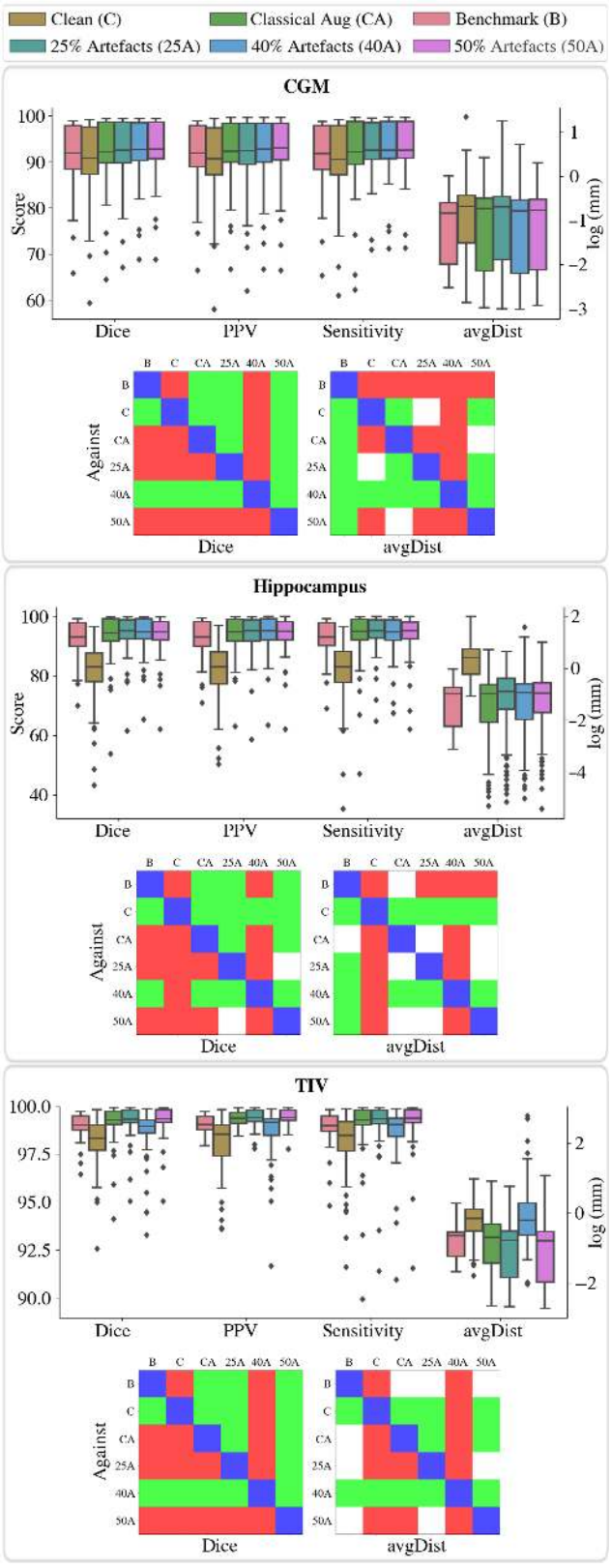


Fig. 5. Segmentation robustness on real-world test-retest data for CGM, hippocampus and TIV across models. Top: For each task, boxplots of different error metrics for the 5 CNN models in addition to a non-deep learning "benchmark" method. Bottom: Bonferroni corrected pairwise Wilcoxon tests for Dice and average distance between column and row models - Green: significantly better, White: no statistically significant difference; Red: significantly worse.

ing to a certain class. Being able to provide in addition a calibrated measure of the uncertainty of a given prediction has become essential in applications for which safety is paramount such as medical applications.

As theorised by Gal and Gharhamani [24], uncertainty can be estimated by sampling at inference time from multiple outputs of the network trained with dropout. Adapting the approach from [25], uncertainty over the segmentation result is obtained as the variance over the predictions made from multiple forward passes of the dropout network. For training, the dropout rate was set at 0.5 everywhere except the initial layer, which was set to 0.05, and the final layer for which no dropout was used. Uncertainty estimates were made from 100 forward passes of the dropout network. Mean and variance results obtained on the CGM and hippocampus segmentation tasks for the aforementioned models trained with dropout are shown in Fig. 6. Considering the uncertainty predictions for CGM segmentation, in models trained with motion augmentation, higher variance of predictions are observed in artefacted regions, especially close to the cortical surface, in comparison to the predicted uncertainty given the clean image. It is clear that uncertainty predictions made by the motion augmented model reflect the presence of motion artefacts in the data. Note that this is a behaviour that the clean and classically augmented models do not exhibit.

To further investigate the behaviour of segmentation uncertainty estimation in the presence of motion artefacts, with respect to the type of augmentation applied at training, per-voxel Kullback-Leibler divergence (KLD) maps comparing the sampled distributions for clean and artefacted images were calculated, as shown in the bottom rows of Fig. 6. By associating KLD with uncertainty, as measured by the sampled variance (std), we can examine this relationship for each model and mode of augmentation, visualised by the histogram plots of uncertainty on the artefacted image vs KLD in Fig. 7.

From the histograms, different modes of association between uncertainty and KLD can be interpreted as follows: 1) low std - low KLD: the model gives similar predictions on clean and artefacted images in a confident fashion; 2) high std - low KLD: the model provides highly similar distributions but overestimates uncertainty 3) low std - high KLD: the model provides mismatching answers with high confidence, a clinically unsafe behaviour 4) high std - high KLD: the probability distributions are different from each other but the model is aware that it cannot ascertain the results with certainty. Note that, in the presence of heavily artefacted images (Fig. 7 a)), models trained on clean data or with only classical augmentation behave unsafely more often, i.e. more predictions with high KLD and low uncertainty. Models trained with motion augmentation were found to be safer.

VI. Artefact Removal

While not the main focus of this work, in this section we demonstrate that our artefact augmentation model can be also used to train CNNs to learn to retrospectively remove, to a limited extent, motion artefacts from MR images. Previously, we have argued that it is generally more useful to learn a

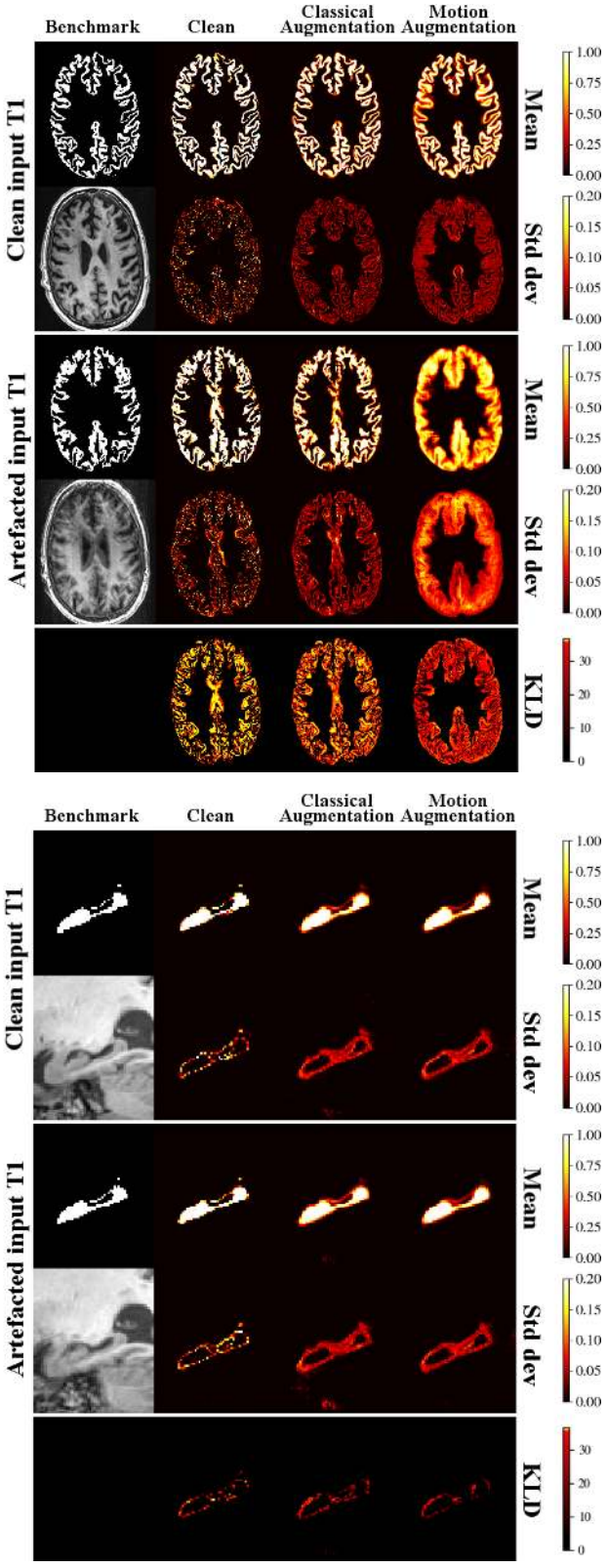
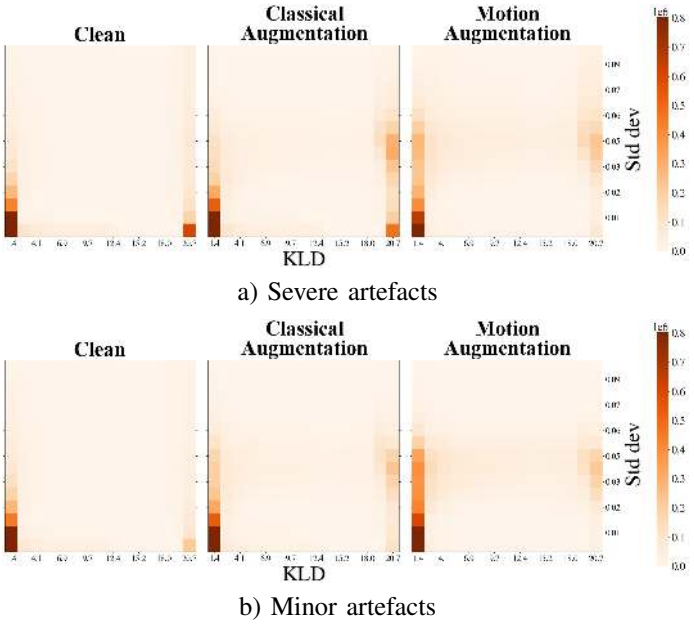


Fig. 6. Per-voxel mean and uncertainty estimations on CGM (top) and hippocampus (bottom) segmentation tasks for clean (no augmentation), classically augmented and motion augmented models for a test-retest pair for which one scan is heavily artefacted. The segmentation produced by a benchmark method is shown for reference. Bottom row of each block: KL-divergence (KLD) between the probability distributions produced by each model on clean and artefacted scans.



a) Severe artefacts

b) Minor artefacts

Fig. 7. Histograms of per-voxel KLD associated to the uncertainty estimates as measured by the sample variance, shown for models trained with different augmentations on a) severe artefacts and b) minor artefacts. Note that the clean and classically augmented models produce a much higher number of estimates with high KLD and high variance (bottom-right corner of the histograms) compared to the motion augmented model.

task such as segmentation in an end-to-end manner so that it is robust to the presence of motion artefacts, rather than learning an intermediate step of an artefact-free image as this is essentially compressing the output of the bottleneck. However, in some cases it may be useful to work with an artefact-corrected image. For instance, a radiologist may wish to see an artefact-corrected image for visual inspection. Furthermore, many non-deep learning algorithms would not work within an end-to-end framework and therefore would require the artefact-corrected images as input.

Using our motion augmentation model, we trained CNNs to remove first synthetic artefacts from 3D image volumes, where the inputs to the network are simulated artefact images and the ground-truth labels are the corresponding artefact-free images. Model performance is evaluated by computing the average reconstruction error between artefacted and clean images since our augmentation model preserves the alignment of the object. We then applied the artefact correction model trained only on simulated artefacts to a dataset of unseen real-world artefacts and similarly evaluated the reconstruction performance. Since our dataset of real-world images containing artefacts have been rigidly aligned to their corresponding ground-truth artefact-free images, we can directly compute the resulting error between them. To quantitatively evaluate the accuracy of the motion-corrected images, we compute the following error metrics between the ground-truth artefact-free images and the artefact-corrected output images from the CNN model: the mean absolute error (MAE) and the structural similarity index (SSIM), where SSIM is computed as given by Equation 3.

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

	Simulated	
	Uncorrected	Corrected
MAE	0.00434 [0.00323, 0.00599]	**0.00343** [0.00271, 0.00452]
SSIM	0.96835 [0.94835, 0.98338]	**0.98989** [0.98193, 0.99377]

	Real-world	
	Uncorrected	Corrected
MAE	0.00302 [0.00231, 0.00340]	0.00297 [0.00230, 0.00338]
SSIM	0.98566 [0.98001, 0.98886]	0.98604 [0.98105, 0.98960]

TABLE I
ARTEFACT REMOVAL PERFORMANCE ON HOLD-OUT SIMULATED AND REAL-WORLD ARTEFACTS PRESENTED AS MEDIAN [1ST QUARTILE, 3RD QUARTILE]. PAIRWISE STATISTICALLY SIGNIFICANT ACCORDING TO PAIRED SAMPLE WILCOXON TESTS (BONFERRONI CORRECTED) IMPROVEMENTS ARE INDICATED IN BOLD.

A. Network Architecture and Implementation Details

As in our previous experiments, we utilised the HighRes3DNet architecture implemented in NiftyNet, however, modified for the regression task with a skip connection joining the input to the output. We trained with an L1 loss function on voxel intensities, a patch size of 80^3 and batch size of 1, on a single GPU with Adam optimiser and a learning rate of 10^{-4} for 100,000 iterations. As this is primarily an investigation of the effect of simulated artefact augmentation, other network parameters were not explored. Better artefact removal performance may be achievable with network/parameter searching and more sophisticated loss functions such as L1 on image gradients and/or SSIM which are commonly used for image reconstruction tasks, but this is beyond the scope of this paper.

B. Simulated Dataset

Using the same 272 scans from the ADNI dataset, we generated 300 random artefact samples per scan, with varying degrees of artefacts. Fig. 8 shows a sample of results for a hold-out synthetic test set. Table I (top) shows the results for MAE and SSIM between the uncorrected and motion-corrected images. For the corrected data, we observe a statistically significant lower median MAE, and a statistically significant higher median SSIM (indicated by the bold values). For both metrics the interquartile ranges are smaller for the corrected data.

C. Real-world Dataset

Using the motion artefact correction model trained using only simulated data, inference was then performed on real-world artefacts from the ADNI dataset of 106 rigidly aligned quality-controlled MPRAGE test-retest image pairs. Fig. 9 shows a sample of our results on real-world artefact images not seen in training.

Table I (bottom) shows the results for MAE and SSIM metrics on the real-world hold-out set. The artefact removal model produces lower MAE and higher SSIM values as desired, however the improvement is only marginal and not statistically significant. Examining the motion-corrected images in Fig. 9, we observe a noticeable change in appearance as the model attempts to remove the artefacts, however mainly by blurring the image and it is unable to recover image details lost by severe motion artefacts. We find that for many artefacts in the real-world dataset, the artefact correction CNN is unable to completely remove the artefact, particularly in comparison to the model's strong performance on simulated data. When computing the MAE, we discovered that the error is dominated by misalignment between the rigidly registered ground-truth artefact-free images and their corresponding artefact images. This is because it can be difficult to rigidly align images containing very severe artefacts, and so even after an attempted registration, some misalignment of the brains still remains. Therefore, in table I (bottom), we only compute results for images that are sufficiently well-aligned by rejecting image pairs with a before-correction error in the 90th percentile of the data (confirmed as misaligned by visual inspection) such that we can compute a reasonable estimate of the error caused

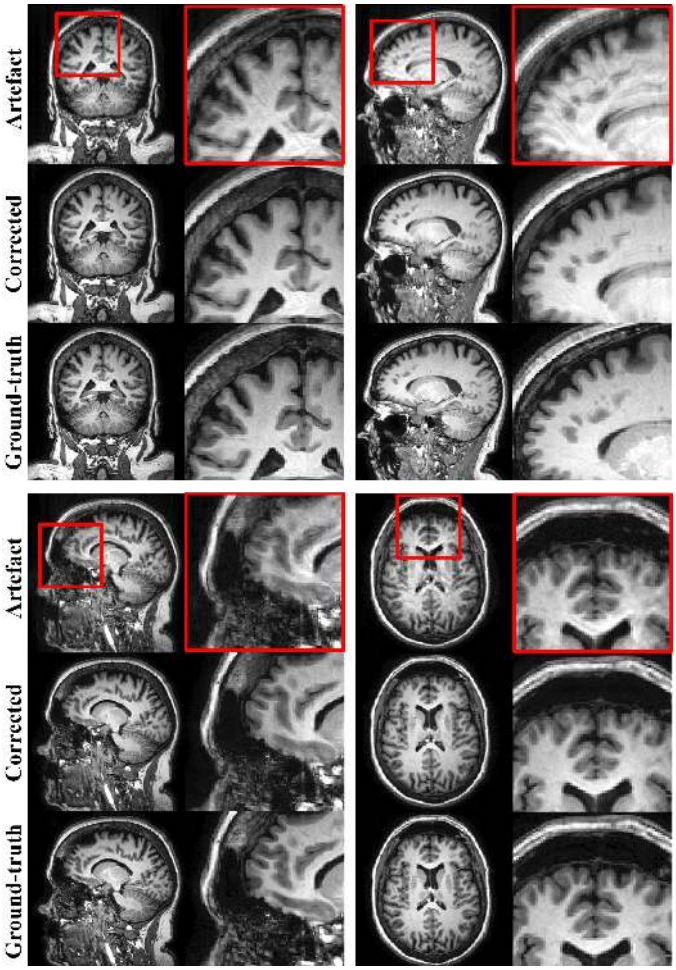


Fig. 8. A sample of artefact removal results for a hold-out synthetic test set. Top row: input artefacted image volumes. Middle row: artefact-corrected images as outputted from the CNN. Bottom row: ground-truth artefact-free images. Best viewed by zooming in on digital copy.

by the artefacts alone. This results in 98 real-world artefacts. We also mask the image volumes by dilated TIV masks, as shown in Fig. 9, to ignore differences outside of the brain.

The fact that, using CNNs, we are data and GPU memory-bound, means our models have limited capacity. It is likely the network underfits the problem and therefore the choice of network architecture, number of parameters and loss function are key factors in the algorithm's ability to remove specific types of artefacts. Currently, we have chosen a generic form of motion model to cover many and varied types of movement artefacts using a single CNN, but one could tailor the sampling strategy to focus on specific artefact sub-types, such as "ringing" or "ghosting" artefacts. We can, however, note that because of the limited effective receptive field of the proposed network, spatially-limited artefacts such as "ringing", are more easily removed than global artefact such as "ghosting".

Considering the reconstruction results from the artefact removal CNN, we have seen that real-world artefacts can be sometimes only partially removed from images. Consequently, in an effort to illustrate that it is indeed better to learn the segmentation task end-to-end, in Fig. 10 we estimate the uncertainty of the CGM segmentation given the motion-corrected images as input to the network. Using the same dropout method as discussed in Section V, the artefact-corrected images are passed through the segmentation CNN model that has been trained with only classical augmentations (rotation, translation, scaling) since if the motion artefact has been successfully removed from the image then only classical augmentations should be required to account for image appearance variability. The model uncertainty given the artefact-corrected image is estimated from multiple forward passes of the network and is shown in Fig. 10 by the plots of mean and standard deviation (std) over the segmentation predictions. For comparison, we also show the uncertainty of the motion augmented model given the uncorrected image as input. In Fig. 10 we observe that the variance of the classically augmented model output given the motion-corrected image is generally of similar value to the model's variance given the ground-truth artefact-free image, i.e. the classically augmented model is similarly confident in its predictions on the ground-truth artefact-free image as it is given the motion-corrected image. However, the KL-divergence (KLD) between the output distributions of the two images of the classically augmented model is very high, as shown by the regions of high KLD in the KLD map on the bottom row of Fig. 10. This is much higher than the corresponding KLD of the motion augmented model. These regions of high KLD imply that the classically augmented model makes predictions that are very different between the two input images, even though the artefact should have been removed by the artefact-correction CNN if it was successful. Given that the ground-truth image is the correct one, the model's prediction on the motion-

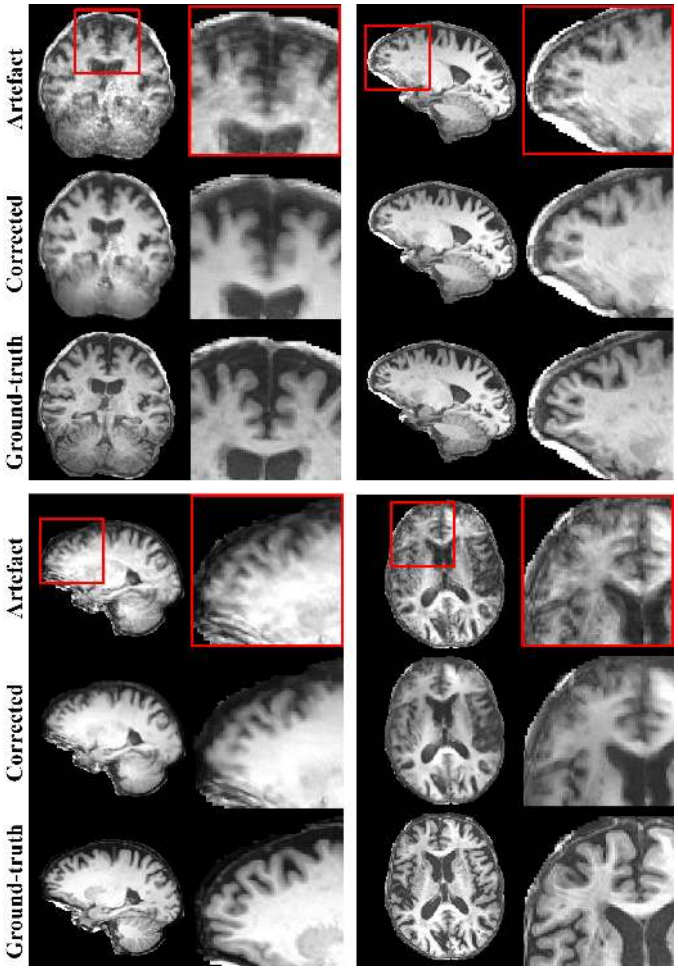


Fig. 9. A sample of artefact removal results for an unseen real-world test set. Top row: input artefacted image volumes. Middle row: corrected image from the output of the CNN model. Bottom row: ground-truth artefact-free images. Best viewed by zooming in on digital copy.

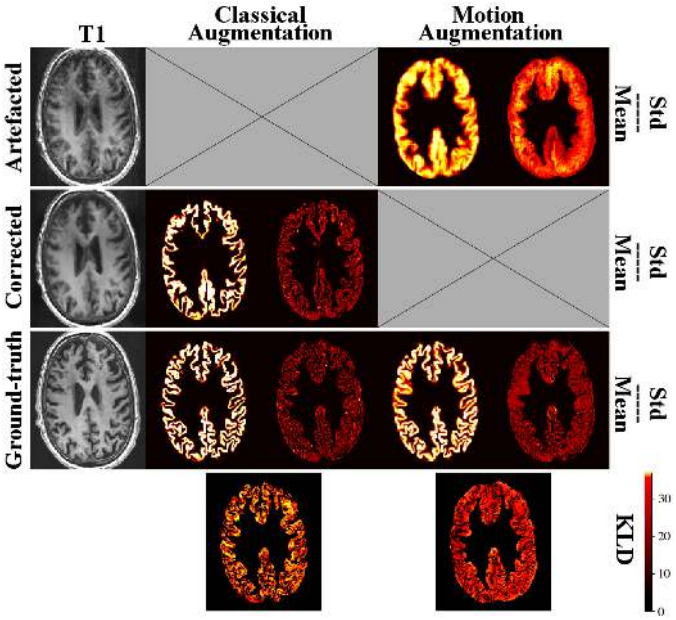


Fig. 10. Per-voxel mean and uncertainty estimations on the CGM segmentation task for classically augmented and motion augmented models. Results are shown for a test-retest pair for which one scan is heavily artefacted and for the corresponding motion-corrected image, as output by the artefact removal CNN. Bottom row: KL-divergence (KLD) between the probability distributions produced by each model.

corrected image must be inaccurate, despite the fact that the artefact has been corrected for. In comparison, the motion augmented model produces output distributions with generally lower KLD overall, even when the input to the network is the original artefact image, and has not been motion-corrected. This implies that the motion augmented model gives more statistically similar results on artefacted images as it does on artefact-free images, and is, therefore, able to more robustly perform the CGM segmentation task given artefacted input data in comparison to trying to remove the artefact first and then attempting the segmentation using a model that has not been trained with artefact augmentation but only classical augmentations.

VII. Discussion

Considering the results on data with synthetic artefacts, in the tasks of CGM, hippocampus and TIV segmentation, models trained with motion artefact augmentation perform generally better than models without any augmentation or with only classical augmentation (rotation, translation and scaling). For CGM and TIV segmentation, in terms of Dice, PPV and sensitivity metrics, the model that observes the most artefacts during training (50% artefacts) consistently performs significantly better than the others, with a lower result variance. For the hippocampus, the benefit of artefact augmentation is less clear. This is likely related to the location of the object (medial brain), thus being less affected by extra-cranial fat ringing artefacts. For example, ringing artefacts mainly impact the cortical surface, and ghosting typically affects the TIV. For the average distance metric, the classically augmented model appears to perform better on CGM and hippocampus, whereas for TIV the artefact model is statistically significantly better.

On real-world data we observe a similar benefit to performance when training with simulated data. In terms of Dice score coefficient, PPV and sensitivity, the motion augmented models mostly perform better. This suggests the proposed motion artefact generation is realistic and contributes to increased robustness to artefacts of models trained with this augmentation. Additionally, it appears that the larger the artefactual variability encountered at training the better the performance of the model.

Limitations

Although artefact augmentation shows promising results for segmentation, there are limitations with the proposed model:

First, our motion augmentation model uses only magnitude images as input to enable its use on a wide variety of input data and supervision problems. Without phase information, generated artefacts are an approximation to the true artefact appearance. Phase information, if available, should be incorporated into the model to improve the realism of generated artefact patterns. This could potentially lead to bias in the neural network outputs, but, as shown through real-world data experiments, even without phase information we observe improved robustness and generalisation to real-world artefacts.

Second, the augmentation model assumes that a valid segmentation exists, but this may not always be true. With heavy artefacts caused by large movements, it is difficult to say with certainty where in space the true segmentation should be. If the subject's head was in one place for 50% of the scan and in another position for the remaining time, where should the ground-truth segmentation be located? In this case, an uncertain segmentation is the only hypothetically correct answer.

Third, our CNN models are parameter-deprived due to memory constraints, as training with artefacts sometimes decreases inference performance on clean data. Note, however, that this drop in performance on clean data is not statistically significant, while often providing significant improvements on artefact data. While performance is a key goal, robustness to data artefacts is paramount to enable the safe clinical translation of such technique.

Fourth, our motion model is randomly sampled from PDFs, but human motion in MRI is not completely random and certain motions are more common, e.g. nodding when the patient swallows, and nor does it capture non-rigid motions. Additionally, motion is much more likely to appear between repetitions, and not within a short echo readout. Therefore the appearance and distribution of artefacts in our simulated dataset is not entirely representative of the distribution of observed real-world artefacts. With more consideration of the types of movements that occur, adaptation of the model could see a potential further increase in performance on real data.

Overall, we must appreciate that the MR imaging process is complex, with many moving parts, making it difficult to simulate accurately. Coil sensitivities/arrangements/count, sequence timings (TR/TE), gradient ramp-up and cooling times, k-space trajectory, B1/B0 inhomogeneities, patient spatial location, local magnetisation transfer effects etc. are all factors that affect artefact appearance. It would be unrealistic, however, given the volume and variability of data that is necessary to train deep learning systems to have a simulation system that encompassed all these factors. The proposed model is only approximate, and any attempt at generating such a simulator would most likely also be approximate even with more complex modelling and at the expense of increased computational time. The question that remains is: what is the sufficient amount of realism that transfers synthetic artefact patterns to real-world data? We have demonstrated in real-world experiments that, while being a rough approximation, the proposed model does confer the robustness that we expect, while ensuring the simulation is scalable and fast, and the data is large and varied.

VIII. Conclusion

Our main contributions are threefold. Firstly, we propose a realistic, fully 3D, motion model of MRI acquisitions to augment training data, improving the performance and robustness of semantic segmentation CNNs to real-world artefacts. Training on simulated artefacts has been shown to successfully translate to improved performance on real-world artefacts, while the performance on artefact-free data is largely unaffected by the use of augmented data during training. Secondly, by training the different tasks end-to-end with motion augmentation, a new internal data representation

is created allowing the model to become robust to the presence of artefact, instead of requiring an explicit intermediate step of artefact removal likely to destroy important image information. Lastly, our augmentation model provides more calibrated and informative uncertainty estimates for segmentation predictions in the presence of real-world motion-corrupted data. This is of utmost importance when addressing the question of safe clinical translation of such models.

What humans deem acceptable scan quality for radiological assessment is different to the quality required for automated analysis. With this in mind, we observe that scan quality is intrinsically related to the task being solved. This observation, as opposed to a human-perceived notion of image-wide scan quality, is a concept rarely recognised by machine learning researchers, systems and datasets.

ACKNOWLEDGMENT

R. Shaw is funded by an EPSRC CASE studentship. C. H. Sudre was funded by an Alzheimer's Society Fellowship (AS-JF-17-011). M. J. Cardoso was funded by the Wellcome Flagship Programme (WT213038/Z/18/Z)" and "Wellcome EPSRC CME (WT203148/Z/16/Z) and the NIHR GSTT BRC. We gratefully acknowledge NVIDIA corporation for the donation of the GPU that was used in the preparation of this work. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research Development, LLC.; Johnson Johnson Pharmaceutical Research Development LLC.; Lumosity; Lundbeck; Merck Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

REFERENCES

[1] M. L. Wood and R. M. Henkelman, "MR image artifacts from periodic motion," Medical Physics, 12(2), 143-151, 1985.

[2] A. Alexander-Bloch, L. S. Clasen, M. G. Stockman, L. H. Ronan, F. M. Lalonde, J. N. Giedd, et al., "Subtle in-scanner motion biases automated measurement of brain anatomy from in vivo MRI," Human Brain Mapping, 37(7), 2385-2397, 2016.

[3] G. R. Wylie, H. M. Genova, J. DeLuca, N. D. Chiaravalloti, and J. F. Sumowski, "Functional magnetic resonance imaging movers and shakers: does subject-movement cause sampling bias?," Human Brain Mapping, 35(1), 1-13, 2014.

[4] M. Zaitsev, J. Maclaren, and M. Herbst, "Motion artifacts in MRI: A complex problem with many partial solutions," Journal of Magnetic Resonance Imaging, 42(4), 887-901, 2015.

[5] M. Usman, D. Atkinson, F. Odille, C. Kolbitsch, G. Vaillant, T. Schaeffter, et al., "Motion corrected compressed sensing for free-breathing dynamic cardiac MRI," Magnetic Resonance in Medicine," 70(2), 504-16, 2013.

[6] F. Godenschweger, U. Kägebein, D. Stucht, U. Yarach, A. Sciarra, R. Yakupov, et al., "Motion correction in MRI of the brain," Physics in Medicine and Biology, 61(5), R32-56, 2016.

[7] D. Atkinson, D. L. G. Hill, P. Stoyle, P. E. Summers, S. Clare, R. Bowtell, et al., "Automatic compensation of motion artifacts in MRI," Magnetic Resonance in Medicine, 41(1), 163-170, 1999.

[8] M. Medley, H. Yan, and D. Rosenfeld, "An improved algorithm for 2-D translational motion artifact correction," IEEE Transactions on Medical Imaging, 10(4), 548-553, 1991.

[9] P. J. Bones, J. R. Maclaren, R. P. Millane, and R. Watts, "Quantifying and correcting motion artifacts in MRI," Proceedings of SPIE, 6316, 2006.

[10] A. Loktyushin, H. Nickisch, R. Pohmann, and Bernhard Schölkopf, "Blind retrospective motion correction of MR images," Magnetic Resonance in Medicine, 70(6), 1608-1618, 2012.

[11] O. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," MICCAI, 2016.

[12] M. Usman, D. Atkinson, F. Odille, C. Kolbitsch, G. Vaillant, T. Schaeffter, et al., "Motion corrected compressed sensing for free-breathing dynamic cardiac MRI," Magnetic Resonance in Medicine, 70(2), 504-516, 2013.

[13] K. Meding, A. Loktyushin and M. Hirsch, "Automatic detection of motion artifacts in MR images using CNNS," IEEE International Conference on Acoustics, Speech and Signal Processing, 811-815, 2017.

[14] B. A. Duffy, W. Zhang, H. Tang, L. Zhao, M. Law, A. W. Toga, et al., "Retrospective correction of motion artifact affected structural MRI images using deep learning of simulated motion," MIDL, 2018.

[15] K. Armanious, C. Yang, M. Fischer, T. Küstner, K. Nikolaou, S. Gatidis, et al., "MedGAN: Medical Image Translation using GANs," International Conference on Learning Representations, 2018.

[16] K. Pawar, Z. Chen, N. J. Shah, and G. F. Egan, "MoCoNet: Motion Correction in 3D MPRAGE images using a Convolutional Neural Network approach," International Conference on Learning Representations, 2018.

[17] R. Shaw, C. H. Sudre, S. Ourselin, and M. J. Cardoso, "MRI k-Space Motion Artefact Augmentation: Model Robustness and Task-Specific Uncertainty," Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning, 102, 427-436, 2019.

[18] M. Alexa, "Linear combination of transformations," ACM Transactions on Graphics, 21(3), 380-387, 2002.

[19] W. Li, 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", IPMI, 2017.

[20] E. Gibson, W. Li, C. H. Sudre, L. Fidon, D. I. Shakir, G. Wang, et al., "NiftyNet: a deep-learning platform for medical imaging," Computer Methods and Programs in Biomedicine, 2018.

[21] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations," Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, 240-248, 2017.

[22] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," International Conference on Learning Representations, 2014.

[23] M. J. Cardoso, M. Modat, R. Wolz, A. Melbourne, D. Cash, D. Rueckert, et al., "Geodesic information flows: spatially-variant graphs and their application to segmentation and fusion," IEEE Transactions on Medical Imaging, 34(9), 1976-1988, 2015.

[24] Y. Gal and Z. Ghahramani, "Dropout As a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," Proceedings of the 33rd International Conference on International Conference on Machine Learning, 48, 1050-1059, 2016.

[25] Z. Eaton-Rosen, F. J. S. Bragman, S. Bisdas, S. Ourselin, and M. J. Cardoso, "Towards safe deep learning: accurately quantifying biomarker uncertainty in neural network predictions," MICCAI, 2018.

APPENDIX
TEST-RETEST IMAGES