# MRI k-Space Motion Artefact Augmentation: Model Robustness and Task-Specific Uncertainty

**Richard Shaw**[1,2] RICHARD.SHAW.17@UCL.AC.UK
**Carole Sudre**[2,1,3] CAROLE.SUDRE@KCL.AC.UK
**Sebastien Ourselin**[2] SEBASTIEN.OURSELIN@KCL.AC.UK
**M. Jorge Cardoso**[2] M.JORGE.CARDOSO@KCL.AC.UK
[1] *Department of Medical Physics and Biomedical Engineering, University College London, UK*
[2] *School of Biomedical Engineering and Imaging Sciences, King's College London, UK*
[3] *Dementia Research Centre, Institute of Neurology, UCL, UK*

## Abstract

Patient movement during the acquisition of magnetic resonance images (MRI) can cause unwanted image artefacts. These artefacts may affect the quality of diagnosis by clinicians and cause errors in automated image analysis. In this work, we present a method for generating realistic motion artefacts from artefact-free data to be used in deep learning frameworks to increase training appearance variability and ultimately make machine learning algorithms such as convolutional neural networks (CNNs) robust to the presence of motion artefacts. We model patient movement as a sequence of randomly-generated, 'de-meaned', rigid 3D affine transforms which, by resampling artefact-free volumes, are then combined in k-space to generate realistic motion artefacts. We show that by augmenting the training of semantic segmentation CNNs with artefacted data, we can train models that generalise better and perform more reliably in the presence of artefacted data, with negligible cost to their performance on artefact-free data. We show that the performance of models trained using artefacted data on segmentation tasks on real-world test-retest image pairs is more robust. Finally, we demonstrate that measures of uncertainty obtained from motion augmented models reflect the presence of artefacts and can thus provide relevant information to ensure the safe usage of deep learning extracted biomarkers in clinics.

## 1. 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 (Wood and Henkelman, 1985). Motion artefacts can affect the interpretability of images, potentially affecting the quality of diagnosis by clinicians 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. Additionally, 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 (Wylie et al., 2014). Even if included, biomarker measures may be biased by artefacts leading to spurious findings (Alexander-Bloch et al., 2016).

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 scan trajectory. Movements at the k-space centre

correspond to low image frequencies and result in ghosting artefacts, where the image is repeated, as does quasi-periodic motion e.g. respiration (Zaitsev et al., 2015). Movements toward the edges of the k-space corresponding to the acquisition of high image frequencies, often result in ringing artefacts. Most commonly observed MR motion artefacts result in 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 mostly attempted to design ways of correcting for them (Usman et al., 2013). This work 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 networks potential for generalisation (Çiçek et al., 2016). While classical data augmentation usually involves random geometric transformations and/or intensity changes, biologically and physically plausible augmentation models would be beneficial.

Some prior work has been done in this domain. Meding et al. (Meding et al., 2017) used convolutional neural networks to classify MR magnitude images as artefacted or not. Going beyond the binary classification task, Duffy et al. (Duffy et al., 2018), also using CNNs, attempt to learn how to retrospectively remove artefacts from MR images. Their network, trained on synthetic data proposes an unrealistic motion model that is limited to axial translation. Using a Generative Adversarial Network, Armanious et al. propose MedGAN (Armanious et al., 2018) with the objective of "translating" motion-corrupted MR images to their corresponding motion-free images, but restrict their work to 2D slices. Pawar et al. (Pawar et al., 2018), with the objective of learning to remove artefacts, model 3D motion in the image domain and reconstruct 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 clean 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 present artefacts thus resulting in data that is unusable for further processing. Additionally, end-to-end learning allows for model artefact-induced task uncertainty to be learned directly from raw artefacted inputs.

## 2. Motion Artefact Model

Our proposed method for generating motion artefacts is illustrated in Figure 1. The procedure is summarised by the following steps: (1) Generate a random movement model by sampling different probability distribution functions. (2) De-mean the generated movement transforms. (3) Resample the artefact-free volume according to the de-meaned movement model. (4) Reconstruct a composite k-space from the k-spaces of multiple resampled volumes. (5) Transform the k-space to the image domain to produce final artefacted sample.

Taking each step in turn, we first sample movements from PDFs, modelling head motion as a sequence of independent small and large motions (e.g. twitches/nodding). We sample $N$ movements from a Poisson distribution - small movements are assumed to occur more often and large move-

ments less frequently - and the time $t$ in k-space at which each movement occurs from a uniform distribution (assuming k-space scans in the phase encoding direction). Modelling each movement as a 3D affine $A$ matrix comprising a rigid 3D rotation and translation in the image domain, the rotation is sampled from between $(-30°, 30°)$ and the translation between (-10mm, 10mm) in all three axes. The sequence of movement transforms $\{A\}_{i=0}^{N}$ is combined incrementally in log-Euclidean space (Alexa, 2002), allowing for the linear combination of transforms. To apply affine matrix $A$ in log-Euclidean space, we use the matrix exponential $\exp_M(A) = \sum_{k=0}^{\infty} \frac{A^k}{k!}$, and corresponding matrix logarithm.

With the motion model defined, the second step is to 'de-mean' the movements. When applying our augmentation model to a clean MR volume $I_0$, we expect the barycenter of the imaged object to remain in approximately the same position within the 3D volume. This is achieved by 'de-meaning' each affine transform $A_i$ by pre-multiplying by the inverse of the average transform $A_{avg}$, computed as the weighted sum of the sequence $N$ affine transformations in log-Euclidean space, given by Equation (1),

$$A_{avg} = \exp_M(\sum_{i=0}^{N} \hat{w}_i \log_M(A_i)), \tag{1}$$

where $\hat{w}_i$ is a 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 of the $i$-th movement, transforming back to the image domain, and summing the resulting voxel intensities, as given by $w_i = \sum_{voxels} \mathscr{F}^{-1}(M_i \odot \mathscr{F}(I_0))$, with the weights then normalised to sum to 1.

The third step is to apply each 'de-meaned' affine transform to the original artefact-free image volume $I_0$ and resample using b-spline interpolation. We always resample the original image to reduce propagating interpolation errors and we apply edge-padding to mitigate edge effects. Fol-

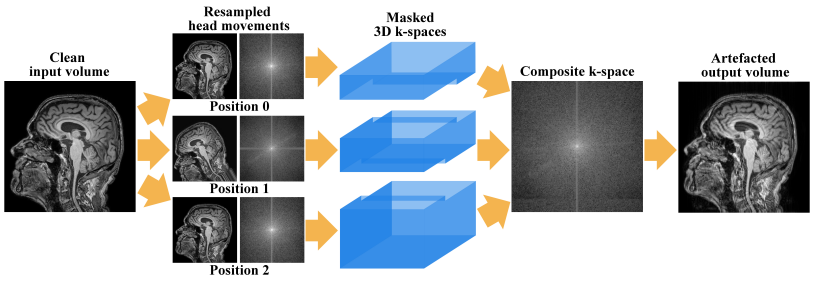


Figure 1: Motion artefact augmentation framework: The artefact-free volume is resampled according to a randomly sampled movement model, defined by a sequence of 'de-meaned' 3D affine transforms. Their 3D Fourier transforms are combined to form a composite k-space, which is transformed back to the image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

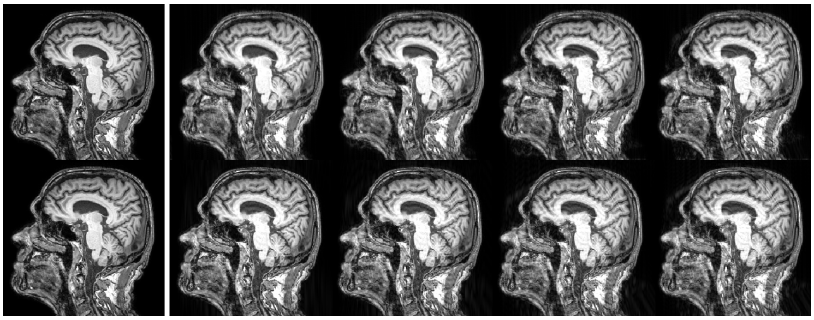

Figure 2: Simulated motion artefacts due to nodding motion. Top row: artefacts due to increasing amounts of movement from left to right. Bottom row: artefacts due to changing the time at which the movement occurs, earlier in the k-space scan trajectory from left to right, therefore retaining higher image frequencies. Best viewed zoomed in on digital copy.

lowing each transformation, we compute the 3D Fourier Transform of each resampled image. 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. Finally, the inverse 3D Fourier Transform of the composite k-space is derived, and the magnitude image provides the final artefacted sample. Examples of our artefact augmentation are shown in Figure 2.

## 3. 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).

### 3.1. Network architecture and implementation details

We used the HighResNet (Li et al., 2017) architecture implemented in NiftyNet (Gibson et al., 2018), with Dice loss (Sudre et al., 2017), patch size of $80^3$ and batch size 1, trained on a single GPU with Adam optimiser (Kingma and Ba, 2014) 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 training performance, so we use weighted patch sampling with higher weight at regions defined by the blurred ground-truth label, 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 reaching 100,000 iterations.

### 3.2. Simulated Dataset

For experiments on simulated data, we use 272 MPRAGE scans from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and generate 15 artefacted volumes per scan. The data was split into

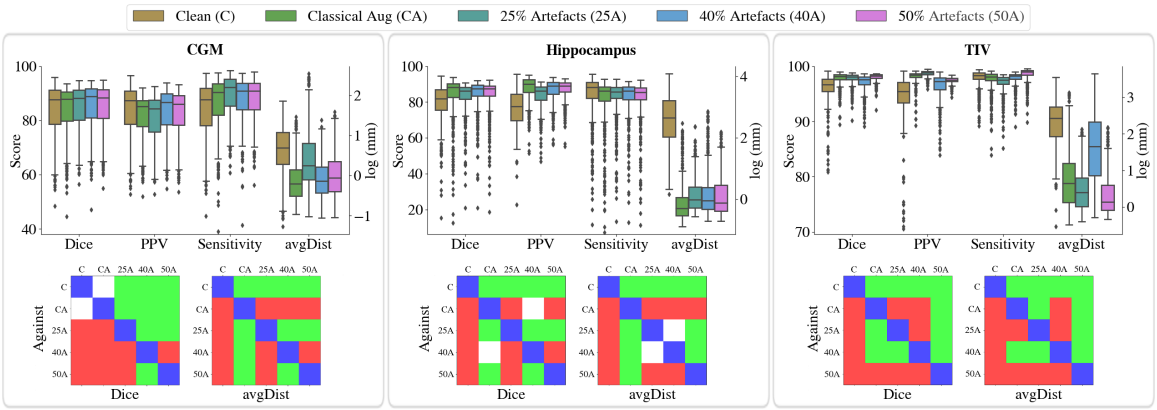


Figure 3: Segmentation results on simulated data for CGM, hippocampus and TIV across models. Top: Boxplots of different error metrics for 5 models with different augmentation: clean, classically augmented, 25%, 40%, 50% artefacts. Bottom: Bonferroni corrected pairwise Wilcoxon tests for Dice and average distance between column and row models - Green: significantly better, White: no statistically significant difference; Red: significantly worse.

80% training, 10% validation and 10% testing and 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 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 these three tasks across all models is evaluated with Dice score, positive predictive value, sensitivity and average distance metrics and presented in the first row of Figure 3. Results of Bonferroni corrected matched pair Wilcoxon tests between models are presented on the bottom row.

### 3.3. Real-world Dataset

Robustness of CNNs trained with the proposed motion augmentation to real-world motion artefacts was then evaluated in a test-retest setting. 106 quality-controlled pairs of MPRAGE test-retest images from the ADNI dataset on which only one of the images was considered artefacted were used for this purpose. Image pairs were rigidly registered together in a groupwise space to avoid interpolation bias. For comparison purposes, a benchmark label fusion algorithm was used to perform the segmentation tasks on each pair of images (Cardoso et al., 2015). For each trained model and the benchmark method, Dice score, positive predictive value, sensitivity and average distance were used as evaluation measure between test and retest images, with the results obtained on the clean image being used as reference. Figure 4 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.

## 4. Task-specific Uncertainty Estimation

Deep learning models for segmentation tasks classically provide for each voxel a point-estimate probability of belonging 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 (Gal and Ghahramani, 2016), uncertainty can be estimated by sampling at inference time from multiple outputs of the network trained with dropout. Adapting the approach from (Eaton-Rosen et al., 2018), 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. Mean and variance results obtained on the CGM segmentation task for the aforementioned models trained with dropout are shown in Figure 5. In models trained with motion augmentation, higher variance of predictions are observed in artefacted regions, especially close to the cortical surface.

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 row of Figure 5. 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 Figure 6.

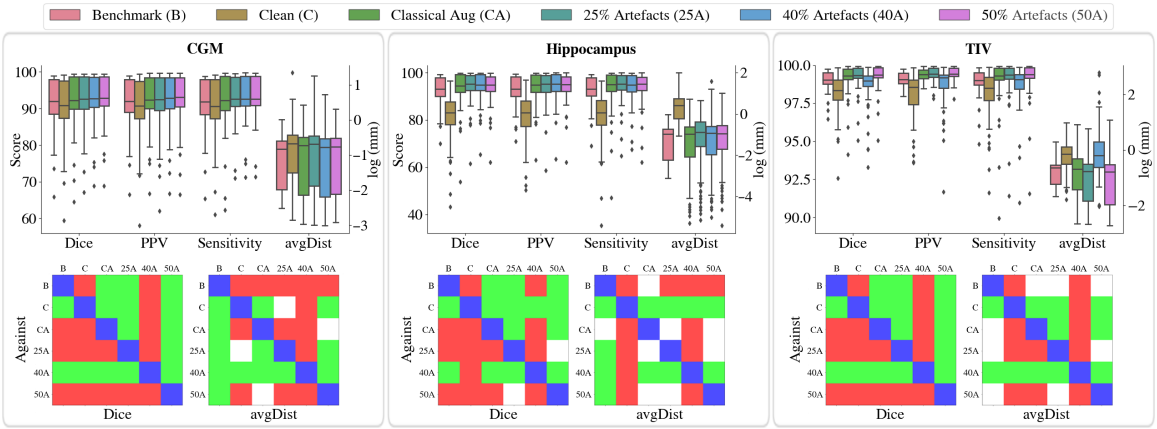


Figure 4: Segmentation robustness on real-world test-retest data for CGM, hippocampus and TIV across models. Top: For each task, boxplots of different error metrics for the 5 models in addition to a benchmark method. Bottom: Bonferroni corrected pairwise Wilcoxon tests for Dice and average distance between column and row models - Green: significantly better, White: no statistically significant difference; Red: significantly worse.

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

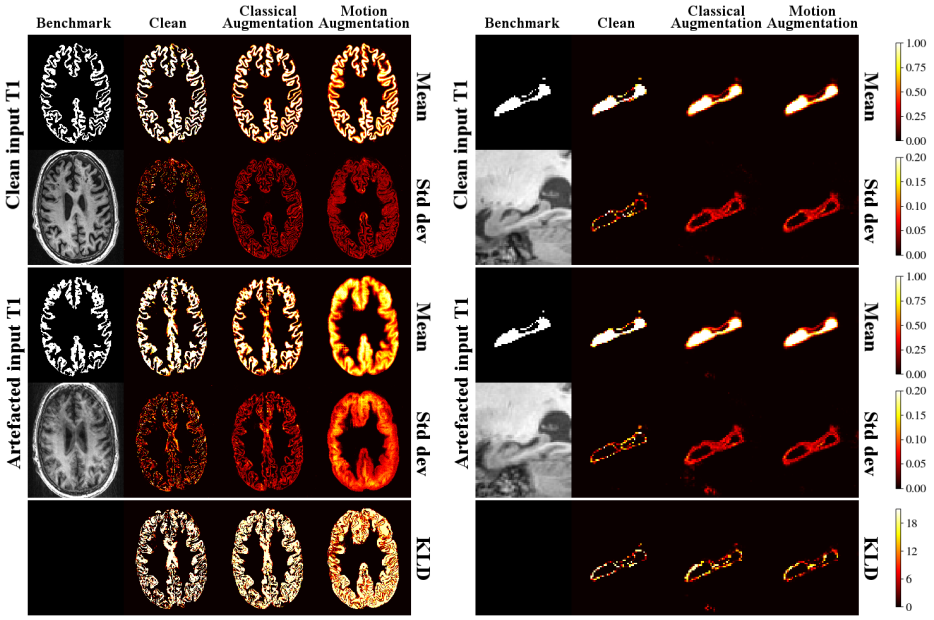


Figure 5: Per-voxel mean and uncertainty estimations on CGM and hippocampus segmentation tasks for clean (no augmentation), classically augmented and motion augmented models for a test-retest pair for which one scan is heavily artefacted. The segmentation produced by a benchmark method is shown for reference. Bottom row: KL-divergence (KLD) between the probability distributions produced by each model on clean and artefacted scans.

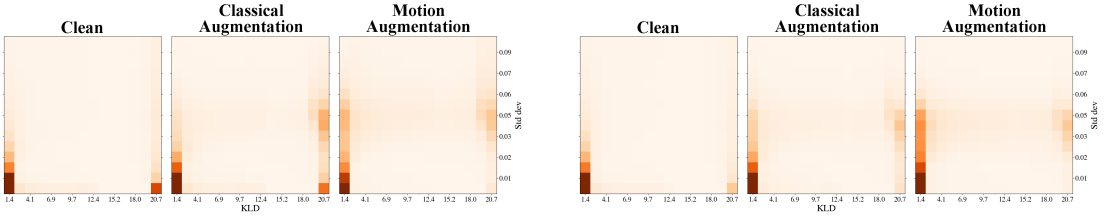


Figure 6: Histograms of per-voxel KLD associated to the uncertainty estimates as measured by the sampled variance, shown for models trained with different augmentations on (right) heavily artefacted and (left) lightly artefacted data.

presence of heavily artefacted images (Figure 6 left), models trained on clean (artefact-free) 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.

## 5. Discussion and Conclusion

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

First, the augmentation model assumes a valid segmentation exists, but this may not always be true. With heavily artefacted scans caused by extreme movements, it is difficult to say with certainty where 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.

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

Third, 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. Therefore our simulated dataset is not entirely representative of the distribution of observed motion artefacts. With more consideration of the types of movements that occur, adaptation of the model could see potential increased performance on real data.

Notwithstanding these observations, our main contributions are threefold: Firstly, 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 noise, 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.

## References

Marc Alexa. Linear combination of transformations. In *SIGGRAPH*, 2002.

Aaron Alexander-Bloch, Liv S. Clasen, Michael Stockman, Lisa Ronan, Francois M. Lalonde, Jay N. Giedd, and Armin Raznahan. Subtle in-scanner motion biases automated measurement of brain anatomy from in vivo mri. *Human brain mapping*, 37 7:2385–97, 2016.

Karim Armanious, Chenming Yang, Marc Fischer, Thomas Küstner, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. Medgan: Medical image translation using gans. *CoRR*, 2018.

M. Jorge Cardoso, Marc Modat, Robin Wolz, Andrew Melbourne, David M. Cash, Daniel Rueckert, and Sébastien Ourselin. Geodesic information flows: Spatially-variant graphs and their application to segmentation and fusion. *IEEE Transactions on Medical Imaging*, 34:1976–1988, 2015.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In *MICCAI*, 2016.

Ben A. Duffy, Wenlu Zhang, Haoteng Tang, Lu Zhao, Meng Law, Arthur W. Toga, and Hosung Kim. Retrospective correction of motion artifact affected structural mri images using deep learning of simulated motion. *MIDL*, 2018.

Zach Eaton-Rosen, Felix J. S. Bragman, Sotirios Bisdas, Sébastien Ourselin, and M. Jorge Cardoso. Towards safe deep learning: Accurately quantifying biomarker uncertainty in neural network predictions. In *MICCAI*, 2018.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, volume 48 of *ICML*, pages 1050–1059, 2016.

Eli Gibson, Wenqi Li, Carole H. Sudre, Lucas Fidon, Dzhoshkun I. Shakir, Guotai Wang, Zach Eaton-Rosen, Robert Gray, Tom Doel, Yipeng Hu, Tom Whyntie, Parashkev Nachev, Marc Modat, Dean C. Barratt, Sébastien Ourselin, M. Jorge Cardoso, and Tom Vercauteren. Niftynet: a deep-learning platform for medical imaging. *Computer Methods and Programs in Biomedicine*, 158:113–122, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Wenqi Li, Guotai Wang, Lucas Fidon, Sébastien Ourselin, M. Jorge Cardoso, and Tom Vercauteren. On the compactness, efficiency, and representation of 3d convolutional networks: Brain parcellation as a pretext task. In *IPMI*, 2017.

Kristof Meding, Alexander Loktyushin, and Michael Hirsch. Automatic detection of motion artifacts in mr images using cnns. In *42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, pages 811–815, 2017.

Kamlesh Pawar, Zhaolin Chen, N Jon Shah, and Gary Egan. Moconet: Motion correction in 3d mprage images using a convolutional neural network approach. *arXiv e-prints*, art. arXiv:1807.10831, 2018.

Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *DLMIA/ML-CDS@MICCAI*, 2017.

Muhammad Usman, David Atkinson, Freddy Odille, Christoph Kolbitsch, Ghislain Vaillant, Tobias Schaeffter, Philip G. Batchelor, and Claudia Prieto. Motion corrected compressed sensing for free-breathing dynamic cardiac mri. *Magnetic resonance in medicine*, 70 2:504–16, 2013.

Michael L. Wood and Mark Henkelman. Mr image artifacts from periodic motion. *Medical Physics*, 12(2):143–151, 1985.

Glenn R. Wylie, Helen M. Genova, John DeLuca, Nancy D. Chiaravalloti, and James F. Sumowski. Functional magnetic resonance imaging movers and shakers: does subject-movement cause sampling bias? *Human brain mapping*, 35 1:1–13, 2014.

Maxim Zaitsev, J. Piers Maclaren, and Michael F. Herbst. Motion artifacts in mri: A complex problem with many partial solutions. *Journal of magnetic resonance imaging : JMRI*, 42(4): 887–901, 2015.