Combined Denoising and Suppression of Transient Artifacts in Arterial Spin Labeling MRI Using Deep Learning

Patrick W. Hales, PhD,[1*] Josef Pfeuffer, PhD,[2] and Chris A. Clark, PhD[1]

Background: Arterial spin labeling (ASL) is a useful tool for measuring cerebral blood flow (CBF). However, due to the low signal-to-noise ratio (SNR) of the technique, multiple repetitions are required, which results in prolonged scan times and increased susceptibility to artifacts.
Purpose: To develop a deep-learning-based algorithm for simultaneous denoising and suppression of transient artifacts in ASL images.
Study Type: Retrospective.
Subjects: 131 pediatric neuro-oncology patients for model training and 11 healthy adult subjects for model evaluation.
Field Strength/Sequence: 3T / pseudo-continuous and pulsed ASL with 3D gradient-and-spin-echo readout.
Assessment: A denoising autoencoder (DAE) model was designed with stacked encoding/decoding convolutional layers. Reference standard images were generated by averaging 10 pairwise ASL subtraction images. The model was trained to produce perfusion images of a similar quality using a single subtraction image. Performance was compared against Gaussian and non-local means (NLM) filters. Evaluation metrics included SNR, peak SNR (PSNR), and structural similarity index (SSIM) of the CBF images, compared to the reference standard.
Statistical Tests: One-way analysis of variance (ANOVA) tests for group comparisons.
Results: The DAE model was the only model to produce a significant increase in SNR compared to the raw images ($P < 0.05$), providing an average SNR gain of 62%. The DAE model was also effective at suppressing transient artifacts, and was the only model to show a significant improvement in accuracy in the generated CBF images, as assessed using PSNR values ($P < 0.05$). In addition, using data from multiple inflow time acquisitions, the DAE images produced the best fit to the Buxton kinetic model, offering a 75% reduction in the fitting error compared to the raw images.
Data Conclusion: Deep-learning-based algorithms provide superior accuracy when denoising ASL images, due to their ability to simultaneously increase SNR and suppress artifactual signals in raw ASL images.
Level of Evidence: 3
Technical Efficacy Stage: 1

J. MAGN. RESON. IMAGING 2020.

ARTERIAL SPIN LABELING (ASL) is a non-invasive imaging modality that provides a powerful means of measuring cerebral blood flow (CBF).[1,2] One of the key advantages of ASL is that it utilizes water in the blood as an endogenous tracer to measure perfusion, eliminating the need for an exogenous contrast agent. Instead, a series of label (L) and control images (C) are acquired, in which inflowing blood–water proximal to the imaging volume is "tagged" using RF pulses during the label acquisition. After a delay, to allow labeled blood to flow into the tissue (the postlabeling delay, PLD), the perfusion signal (dM) is determined by the pairwise subtraction of control and label images (dM = C – L). These are subsequently converted into CBF images, in

View this article online at wileyonlinelibrary.com. DOI: 10.1002/jmri.27255

Received Mar 23, 2020, Accepted for publication May 28, 2020.

*Address reprint requests to: P.W.H., Developmental Imaging & Biophysics Section, UCL Great Ormond Street Institute of Child Health, 30 Guilford Street, London WC1N 1EH, UK. E-mail: p.hales@ucl.ac.uk
Contract grant sponsor: Children with Cancer UK; Contract grant number CwCUK-15-203.
Level of EvidenceTechnical Efficacy Stage

From the [1]Developmental Imaging & Biophysics Section, UCL Great Ormond Street Institute of Child Health, London, UK; and [2]MR Application Development, Siemens Healthcare GmbH, Erlangen, Germany

physiological units of ml / 100 g / minutes, using the method described previously.[3]

Due to the quantitative nature of the technique, the lack of exposure to ionizing radiation, and the avoidance of a contrast agent injection, ASL has excellent clinical potential. However, an inherent limitation is the comparatively low signal-to-noise ratio (SNR) of the technique. This is due to a number of factors. Firstly, T_1 recovery of the tagged bolus during the PLD reduces the signal available from the tracer itself. In addition, in normal gray matter, perfusion replaces only ~1% of the brain water with in-flowing blood-water every second.[3] As such, inflowing blood can only perturb a very small fraction of the total magnetization in a typical voxel, and unwanted signal fluctuations in the static tissue can easily outweigh the perfusion signal. To counteract this, background suppression pulses are often used to null the signal from the static tissue prior to image acquisition.[4] Nonetheless, generally multiple repetitions of an ASL acquisition must be acquired in order to provide sufficient SNR, which leads to increased scan times.

In addition to the inherent limitations in SNR, ASL images can be corrupted by a number of artifacts.[5,6] Some of these are related to the acquisition protocol and/or the physiology of the subject, such as arterial transit time artifacts resulting from an insufficiently long PLD.[6] Others are transient, and may occur sporadically during the series of repetitions. These can include artifacts related to subject motion,[6] and cerebrospinal fluid (CSF) "shine-through" in the ventricles due to RF instabilities.[5,7] As these artifacts typically occur in only a small number of the total repetitions, their impact is less conspicuous after signal averaging (Fig. 1).

A number of postprocessing techniques have been investigated to improve image quality in ASL data. Techniques to suppress transient artifacts include outlier rejection to remove hardware instabilities and motion-corrupted repetitions,[7–11] physiological noise correction,[12] and temporal filtering techniques.[13–15] Techniques to increase SNR have focused on established image denoising techniques. These include Gaussian smoothing,[13,16,17] which provides a simple method for increasing SNR in noisy data, albeit at the cost of a loss of sharpness in the resulting image. More complex methods include techniques such as non-local means (NLM) filtering.[18] The principle of NLM is to average the value of a given voxel with values of other voxels in a limited neighborhood, provided that the patches centered on the other voxels are similar enough to the patch centered on the voxel of interest. This provides effective image denoising, while potentially preserving fine structures and details in the image. Additional denoising strategies have also shown promising results, such as wavelet-based techniques,[13,19] Wiener filters,[13] adaptive filters,[13] and total generalized variation regularization.[15]

In recent years, deep learning has emerged as a powerful tool for image processing and reconstruction. Within this field, convolutional neural networks (CNNs) have become a popular choice for processing imaging data, due to their ability to learn important features of images in a translationally invariant way. These techniques have been successfully applied to image denoising,[20–22] and several studies have applied deep-learning approaches to improving SNR in ASL images.[23–28] Kim et al developed a denoising CNN with two pathways, for extracting local low-level features and large-scale global features in parallel.[23] This was shown to provide improvements in SNR and CBF accuracy in both single-PLD and Hadamard-encoded multiple-PLD data. Ulas et al. developed a CNN that was trained using a custom loss function, which enforced CBF estimates to be close to model-based reference values.[24] Xie et al. recently developed a model combining dilated convolutions with wide activation residual blocks, which provided improved denoising compared to existing CNN architectures.[26] Owen et al. introduced a joint filtering CNN model, in which maps of the mean and temporal variance of the ASL signal were used as dual inputs, in order to improve SNR and partially suppress transient artifacts.[28] Finally, Gong et al. demonstrated an unsupervised deep-learning-based framework that incorporates a subject's T_1-weighted anatomical image as a structural prior.

The aforementioned studies have shown promising results for denoising low-SNR ASL images. However, previous studies have generally applied averaging over a subset of the total acquired repetitions, in order to generate low-SNR inputs for model training. Some previous studies have also applied motion correction[24,26,28] and Gaussian smoothing[24,26] to the input data. In doing so, the presence of transient artifacts in the input data will be reduced, and the ability of these models to identify and suppress these artifacts may be compromised.

The purpose of this study was to develop a deep-learning-based denoising autoencoder (DAE) model for simultaneous denoising and suppression of transient artifacts in ASL images. We aimed to develop a DAE model that could provide both effective denoising as well as differentiate between abnormal ASL signal associated with pathology, and that associated with transient artifacts, using just a single ASL acquisition (rather than relying on multiple repetitions). Having developed this model, we aimed to evaluate its performance in pseudo-continuous ASL (pCASL) and multiple inflow-time (multi-TI) pulsed-ASL (PASL) data acquired in healthy volunteers.

Materials and Methods

Arterial Spin Labeling Acquisition

All ASL datasets were acquired using a 3 T MRI scanner (Magnetom Prisma, Siemens Healthcare, Erlangen, Germany), equipped with a 20-channel head receive coil. PCASL data were acquired using a prototype sequence (Siemens Healthcare), with background suppression RF pulses and a 3D gradient-and-spin-echo (GRASE) readout. The

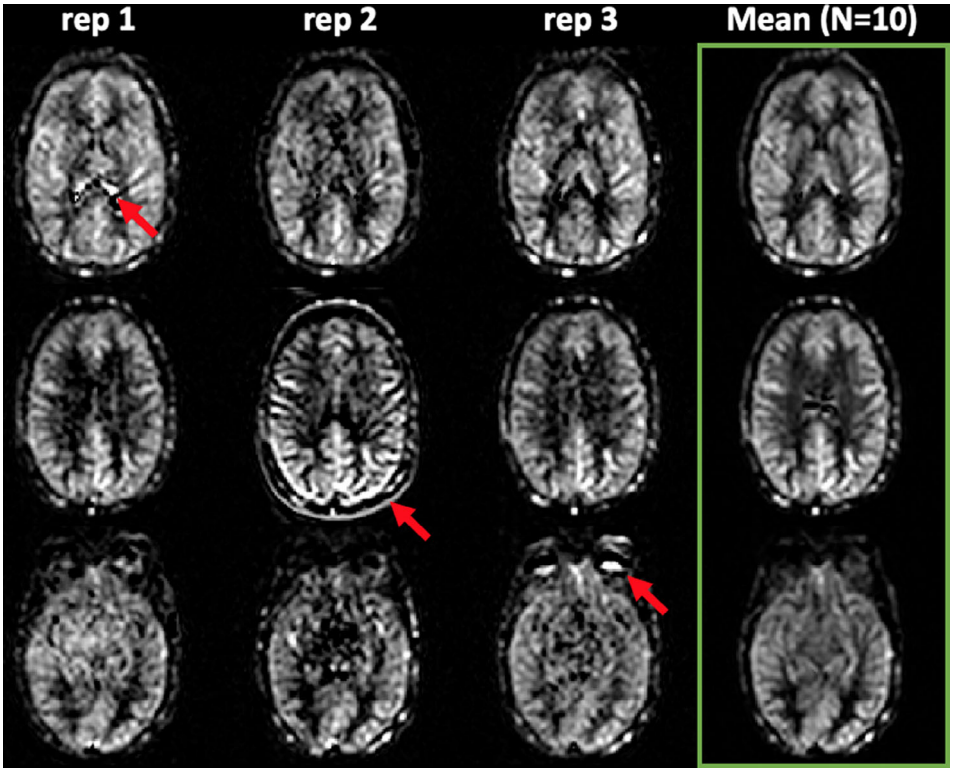

FIGURE 1: Illustration of transient artifacts affecting image quality in ASL datasets. Individual dM images for three repetitions are shown, along with the corresponding image after averaging over 10 repetitions (green box). Individual artifacts are illustrated with red arrows. Top row: CSF shine-through, demonstrating artifactual high signal in the lateral ventricles. Middle row: subject motion artifact, resulting in artifactual signal modulation within the brain, and a peripheral ring of high signal intensity. Bottom row: increased dM signal due to the subject's eye motion. Given the transient nature of these artifacts, their impact is less pronounced after averaging over multiple repetitions (right column). Note, the windowing used here, and in all subsequent dM and CBF images, has a minimum value of zero.

labeling duration was 1800 msec, with a 1500 msec post-labeling delay, and 10 repetitions were acquired. Additional sequence parameters were: relaxation time (TR) = 4620 msec, echo time (TE) = 21.8 msec, field of view = 220 mm, matrix size = 64 x 62, in-plane resolution = 1.7 x 1.7 mm (after zero-filling), number of partitions = 24, slice thickness = 4.0 mm, turbo factor = 12, echo-planar imaging (EPI) factor = 31, segments = 2 (with parallel imaging, generalized autocalibrating partial parallel acquisition [GRAPPA] acceleration factor = 2). A proton-density-weighted (M_0) image was also acquired (TR = 4000 msec), with identical readout to the ASL acquisition but with the labeling and background suppression RF pulses removed, for CBF quantification. Total acquisition time was 3 minutes 19 seconds.

Multi-TI PASL data were acquired using the same prototype sequence. Acquisitions were acquired at 10 TIs, ranging from 350–2600 msec in 250 msec steps, with a single acquisition per TI. The TR was 3300 msec; all other readout parameters were identical to the pCASL acquisition. Q2TIPS RF pulses[29] were applied 700 msec after the labeling pulse to define the temporal width of the bolus. The total acquisition time was 2 minutes 25 seconds.

Training, Validation, and Testing Data

Retrospective anonymized pCASL data were accrued from the clinical database of ASL acquisitions acquired as part of the clinical imaging of pediatric patients at our institution, between 2016–2019. Images that had been severely corrupted due to susceptibility artifacts caused by implants or dental braces, or significant patient motion, had already been excluded prior to entry into the database. Institutional ethical approval with waived consent was granted for retrospective access to this database for this study. The training dataset comprised a cohort of 131 treatment-naïve pediatric neuro-oncology patients (mean age = 7.1, range = 0.4–17.1 years), all of whom received the pCASL acquisition described above as part of their clinical imaging. Following model training, illustrative additional clinical examples from the same database were used as part of the model testing. These included ASL images from a further three neuro-oncology patients (patient #1: 13 years, diffuse astrocytoma; patient #2: 0.9 years, pilocytic astrocytoma; patient #3: 4 years, glioblastoma multiforme), and an additional patient with Sturge–Weber syndrome (patient #4: 11 years). ASL data for patient #1 was acquired at 3T using the protocol described above. ASL data for patients #2 and #3 were acquired with a Siemens Avanto 1.5 T MRI scanner using a similar pCASL protocol to that described above, but with thicker slices (5.0 mm), and no zero-filling. ASL data for patient #4 were acquired at 3 T, again using a similar pCASL protocol to that described above, but with a PLD of 2000 msec.

In order to produce the reference images for each subject, the individual control and label images from all repetitions acquired in

that subject were first coregistered using an affine transformation with 12 degrees of freedom, using the *flirt* algorithm in FSL.[30] This was done to correct any artifacts resulting from subject motion between the control and label acquisitions. The individual difference images (dM) were then calculated for all repetitions, using the motion-corrected control and label images. Following this, the dM values across all repetitions were converted to z-scores, on a voxel-wise basis. Averaging was then performed, and outliers were excluded by only averaging over individual dM values within a voxel with a z-score less than 3.0, to create the final reference image (dM_{mean}). This was done to exclude transient artifacts from the signal averaging, which typically occur in only a small number of repetitions, and can be localized to specific regions of the brain, such as the CSF.[5,7] Lastly, the first and last axial slices were excluded for each subject, to remove registration and wrap-around artifacts from the 3D-GRASE acquisition.

The above steps resulted in a set of 220 "raw" difference images (dM_{raw}) per subject (22 slices x 10 repetitions), in which no motion correction or outlier rejection was applied. Each dM_{raw} image was matched to the corresponding mean image (dM_{mean}), after correction of motion and transient artifacts as described above, which represented the reference standard in this study. Over the entire cohort, this provided a set of 28,820 noisy dM_{raw} images (single repetition), each matched to their corresponding reference standard images. 80% of this dataset was used for model training, with 20% retained for model validation. The mean and standard deviation (SD) of the raw image set were used for Z-normalization of all images before they were entered into the model.

As the training dataset consisted of ASL images acquired in pediatric patients with brain tumors, following training we evaluated the model's performance in healthy adult volunteers, to determine its performance under normal conditions (i.e. adult subjects with no pathology). In addition, the trained model was evaluated using both pCASL data and multi-TI PASL data. To achieve this, new pCASL datasets were acquired in 11 healthy adult subjects (mean age 32 years, range 23–40 years). Additional multi-TI PASL data (using the protocol described above) were acquired in seven healthy subjects (mean age 30 years, range 21–40 years). All subjects provided informed written consent, and institutional ethical approval was granted to use these data.

Model Architecture

A schematic of the DAE model architecture is shown in Fig. 2. The encoder component consisted of three convolution steps. Each convolution step employed 64 filter layers, each of which applied a 3 x 3 kernel, with padding used to maintain consistent image dimensions between the input and output. Following each convolution step, a rectified linear unit (ReLU) activation layer was added, followed by a 2 x 2 max pooling layer, in order to subsample the output by a factor of two. The decoder component mirrored the encoder architecture, with 2 x 2 upsampling used between convolution operations, in order to reconstruct an output image with the same dimensions as the input. The last convolution step consisted of one filter layer only, with no ReLU activation, to produce the final image. Skip connections were added between the first two convolution steps on the encoding side and their counterparts on the decoding side. This allows image details captured in the feature maps from the encoding components to be concatenated with the feature maps produced during decoding, improving image restoration and the ability to train deeper networks.[31]

Model Training

A batch size of 100 was used for model training. We employed the *RMSProp* optimizer with default settings in *Keras* (using the *TensorFlow* backend) to update the network's weights, and the mean squared error (MSE) was used for the loss function. Training was preformed over 100 epochs, with an early-stopping criterion to interrupt the training when the loss in the validation data failed to improve over 10 consecutive epochs. We additionally trained the model using subsets of 25, 50, and 75% of the total training data, in order to investigate the number of training datasets needed to train the model. For each subset, as before, 80% of the data were used for training, with 20% retained for validation. In order to compare the training performance across these subsets in a fair manner, the validation loss function values were normalized to the total number of validation datasets in each subset.

Comparison With Alternative Denoising Methods

Two alternative denoising techniques, Gaussian and NLM filtering, were used to compare the performance of the DAE against more established methods. A subset of 500 training datasets was used to optimize the parameters for these alternative denoising methods, with the filter parameters that gave the minimum root-mean-square error (RMSE) between the denoised and the reference standard images being optimal. For the Gaussian filter, the optimum window size was determined. For the NLM filter, the patch size, patch distance, and cutoff distance were optimized. All filters were applied using Python 3.7: the *cv2* package was used for the Gaussian filter, and the *skimage* package was used for the NLM filter. Multi-parametric optimization of the NLM filter was performed using non-linear least-squares minimization, using the *lmfit* package.

Model Testing

SINGLE PLD PCASL DATA. The pCASL data acquired in 11 healthy subjects was used to test the efficacy of the DAE, Gaussian, and NLM models on un-seen data. The first repetition from each subject's raw dM dataset (using all axial slices) was used as the noisy input to the models (dM_{raw}). The denoised version of this was calculated for each model (dM_{Gauss}, dM_{NLM}, dM_{DAE}), for comparison with the reference standard dM_{mean} images. These dM images were then used to calculate CBF maps for each dataset (CBF_{raw}, CBF_{Gauss}, CBF_{NLM}, CBF_{DAE}, CBF_{mean}), using the standard method described previously,[3] with $\lambda = 0.9$, $\alpha = 0.85$, and $T_{1bl} = 1.65$ s. The CBF_{mean} map was used as the reference standard, against which alternative CBF maps were compared.

MULTI-TI PASL DATA. The performance of each model was also evaluated on un-seen, multi-TI PASL data, acquired in seven healthy subjects as described above. Here, the raw multi-TI difference images were denoised using each model. These datasets were then fit to the Buxton kinetic model,[32] with CBF and bolus arrival time (BAT) as fitted parameters. The temporal width of the bolus was

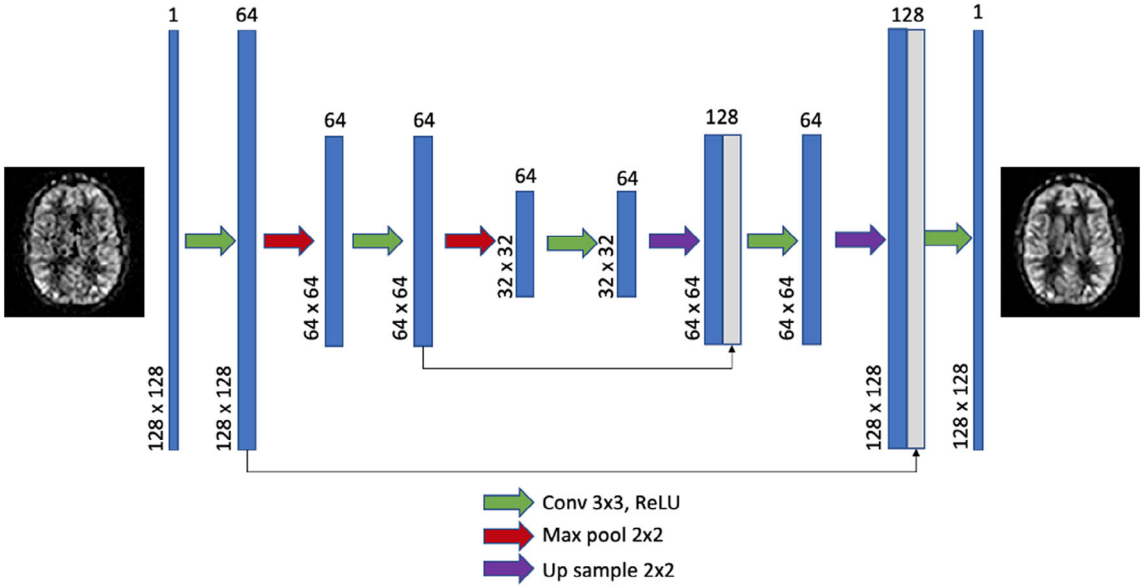

FIGURE 2: Architecture of the denoising autoencoder model, with an example low-SNR, single-repetition dM_{raw} image (left), and the corresponding high-SNR dM_{mean} image (right; same axial slice averaged over 10 repetitions). Image dimensions are shown for each step, along with the number of filter layers used. Skip connections are illustrated as horizontal lines, convolution operations (with subsequent ReLU activation) as green arrows, and max-pooling/upsampling operations as red/purple arrows, respectively.

fixed at 700 msec, due to the use of Q2TIPS saturation pulses during acquisition. Model fitting was performed using the *lmfit* Python package, and the goodness of fit in each voxel was calculated using χ^2 values (sum of squared residuals between the observed and fitted values over all TIs). Models were compared by calculating the mean χ^2 over all brain voxels within each subject.

EVALUATION METRICS. In order to compare the denoising performance of each model, the SNR of each dM dataset was calculated. As the images were acquired using parallel imaging, we used the "difference" method for calculating SNR,[33,34] utilizing the individual ASL repetitions acquired in each subject. First, the *bet* algorithm in FSL[35] was used to define a brain mask for each subject, using the M_0 calibration image, which provided the region of interest (ROI) over which SNR was measured. Following the method described previously,[34] SNR in this ROI was defined using the dM images from two consecutive repetitions (dM_i and dM_{i+1}, where i is the repetition index), using the following relationship:

$$SNR_{ROI} = \frac{(mean(dM_i + dM_{i+1}))_{ROI}}{(\sqrt{2} \cdot std(dM_i - dM_{i+1}))_{ROI}} \quad (1)$$

Equation (1) was applied to all available pairs of dM images across the 10 repetitions, and the mean value of these was taken to represent the final SNR value.

Similar to previous studies,[24–26,28] the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) of the CBF images were used as additional evaluation metrics. PSNR was used to define the accuracy of each CBF image in comparison to the reference standard, and was defined as PSNR = $20 \cdot \log_{10} \cdot (CBF_{max}/RMSE)$. Here, CBF_{max} is the maximum value in the reference standard CBF image, and *RMSE* is the root mean square error between each CBF image and the reference standard (i.e., the average value across all brain voxels of $\sqrt{(CBF_{ref}\text{-}CBF)^2}$. Higher values of PSNR indicate that CBF images are more accurate when compared to the reference standard. SSIM was used to quantify the visual quality of CBF maps in comparison to the reference standard.[36] SSIM is thought to mimic the perceived quality of an image by a human observer, with values of 0 indicating no similarity, and 1.0 indicating perfect similarity. The *skimage* Python package was used to calculate SSIM values, using the default settings. In addition, as denoising methods can often result in increased blur in the resulting image, the level of "focus" in each dataset was quantified using the modified Laplacian method.[37,38] Here, the mean value of the dM image convolved with a Laplacian kernel (applied in the x and y directions independently) was used to estimate the amount of edges present in an image, and provide an estimate of "focus," with higher values indicating increased sharpness of the image.[37] This was implemented using the *cv2* package in Python.

Influence of Signal Averaging Prior to Denoising

In order to determine how the SNR of the input images influenced the performance of the denoising models, the individual repetitions acquired in the testing datasets were used to perform signal averaging prior to denoising. In each subject, the following datasets were created using the set of 10 repetitions (NSA = number of signal averages): NSA = 2 (5 repetitions), NSA = 3 (3 repetitions), NSA = 4 (2 repetitions), NSA = 5 (2 repetitions). The DAE, Gaussian and NLM filters were applied to each of these datasets, and the denoising performance (quantified using SNR) and accuracy (as compared to the reference standard images based on NSA = 10, and quantified using PSNR) of the resulting images were measured.

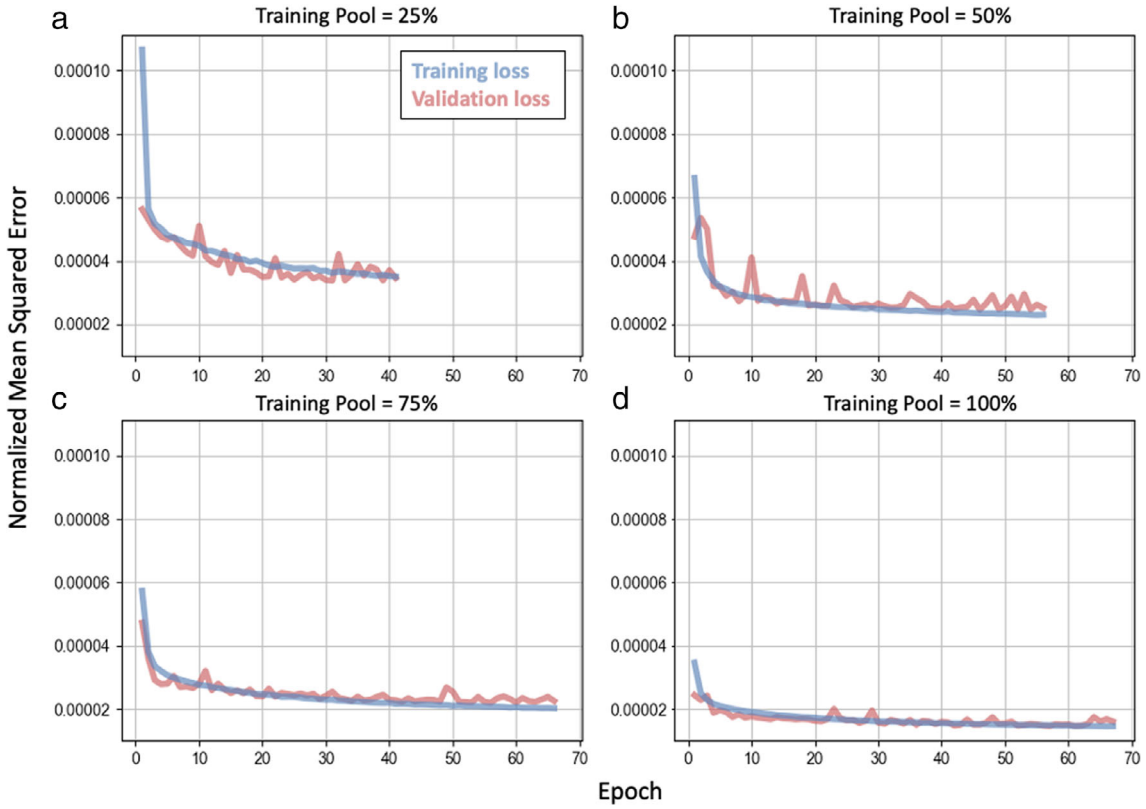

FIGURE 3: Plots of the training and validation loss as a function of epoch, during model training. The loss function was mean-squared-error, which was normalized to the number of images in the validation pool. The training dataset was split into four subsets, containing 25% (a), 50% (b), 75% (c), and 100% (d) of the full available training data (*N* = 28,820 images). An early-stopping criterion was applied to interrupt training when the loss in the validation data failed to improve over 10 consecutive epochs.

Correction of Motion Artifacts Using the DAE

In order to illustrate the ability of the DAE to correct small motion artifacts in the dM_{raw} images, the ASL data from one of the test subjects was used to simulate the effect of motion in the raw data. Using the first repetition from this subject, a spatial mismatch was created between the control and label images, by applying a range of in-plane rotations (ranging from 0.2–3.0 degrees), as well as translations in the x- and y-directions (ranging from 0.2–3.0 mm), to the control image only. A motion-corrupted dM_{raw} dataset was created for each of these (shifted control image - label image), after which the DAE model was applied. CBF maps were created using the non-motion-corrupted dM_{raw} dataset (CBF_{ref}), the motion-corrupted dM dataset (CBF_{MC}), and the motion-corrupted dM dataset after applying the DAE ($CBF_{MC\text{-}DAE}$). The mean absolute CBF error, using CBF_{ref} as the reference, was calculated across all brain voxels in the CBF_{MC} and $CBF_{MC\text{-}DAE}$ images, in order to quantify the level of CBF error introduced by the motion artifact, and the extent to which this was corrected using the DAE model.

Application of the DAE in Additional Clinical Examples

The DAE model was applied to patients 1–4 (described above), in order to illustrate its use in additional, un-seen clinical ASL images. In patients in whom an abnormal CBF hyperintensity was present as a result of their tumor, the ability of the DAE to retain this signal after denoising was examined. This was performed by converting the CBF_{raw} and CBF_{DAE} images to z-score maps, based on the mean and standard deviation of the CBF values across all brain voxels in a given patient. This was used to highlight regions of perfusion abnormality (high z-score) both before and after application of the DAE model.

Statistics

The *SciPy* Python package was used for all statistical analysis. For comparison of evaluation metrics between models, the Levine test was used to test for equal variances.[39] In cases of equal variance, one-way analysis of variance (ANOVA) tests were used for group comparisons, followed by a Tukey honestly significant difference post-hoc test. For unequal variance, a Welch ANOVA test was used, followed by a Games–Howell post-hoc analysis. All P values were reported after correcting for multiple comparisons, with significance defined as $P < 0.05$.

Results

Model Training and Filter Optimization

The training of the DAE model was performed using the UCL High Throughput Computing Facility, using compute nodes equipped with nVidia Tesla V100 GPUs and 192 GB of RAM per node. Typical training time was 25 minutes. Plots of the training and validation loss during training, using the full training dataset, as well for model training using

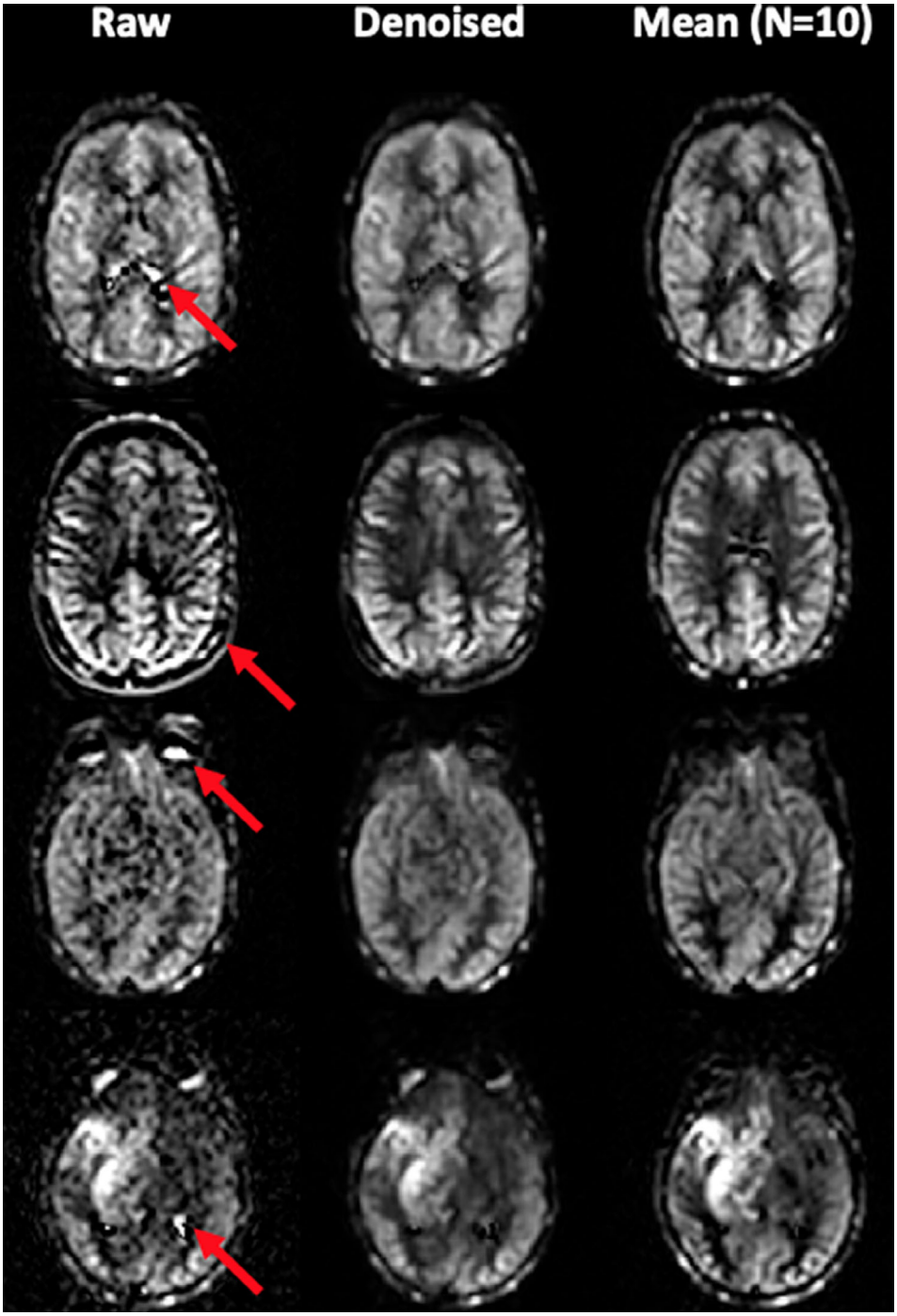

FIGURE 4: Example of the application of the DAE on un-seen data. The illustrative examples shown in Fig. 1 are shown here before and after denoising with the DAE, along with a further clinical example (bottom row; pediatric brain tumor patient [patient #1], showing diffuse hyperperfusion in the right temporal lobe). dM images from a single repetition are shown in both their raw form (left column) and after denoising with the DAE (middle column). The equivalent image after averaging over 10 repetitions is shown in the right column. Transient artifacts are indicated with arrows.

subsets of 25, 50, and 75% of the available training data, are shown in Fig. 3. The early stopping criteria, after which the model is no longer showing improving performance in the validation data, and is starting to "overfit" to the training data, was met at epoch 41 using 25% of the training data. This increased with larger training datasets, with early-stopping being reached at epoch 67 using 100% of the training data. Normalized mean-squared-error in the validation data also decreased with increasing size of the training dataset, ranging from 3.4×10^{-5} (25% training dataset) to 1.5×10^{-5} (100% training dataset; see Fig. 3). Combined, this indicated improved performance of the model when

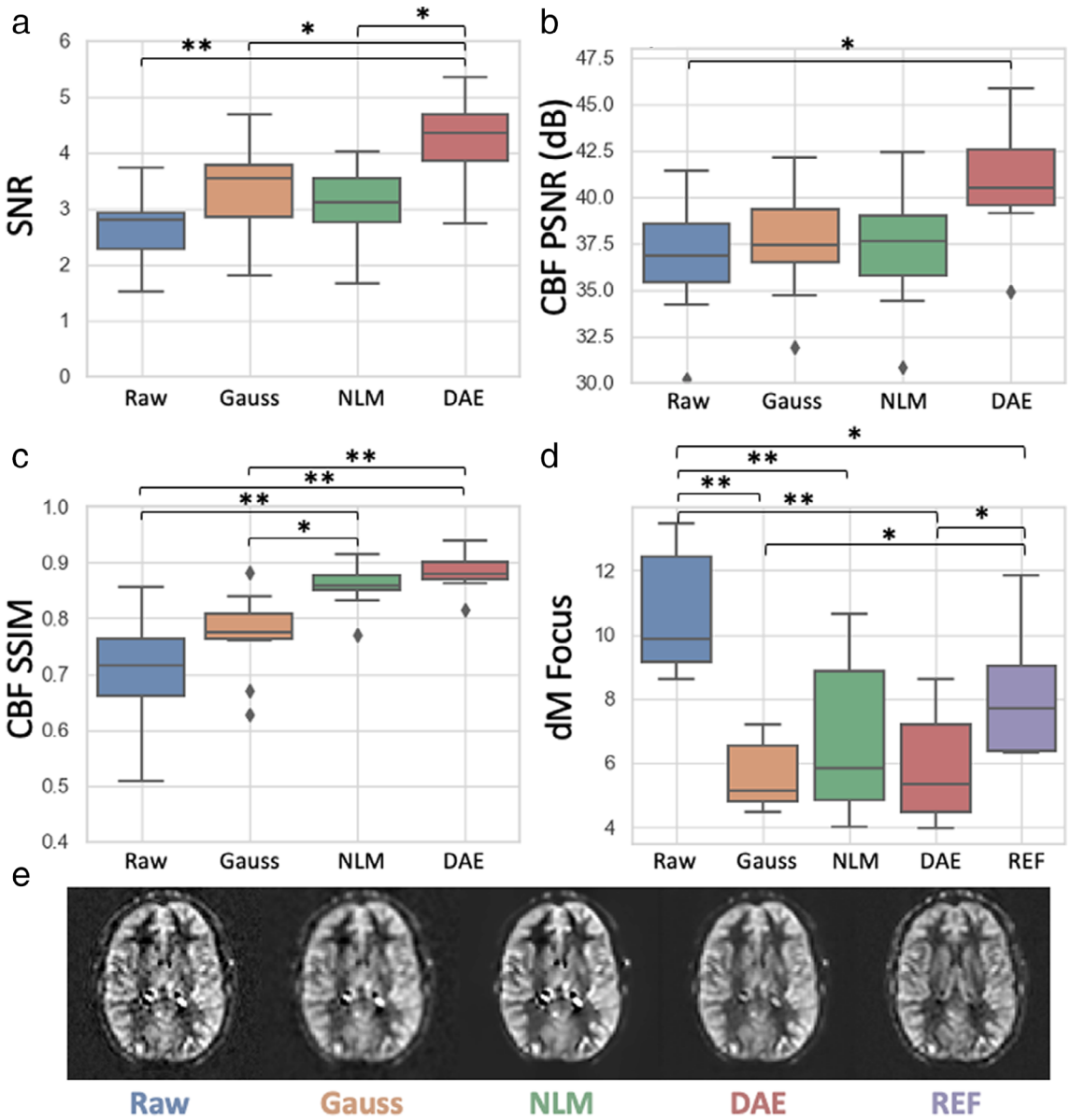

FIGURE 5: (a) Box-and-whisker plot of SNR in the dM images acquired in 11 healthy subjects, using the raw, Gaussian, non-local means (NLM), and denoising autoencoder (DAE) datasets. (b) Peak SNR values, obtained using CBF images calculated from dM_{raw}, dM_{Gauss}, dM_{NLM}, and dM_{DAE} datasets, compared to the reference standard CBF images. (c) Structural similarity index (SSIM) values, demonstrating the visual similarity assessment between CBF maps generated by the different models, in comparison to the reference standard. (d) dM "focus" values for each dataset, indicating the level of blurring introduced by the denoising or signal averaging. Higher focus values indicate a sharper image. (e) Example dM images from each model, in an axial slice from one representative subject. The artifactual CSF hyperintensity seen in the raw dM image remains prominent in the dM_{Gauss} and dM_{NLM} images, but is attenuated in the dM_{DAE} image, which more closely resembles the reference standard. Significant differences between groups (*$P < 0.05$, **$P < 0.001$) are illustrated in each plot.

trained using increasingly large datasets. The model trained using the full training dataset was used for the rest of this study, and is publically available[1].

For the Gaussian filter, the optimum window size was five voxels (standard deviation = 1.9 mm). For the NLM filter, the optimum patch size was six voxels, the optimum patch distance was 13 voxels, and optimum cutoff distance was 6.0.

Exemplary results from the DAE in the un-seen data (not used during training) are shown in Fig. 4, demonstrating the model's ability to suppress the transient artifacts illustrated in Fig. 1. The clinical example (patient #1) shown on the bottom row of Fig. 4 illustrates a bright artifactual signal in the lateral ventricle, which could be misinterpreted as a metastasis of the tumor in the temporal lobe. This artifactual signal is suppressed in both the reference standard and the denoised image.

Model Testing: pCASL Data in Healthy Subjects

The mean SNR of the dM_{raw}, dM_{Gauss}, dM_{NLM}, and dM_{DAE} images acquired in 11 healthy subjects is shown in Fig. 5a.

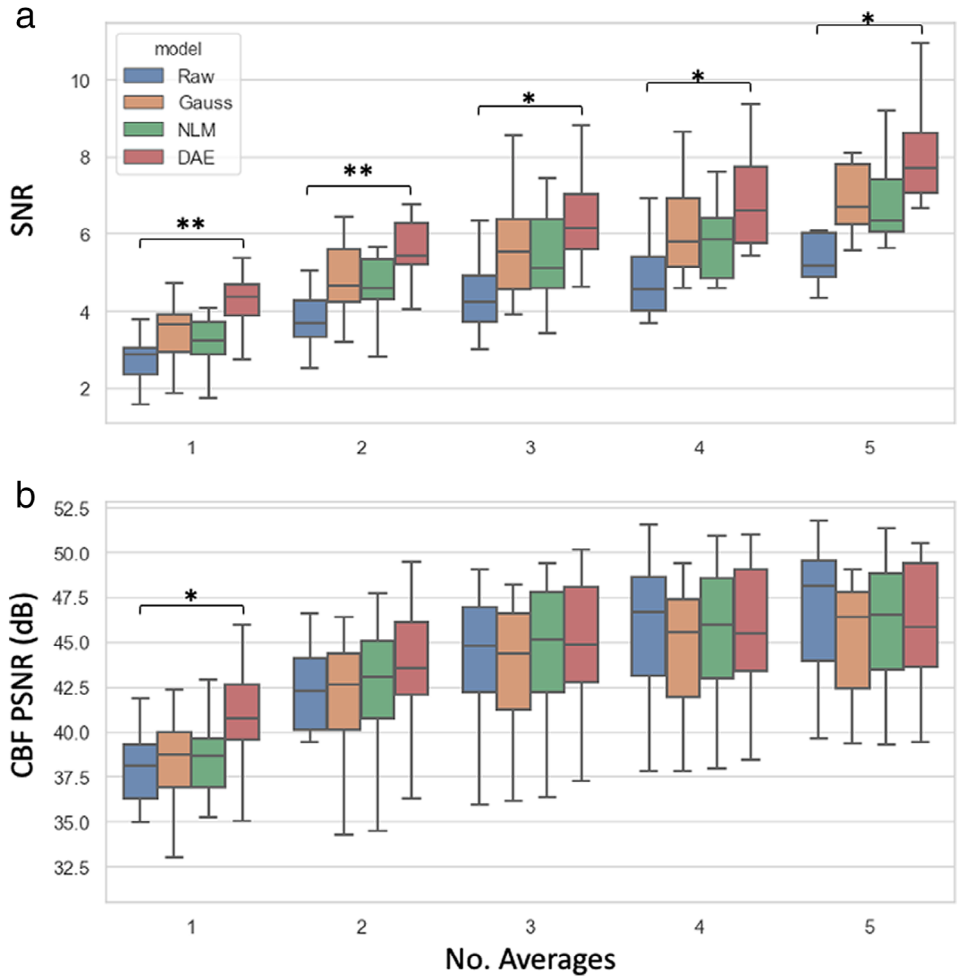

FIGURE 6: Influence of signal averaging prior to denoising for the different models. SNR values are shown in (a), and CBF PSNR values in (b), for input data ranging between 1–5 signal averages. Significant differences between groups (*$P < 0.05$, **$P < 0.001$) are illustrated in each plot.

Mean SNR was 2.6 ± 0.6 (± SD) in the raw images (range 1.5–3.7). The DAE was the only model to produce denoised images with significantly higher SNR than the raw images (4.2 ± 0.7, $P < 0.001$), representing an average gain of 62%. This was significantly higher than the gain in SNR offered by the Gaussian (27%) and NLM (15%) models (Fig. 5a, $P < 0.05$ for both comparisons).

The accuracy of the CBF images produced by each model, in comparison with the reference standard, is shown by the CBF PSNR values in Fig. 5b. PSNR was the highest in the CBF images produced using the DAE model (mean PSNR = 41.0 ± 2.9 dB), and this was the only model to produce a significant increase in PSNR compared to the raw CBF images (mean PSNR [raw] = 37.0 ± 3.1 dB, $P < .05$).

The structural similarity of the CBF images against the reference standard was lowest for the CBF_{raw} images (0.70 ± 0.10), and highest for the CBF_{DAE} images (0.88 ± 0.31), followed by the CBF_{NLM} images (0.86 ± 0.036; Fig. 5c). Both the DAE and NLM models resulted in significant increases in CBF SSIM compared to the CBF_{raw} images ($P < 0.001$).

In all denoised dM images, as well as the reference standard dM_{mean} images, focus values were significantly lower than those in the raw images (Fig. 5d, $P < 0.05$). As such, some degree of blurring was added, either as the result of signal averaging (in the dM_{mean} images), or from the denoising process. There was no significant difference between the dM focus values in the dM_{Gauss}, dM_{NLM}, or dM_{DAE} images; however, the dM_{NLM} images were the only ones not to show significantly lower focus values than the dM_{mean} images, indicating a marginally better performance of the NLM model in terms of image blurring.

Model Testing: Influence of Signal Averaging Prior to Denoising

Plots of dM SNR and CBF PSNR values, using input data obtained after averaging over 1–5 repetitions, are shown in Fig. 6. As expected, the SNR increased by a factor of approximately √NSA in the raw data, ranging from 2.6 ± 0.6 at

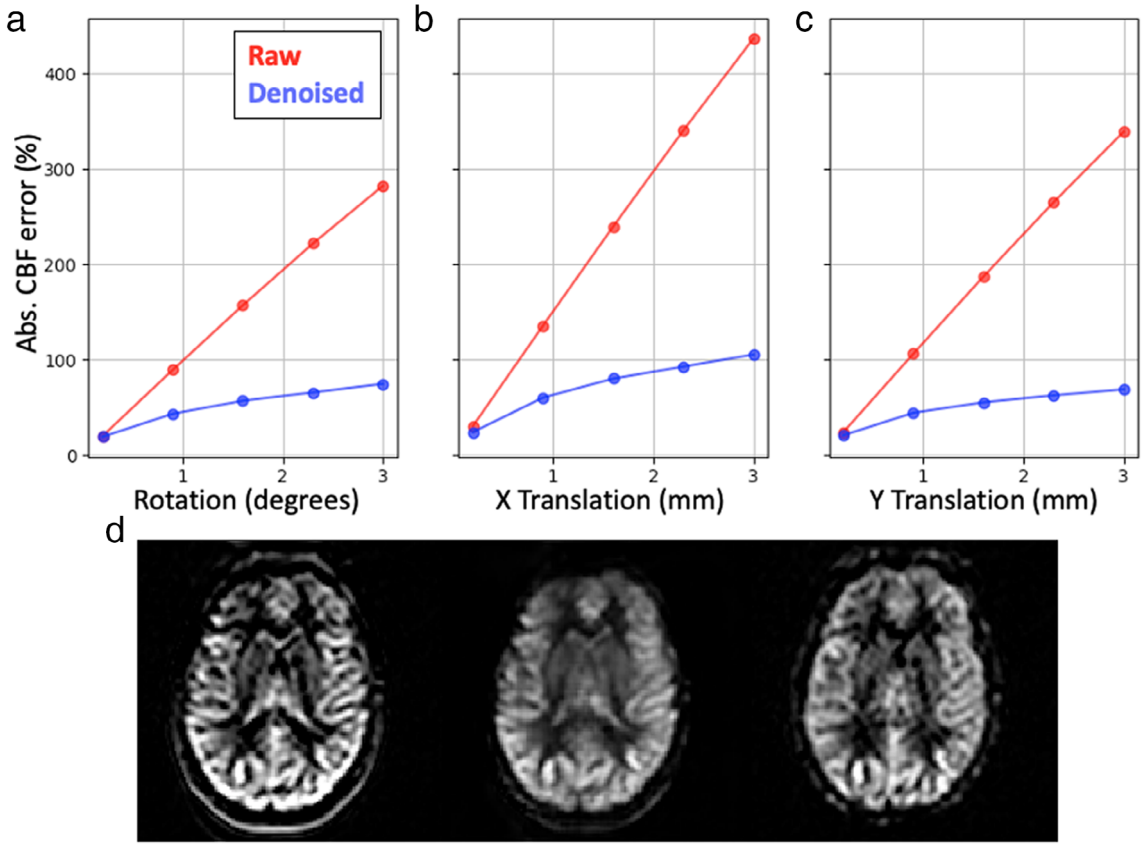

FIGURE 7: Simulation of motion artifacts before and after application of the DAE model. The simulated motion comprised a range of rotations (a), and translations in the x (b) and y (c) directions. The mean absolute percentage error in CBF quantification throughout the brain is shown in a–c for the raw motion-corrupted images (red lines) and motion-corrupted images after application of the DAE model (blue lines). An illustrative example of the dM images after a simulated translation of 0.5 mm in the +y direction is shown (d). Here, the raw motion-corrupted image is shown on the left, the same image after application of the DAE is shown in the middle, and the reference image (without any simulated motion) is shown on the right.

NSA = 1 to 5.7 ± 1.3 at NSA = 5 (Fig. 6a). Across all averaging levels, the DAE was the only model to provide a significant increase in SNR compared to the raw images ($P < 0.001$ for NSA = 1–2, $P < 0.05$ for NSA = 3–5). In terms of CBF PSNR, although the DAE model was the only model to provide a significant improvement over the CBF_{raw} images at NSA = 1; there was no significant difference between the PSNR values for any of the CBF images at NSA = 2–5 (Fig. 6b), and the PSNR values of all the denoising models appeared to plateau after NSA = 3. As such, although improvements in SNR occurred across the full range of NSA values, in terms of combined improvements in CBF accuracy as well as denoising, the DAE model appears to be most useful for raw data acquired with between 1 and 3 signal averages.

Model Testing: Correction of Motion Artifacts Using the DAE

The error in the CBF maps as a result of simulated subject motion, both before and after application of the DAE model, are shown in Fig. 7. As the CBF maps were calculated from a single repetition, even small levels of motion between the label and control acquisition can result in very large errors in CBF quantification. After application of the DAE model, the CBF error, while still large, was markedly reduced. For instance, for a rotation of 1.6 degrees, the average absolute CBF error across all brain voxels was 157% using the raw motion-corrupted images, which reduced to 57% after application of the DAE. Similarly, for a translation of 1.6 mm in the x direction, the absolute CBF error was 239% using the raw images, reducing to 80% after application of the DAE. The full results are given in Fig. 7, along with an illustrative example of the motion-corrupted and denoised images.

Model Testing: Multi-TI PASL Data in Healthy Subjects

The mean voxelwise χ^2 values, after fitting the Buxton kinetic model to multi-TI PASL data in seven healthy subjects, are shown in Fig. 8a. Model fitting using the dM_{Gauss}, dM_{NLM}, and dM_{DAE} datasets resulted in significantly lower voxelwise χ^2 values compared to model fitting using the dM_{raw} datasets ($P < 0.05$, all comparisons). The Buxton fit to the dM_{DAE} images produced significantly lower χ^2 values than all other dM images ($P < 0.05$, all comparisons). Example CBF, BAT,

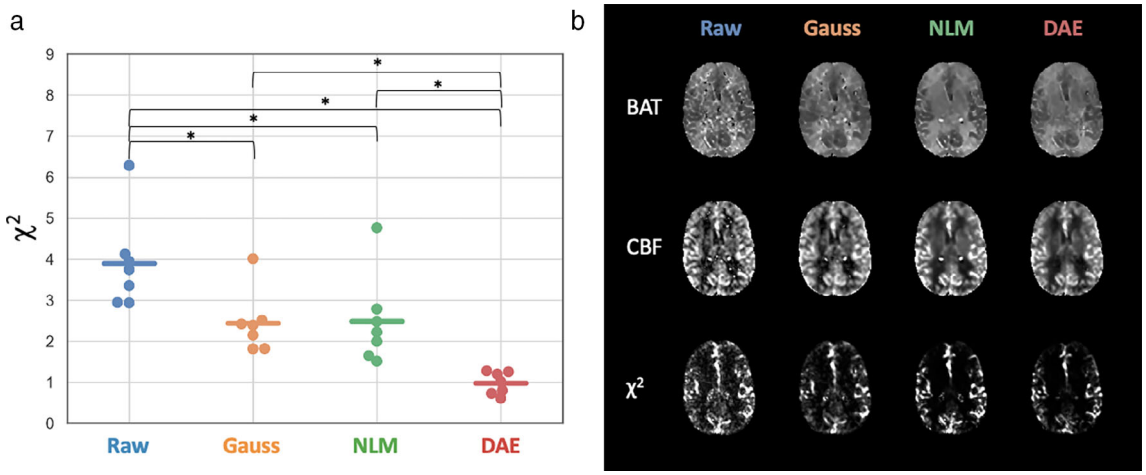

FIGURE 8: (a) Mean, voxelwise χ^2 values (sum of squared residuals), after fitting the Buxton kinetic ASL model to dM images from the raw, Gaussian, non-local means (NLM), and denoising autoencoder (DAE) datasets. Data points represent the mean voxelwise χ^2 values throughout the brain in individual subjects. Significant differences between groups are illustrated (*$P < 0.05$). (**b**) Example fitted maps of bolus arrival time (BAT), cerebral blood flow (CBF), and χ^2 values, in a representative subject.

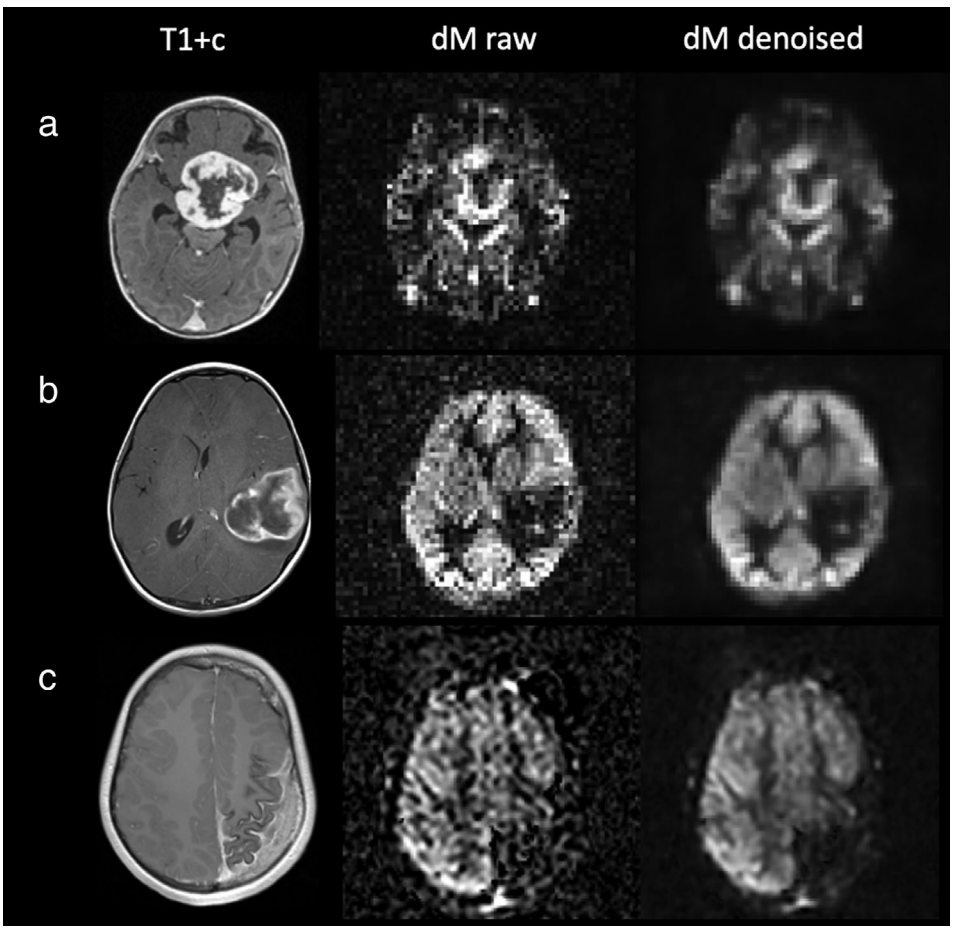

FIGURE 9: Clinical examples of implementation of the denoising autoencoder (DAE). T_1-weighted images after injection of gadolinium contrast agent (T_1 + c) are shown in the left column. Raw pCASL dM images are shown in the center column, with the equivalent denoised dM image shown on the right. Data shown in (a) and (**b**) were acquired at 1.5T (without zero-filling), data shown in (**c**) were acquired at 3T. (a) Patient #2: a 0.9-year-old with a pilocytic astrocytoma; (b) Patient #3: 4-year-old with a glioblastoma multiforme; (c) Patient #4: 11-year-old with Sturge–Weber syndrome, demonstrating a cortical angioma and marked atrophy.

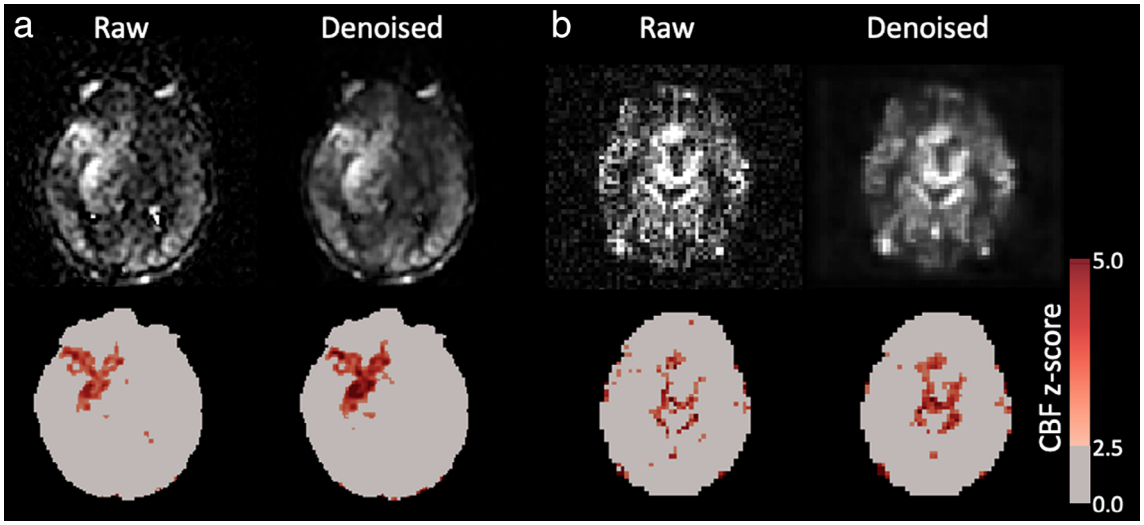

FIGURE 10: Raw and denoised dM images for patient #1 (a), and patient #2 (b), with the corresponding CBF z-score maps shown below. Only CBF z-scores > 2.5 are shown, to highlight regions of abnormal blood flow.

and χ^2 maps in an axial slice from a representative subject are shown in Fig. 8b.

Clinical Examples

Further examples of the DAE model applied to clinical ASL images, none of which were used during model training, are shown in Fig. 9. Furthermore, Fig. 10 illustrates the CBF z-score maps for patients #1 and #2, in which the patient's tumor resulted in a region of hyperperfusion. As shown in these illustrative examples, following denoising, the abnormal signal associated with pathology is indeed retained, and is in fact more prominent as a result of the denoising of the signal throughout the brain. The examples shown here show promise for the DAE for improving the conspicuity of perfusion abnormalities in noisy clinical ASL scans; however, further work is needed to investigate the clinical potential of this.

Discussion

In this work we have developed a deep-learning model for denoising ASL images, based on an autoencoder architecture. Our model was effective at both increasing SNR and suppressing transient artifacts in low-SNR ASL images, producing CBF images with the greatest accuracy in comparison to the reference standard. This is due to the ability of our model to not only learn how to denoise images, but to identify artifactual signals in a single image. In comparison, traditional denoising approaches such as Gaussian and NLM filters can be effective at improving SNR, but cannot learn to separate a prominent artifactual signal from a "true" signal. As such, transient artifacts remain in the denoised CBF images, which results in reduced accuracy.

As the SNR of the input images was increased, the DAE model continued to provide significant improvements in SNR. However, as signal averaging also reduces the prominence of transient artifacts in the input images, the improvement in CBF accuracy after denoising tended to level off as the number of averages was increased. As such, we believe the DAE model is most beneficial when applied to raw data acquired with a small number (~1–3) of averages. In this regard, the model is particularly well suited to multi-TI ASL data, as typically fewer signal averages are acquired per TI in these acquisitions, in exchange for a wider coverage of inflow times. Our results demonstrate that the DAE model performed well on multi-TI PASL data, providing dM images that had the best fit to the widely used Buxton kinetic model. This represents a promising future application for our proposed model.

By training on the large database of clinical pCASL scans, a further aim of this study was to produce a model that could differentiate between abnormal signals associated with pathology, and fluctuating abnormal signals associated with transient artifacts. Our model performed well in this regard, producing denoised images in which artifactual signals were suppressed, while pathological signals remained, and even appeared more prominent. In addition, despite being trained on clinical pediatric datasets, our results also suggest that the model performs well in healthy adult data, indicating that our model performs well under both pathological and non-pathological conditions.

In comparison to previous work, in this study no averaging, motion correction, or smoothing was applied to the noisy images used as inputs during training. This was done to maximize the conspicuity of transient artifacts in the noisy images, so that the model could effectively learn to suppress these, in conjunction with increasing SNR. One previous study also focused on a deep-learning approach for joint denoising and suppression of transient artifacts[28]; however, this required joint inputs relating to the mean and standard

deviation of the ASL signal over multiple repetitions. In comparison, our proposed model can suppress transient artifacts in single subtraction images alone. Also, in contrast to some previous studies, our model does not rely on additional anatomical T_1-weighted images[25] or a CBF signal model prior,[24] which should improve the generalizability of our model.

Limitations

A potential limitation of our study was that we employed a relatively simple model architecture compared to some recent studies in this area.[26] However, the performance of our model, in terms of PSNR and SSIM values, compares favorably with previous work.[25,26] In addition, the aim of this study was to develop and train the model, and test its performance in healthy datasets, rather than perform an in-depth assessment of its diagnostic utility under different pathological conditions. As such, further work should focus on a systematic subjective assessment of denoised images in different clinical scenarios, in order to fully explore the potential benefits of this model. Additionally, validation against an external standard for CBF quantification would be beneficial.

Conclusion

We have proposed a deep-learning-based framework for simultaneous denoising and suppression of transient artifacts in ASL images. The model works effectively on low-SNR ASL data acquired without signal averaging, and produces CBF maps that show good agreement with those acquired with 10 signal averages. As such, our model could provide a significant saving in the scan time required to acquire ASL data.

ACKNOWLEDGMENTS

The authors acknowledge the use of the UCL Myriad High Throughput Computing Facility (Myriad@UCL), and associated support services. The work is funded by Children with Cancer UK. We also acknowledge support from the following: National Institute for Health Research Biomedical Research Centre at Great Ormond Street Hospital for Children NHS Foundation Trust and University College London.

FOOTNOTES

[1]https://github.com/patrickhales/asl-denoising

REFERENCES

1. Detre JA, Leigh JS, Williams DS, Koretsky AP. Perfusion imaging. Magn Reson Med 1992;23:37-45.
2. Williams DS, Detre JA, Leigh JS, Koretsky AP. Magnetic resonance imaging of perfusion using spin inversion of arterial water. Proc Natl Acad Sci U S A 1992;89:212-216.
3. Alsop DC, Detre JA, Golay X, et al. Recommended implementation of arterial spin-labeled perfusion MRI for clinical applications: A consensus of the ISMRM perfusion study group and the European consortium for ASL in dementia. Magn Reson Med 2014;73:102-116.
4. Ye FQ, Frank JA, Weinberger DR, McLaughlin AC. Noise reduction in 3D perfusion imaging by attenuating the static signal in arterial spin tagging (ASSIST). Magn Reson Med 2000;44:92-100.
5. Deibler AR, Pollock JM, Kraft RA, Tan H, Burdette JH, Maldjian JA. Arterial spin-labeling in routine clinical practice, part 1: Technique and artifacts. AJNR Am J Neuroradiol 2008;29:1228-1234.
6. Amukotuwa SA, Yu C, Zaharchuk G. 3D Pseudocontinuous arterial spin labeling in routine clinical practice: A review of clinically significant artifacts. J Magn Reson Imaging 2016;43:11-27.
7. Tan H, Maldjian JA, Pollock JM, et al. A fast, effective filtering method for improving clinical pulsed arterial spin labeling MRI. J Magn Reson Imaging 2009;29:1134-1139.
8. Maumet C, Maurel P, Ferré J-C, Barillot C. Robust estimation of the cerebral blood flow in arterial spin labeling. Magn Reson Imaging 2014;32:497-504.
9. Dolui S, Wang Z, Shinohara RT, Wolk DA, Detre JA. Alzheimer's disease neuroimaging initiative: Structural correlation-based outlier rejection (SCORE) algorithm for arterial spin labeling time series. J Magn Reson Imaging 2017;45:1786-1797.
10. Li Y, Dolui S, Xie D-F, Wang Z. Alzheimer's disease neuroimaging initiative: Priors-guided slicewise adaptive outlier cleaning for arterial spin labeling perfusion MRI. J Neurosci Methods 2018;307:248-253.
11. Shirzadi Z, Crane DE, Robertson AD, et al. Automated removal of spurious intermediate cerebral blood flow volumes improves image quality among older patients: A clinical arterial spin labeling investigation. J Magn Reson Imaging 2015;42:1377-1385.
12. Behzadi Y, Restom K, Liau J, Liu TT. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. Neuroimage 2007;37:90-101.
13. Wells JA, Thomas DL, King MD, Connelly A, Lythgoe MF, Calamante F. Reduction of errors in ASL cerebral perfusion and arterial transit time maps using image de-noising. Magn Reson Med 2010;64:715-724.
14. Zhu H, Zhang J, Wang Z. Arterial spin labeling perfusion MRI signal denoising using robust principal component analysis. J Neurosci Methods 2018;295:10-19.
15. Spann SM, Kazimierski KS, Aigner CS, Kraiger M, Bredies K, Stollberger R. Spatio-temporal TGV denoising for ASL perfusion imaging. Neuroimage 2017;157:81-96.
16. Wang J, Aguirre GK, Kimberg DY, Detre JA. Empirical analyses of null-hypothesis perfusion FMRI data at 1.5 and 4 T. Neuroimage 2003;19:1449-1462.
17. Fazlollahi A, Bourgeat P, Liang X, et al. Reproducibility of multiphase pseudo-continuous arterial spin labeling and the effect of post-processing analysis methods. Neuroimage 2015;117:191-201.
18. Buades A, Coll B, Morel J-M: A non-local algorithm for image denoising. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). Vol. 2; 2005, p 60–65.
19. Bibic A, Knutsson L, Ståhlberg F, Wirestam R. Denoising of arterial spin labeling data: Wavelet-domain filtering compared with Gaussian smoothing. Magma 2010;23:125-137.
20. Jiang D, Dou W, Vosters L, Xu X, Sun Y, Tan T. Denoising of 3D magnetic resonance images with multi-channel residual learning of convolutional neural network. Jpn J Radiol 2018;36:566-574.
21. Manjón JV, Coupe P. MRI Denoising using deep learning. In: Bai W, Sanroma G, Wu G, Munsell BC, Zhan Y, Coupé P, editors. *Patch-based techniques in medical imaging*. Cham, Switzerland: Springer International Publishing; 2018. p 12-19. [Lecture Notes in Computer Science.].
22. Benou A, Veksler R, Friedman A, Riklin Raviv T. Ensemble of expert deep neural networks for spatio-temporal denoising of contrast-enhanced MRI sequences. Med Image Anal 2017;42:145-159.

23. Kim KH, Choi SH, Park S-H. Improving arterial spin labeling by using deep learning. Radiology 2018;287:658-666.

24. Ulas C, Tetteh G, Kaczmarz S, Preibisch C, Menze BH. DeepASL: Kinetic model incorporated loss for denoising arterial spin labeled MRI via deep residual learning. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G, editors. *Medical image computing and computer assisted intervention — MICCAI 2018*. Cham, Switzerland: Springer International Publishing; 2018. p 30-38. [Lecture Notes in Computer Science.].

25. Gong K, Han P, El Fakhri G, Ma C, Li Q. Arterial spin labeling MR image denoising and reconstruction using unsupervised deep learning. NMR Biomed 2019;e4224. https://doi.org/10.1002/nbm.4224.

26. Xie D, Li Y, Yang H, et al. Denoising arterial spin labeling perfusion MRI with deep machine learning. Magn Reson Imaging 2020;68: 95-105.

27. Xie D, Bai L, Wang Z: Denoising arterial spin labeling cerebral blood flow images using deep learning. arXiv:180109672 [cs] 2018.

28. Owen D, Melbourne A, Eaton-Rosen Z, et al. Deep convolutional filtering for spatio-temporal denoising and artifact removal in arterial spin labeling MRI. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G, editors. *Medical image computing and computer assisted intervention — MICCAI 2018*. Cham, Switzerland: Springer International Publishing; 2018. p 21-29. [Lecture Notes in Computer Science.].

29. Luh W-M, Wong EC, Bandettini PA, Hyde JS. QUIPSS II with thin-slice TI periodic saturation: A method for improving accuracy of quantitative perfusion imaging using pulsed arterial spin labeling. Magn Reson Med 1999;41:1246-1254.

30. Jenkinson M, Smith S. A global optimisation method for robust affine registration of brain images. Med Image Anal 2001;5:143-156.

31. Mao X-J, Shen C, Yang Y-B: Image restoration using convolutional auto-encoders with symmetric skip connections. arXiv 2016; 1606.08921 [cs.CV].

32. Buxton RB, Frank LR, Wong EC, Siewert B, Warach S, Edelman RR. A general kinetic model for quantitative perfusion imaging with arterial spin labeling. Magn Reson Med 1998;40:383-396.

33. Dietrich O, Raya JG, Reeder SB, Reiser MF, Schoenberg SO. Measurement of signal-to-noise ratios in MR images: Influence of multichannel coils, parallel imaging, and reconstruction filters. J Magn Reson Imaging 2007;26:375-385.

34. Schönberg SO, Dietrich O, Reiser MF. *Parallel imaging in clinical MR applications*. Berlin: Springer; 2007 Diagnostic Imaging.

35. Smith SM. Fast robust automated brain extraction. Hum Brain Mapp 2002;17:143-155.

36. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. IEEE Trans Image Process 2004;13:600-612.

37. Pertuz S, Puig D, Garcia MA. Analysis of focus measure operators for shape-from-focus. Pattern Recognit 2013;46:1415-1432.

38. Nayar SK, Nakagawa Y. Shape from focus. IEEE Trans Pattern Anal Mach Intell 1994;16:824-831.

39. *Olkin I: Contributions to probability and statistics; essays in honor of Harold Hotelling*. Stanford, CA: Stanford University Press; 1960.