Improving Sensitivity of Arterial Spin Labeling Perfusion MRI in Alzheimer's Disease Using Transfer Learning of Deep Learning-Based ASL Denoising

Lei Zhang, PhD,[1] Danfeng Xie, PhD,[1] Yiran Li, PhD,[1] Aldo Camargo, PhD,[1]
Donghui Song, MS,[1] Tong Lu, PhD,[2] Jean Jeudy, MD,[1] David Dreizin, MD,[1]
Elias R. Melhem, MD,[1] and Ze Wang, PhD,[1]* Alzheimer's Disease Neuroimaging Initiative[#]

Background: Arterial spin labeling (ASL) perfusion magnetic resonance imaging (MRI) denoising through deep learning (DL) often faces insufficient training data from patients. One solution is to train DL models using healthy subjects' data which are more widely available and transfer them to patients' data.
Purpose: To evaluate the transferability of a DL-based ASL MRI denoising method (DLASL).
Study Type: Retrospective.
Subjects: Four hundred and twenty-eight subjects (189 females) from three cohorts.
Field Strength/Sequence: 3 T two-dimensional (2D) echo-planar imaging (EPI)-based pseudo-continuous ASL (PCASL) and 2D EPI-based pulsed ASL (PASL) sequences.
Assessment: DLASL was trained using young healthy adults' PCASL data (Dataset 1: 250/30 subjects as training/validation set) and was directly transferred (DTF) to PCASL data from Dataset 2 (45 subjects test set) of normal controls (NC) and Alzheimer's disease (AD) groups. DLASL was fine-tuned (DLASLFT) and tested on PASL data from Dataset 3 (103 subjects test set) of NC and AD. An existing non-DL method (NonDL) was used for comparison. Cerebral blood flow (CBF) images from ASL MRI were compared between NC and AD to assess characteristic hypoperfusion (lower CBF) patterns in AD. CBF image quality and CBF map sensitivity for detecting hypoperfusion using peak *t*-value and suprathreshold cluster size are outcome measures.
Statistical Tests: Paired *t*-test, two-sample *t*-test, one-way analysis of variance, and Tukey honestly significant difference, and linear mixed-effects models were used. $P < 0.05$ was considered statistically significant.
Results: Mean contrast-to-noise ratio (CNR) of Dataset 2 showed that DTF outperformed NonDL (AD: 3.38 vs. 2.64, NC: 3.80 vs. 3.36). On Dataset 3, DLASLFT outperformed NonDL measured by mean CNR (AD: 2.45 vs. 1.87, NC: 2.54 vs. 2.17) and mean radiologic score (2.86 vs. 2.44). Image quality improvement was significant on both test sets. DTF and DLASLFT improved sensitivity for detecting AD-related hypoperfusion patterns compared with NonDL.
Data Conclusion: We demonstrated the DLASL's transferability across different ASL sequences and different populations.
Level of Evidence: 3
Technical Efficacy: Stage 2

J. MAGN. RESON. IMAGING 2021.

View this article online at wileyonlinelibrary.com. DOI: 10.1002/jmri.27984

Received May 24, 2021, Accepted for publication Oct 22, 2021.

*Address reprint requests to: Z.W., HSF III Room 1173, 670 W Baltimore Street, Baltimore, MD 21201, USA. E-mail: ze.wang@som.umaryland.edu,
The first two authors contributed equally to this work.

[#] 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 the 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

From the [1] Department of Diagnostic Radiology and Nuclear Medicine, University of Maryland School of Medicine, Baltimore, Maryland, USA; and [2] Department of Mathematics, University of Maryland, College Park, College Park, Maryland, USA

Arterial spin labeling (ASL) perfusion magnetic resonance imaging (MRI) is a noninvasive technique for quantifying cerebral blood flow (CBF)[1] and has been increasingly used in neuroscientific and translational studies for assessing brain function and neurovascular conditions.[2] In ASL MRI, arterial blood is labeled with radio-frequency pulses in locations proximal to the imaging plane. A perfusion-weighted MR image is acquired after the labeled spins perfuse into brain tissues. To remove the background MR signal, a control image is also acquired using the same ASL sequence and acquisition timing but with phase modulations to the labeling pulses so that arterial spins can approximately stay unaffected. Perfusion signal is subsequently calculated through pair-wise subtraction of the spin labeled image (L image) and the spin untagged image (the control image or C image) and is converted into the quantitative CBF in units of mL/100 g/minute.[3] Limited by several factors including the longitudinal relaxation rate (T1) of blood water, labeling efficiency, and the post-labeling delay, the labeled blood signal is inherently weak, resulting in a low signal-to-noise ratio (SNR) and limited spatial resolution.[4]

Over the past decades, various methods have been proposed to improve ASL CBF quantification results through model-based and data-driven approaches.[4–7] Machine learning represent a new direction in this endeavor and the published studies include the use of principal component analysis,[8] independent component analysis,[9] support vector machine,[10] low-rank and sparse decomposition,[6] and the spatio-temporal total generalized variation constrained method.[11] Deep machine learning represents a recent focus in ASL MRI processing because of its high flexibility for modeling the nonlinear transform underlying denoising or other spatial-temporal processing. Deep learning (DL) has been used in ASL MRI to improve SNR, spatial resolution, and temporal resolution in ASL MRI.[4,12–20] For denoising per se, DL has the advantage of learning the highly nonlinear noise removal function based on the between neighboring voxel correlations, non-local data features, as well as the perfusion signal data distributions of a large number of subjects. Consequently, it has shown state-of-art performance when trained and tested in normal health subjects' data.[4] These advantages, however, cannot overwrite a general question that each machine learning algorithm has to face: is the deep neural network learned from one dataset valid for another one? This generalizability question has not been examined in ASL MRI but it is very important for the potential translational application of the DL methods. For example, a DL model learned from young healthy subjects may not work for data acquired from older normal subjects, or patients with Alzheimer's Disease (AD) using the same or different ASL imaging sequences.

The purpose of this study was to evaluate the transferability of the recently developed DL-based ASL MRI denoising method (DLASL)[4] for clinical applications in AD. We chose AD as a test point because ASL MRI has been widely used in AD research and revealed a consistent hypoperfusion pattern in the precuneus and lateral parietal cortex as well as prefrontal cortex in AD.[21,22] We hypothesized that DLASL trained with healthy subjects' data can be transferred to AD ASL MRI and will demonstrate higher sensitivity for probing the hypoperfusion patterns in AD as compared with normal elderly subjects (NC).

Materials and Methods

Datasets

The study and subject consent form were reviewed and approved by the Institutional Review Board (IRB). All participants signed a written consent form before participating in the study. Retrospectively using this dataset was approved by IRB as well. All healthy volunteer scans used in this study were also approved by IRB and with the written consent from the participants. Three different datasets were included in this study.

Dataset 1: This was the same as that reported in the original DLASL paper.[4] It contained ASL MRI from 280 young healthy adults (age: 23–47, 110 females, 170 males). The data were acquired in a Siemens 3 T Trio scanner with a two-dimensional (2D) gradient echo echo-planar imaging readout based pseudo-continuous ASL (PCASL) sequence[23] with the following parameters: 40 control/labeled image pairs with labeling time = 1.5 sec, post-labeling delay = 1.5 sec, field of view (FOV) = 22 cm, matrix = 64 × 64, repetition time/echo time (TR/TE) = 4000/11 msec, 20 slices with a thickness of 5 mm plus 1 mm gap.

Dataset 2: It contained 45 subjects (age: 51–83, 25 females, 20 males), i.e., 21 AD patients and 24 normal elderly control (NC). These data were acquired in the same MR scanner using the same PCASL sequence with the same acquisition parameters as used in Dataset 1.

Dataset 3: The dataset was downloaded from the AD Neuroimaging Initiative (ADNI) (http://adni.loni.usc.edu/) and included structural MRI and ASL MRI from 103 subjects, i.e., (age: 59–87, 54 females, 49 males), i.e., 53 AD patients and 50 NCs. These data were acquired in Siemens 3 T MR scanners using the Siemens product 2D Proximal Inversion with Control of Off-Resonance Effects (PICORE) sequence, which is a pulsed ASL sequence using the quantitative imaging of perfusion using a single subtraction II (Q2TIPs) technique for defining the spin bolus.[24] The acquisition parameters were 52 control/labeled image pairs, TR/TE = 3400/12 msec, TI1/TI2 = 700/1900 msec, FOV = 256 mm, 24 sequential 4 mm thick slices with a 25% gap between the adjacent slices, partial Fourier factor = 6/8, bandwidth = 2368 Hz/pix, and imaging matrix = 64 × 64.

Data Preprocessing

All imaging data were processed using ASLtbx[25] and SPM12 (Wellcome Centre for Human Neuroimaging, London, UK, http://www.fil.ion.ucl.ac.uk/spm). The following steps were included in ASL MRI preprocessing: ASL MRI specific motion correction;[5] temporal denoising;[5] spatial smoothing; CBF quantification; outlier cleaning.[30] Mean CBF map was then calculated from the outlier cleaned CBF image series. To assess the AD vs. NC CBF difference, the mean CBF images and those denoised by DL networks were registered by using deformable image registration algorithm as implemented in SPM

12, with the high-resolution structural MRI and subsequently registered into the Montreal Neurological Institute standard brain (MNI space)[26,27] using SPM12.

DL Network Architecture

The original DLASL model[4] was used in this study. Figure 1 shows the architecture of DLASL. The Dilated Wide Activation Network[4] was used to extract more data features at each block. Rather than training the network using CBF maps in the MNI space[26,27] as in the previous study,[4] DLASL was trained using the CBF maps in the native space (before warping into the MNI space). Considering the reduced number of image slices for network training, the number of wide activation residual blocks was reduced to be three in each pathway to reduce the overfitting risk. Moreover, we used Huber loss[29] in this study because it is more robust to outliers than L2 loss and is more precise and stable than L1 loss during the training.

Model Training and Evaluation

DLASL was mainly trained using the PCASL data in Dataset 1. CBF image slices from 250 subjects were used as the training dataset. CBF image slices from another 30 subjects were used for validation. The model was trained using all axial image slices, each with 64 × 64 pixels. During training, the input to DLASL were mean images of the first 10 CBF images of the entire ASL scan. We performed the

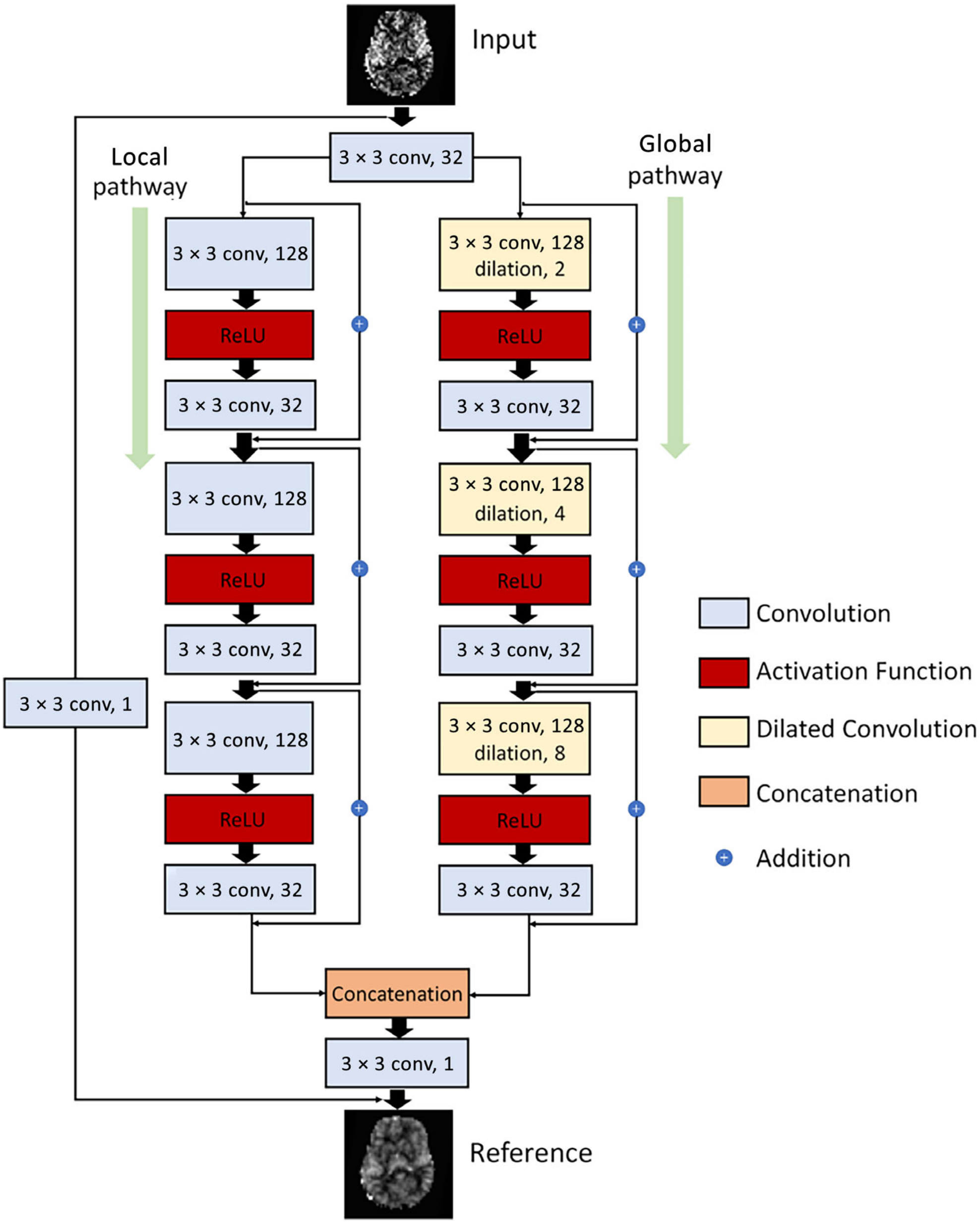

FIGURE 1: Illustration of the architecture of the DLASL. The output of the first layer was fed to both local pathway and global pathway. Each pathway contains three consecutive wide activation residual blocks. Each wide activation residual block contains two convolutional layers (3 × 3 × 128 and 3 × 3 × 32) and one activation function layer. The 3 × 3 × 128 convolutional layers in the global pathway were dilated convolutional layers[28] with a dilation rate of 2, 4 and 8 respectively (a × b × c indicates the property of convolution. a × b is the kernel size of one filter and c is the number of the filters).

priors-guided slice-wise adaptive outlier cleaning[30] by excluding the outlier slices in CBF image series of the whole ASL scan. The reference image was the mean image of the outlier-cleaned CBF image series of the entire ASL scan. After DLASL was trained, the outlier-cleaned mean CBF images were input to the network to generate the DL denoised CBF maps.

ASL MRI in Dataset 1 and Dataset 2 were acquired using the same PCASL sequence with the same acquisition parameters. Except for the CBF value difference between the populations (subjects in Dataset 2 [age: 51–83] were older than those in Dataset 1 [age: 23–47]), these two datasets were both acquired in Siemens 3 T MR scanner using the same PCASL sequence, making a direct model transfer possible. We then directly applied the DLASL network trained with Dataset 1 to the data in Dataset 2 without model retraining.

ASL MRI (Dataset 3) was acquired with a PASL sequence from ADNI (http://adni.loni.usc.edu/), which is known to have lower SNR than PCASL.[31] To consider the effects of the reduced SNR in PASL on the performance of DLASL, we tested the trained DLASL without and with fine-tuning (i.e., retraining the model which was pre-trained with PCASL data by using PASL data). All 103 subjects were used for fine-tuning the model. The input and the reference were the same outlier-cleaned mean CBF maps. Since the total number of ADNI subjects was 103, which is smaller than the number of subjects in Dataset 1, the chance of overfitting the model to the 103 subjects' data is considered very low.

All networks were implemented using the Keras platform.[32] For PCASL data in Dataset 1, the network was trained from scratch using adaptive moment estimation algorithm[33] with a basic learning rate of 0.001. When we applied transfer learning, we used a learning rate of 0.0001 to fine-tune the pre-trained model. Early stopping technique[34] was used to avoid overfitting. If the model's performance on the validation set did not improve after 10 epochs, we stopped the training process. As shown in Fig. 2, we monitored the training loss and validation loss during training/fine-tuning the models. All the models were trained with batches, each containing 64 training samples. The models were trained/fine-tuned with a total of 100 epochs. TensorFlow[35] was used as the backend of Keras to train all the models. All experiments were performed on a PC with Intel(R) Core(TM) i7-5820k CPU @3.30 GHz and a Nvidia GeForce Titan Xp GPU.

For easy reading, we refer to the priors-guided slice-wise adaptive outlier cleaning method,[30] the DLASL direct transfer (without fine-tuning), and DLASL with fine-tuning as NonDL, DTF, and DLASLFT, respectively.

Evaluation Metric

Contrast-to-noise ratio (CNR) was used as the quantitative measurement for image quality assessment of ASL CBF denoising by different methods.[36] A grey matter (GM) and a white matter (WM) mask were defined in the CBF maps by projecting the GM/WM masks defined by the structural MRI-based GM/WM image segmentation. CNR was calculated by:

$$\mathrm{CNR} = \frac{\mathrm{Mean(GM)}}{\mathrm{std(WM)}} \tag{3}$$

where mean(GM) and std(WM) denote mean CBF in the GM mask and std of CBF values in the WM mask, respectively.

Image Quality Assessment

CBF image quality was qualitatively assessed by three blinded independent reviewers, i.e., J.J. (20 years of experience), D.D. (9 years of experience), and E.M. (28 years of experience). The quality score ranges from 1 (worst quality) to 4 (best quality). The score means: "1 = worst quality in terms of severe artifacts, too low or negative perfusion (the black holes in the brain) and abnormally high perfusion value"; "2 = moderate quality in terms of large artifacts, presence of negative perfusion voxels, image intensity in the range of normal perfusion of adults"; "3 = acceptable quality in terms of mild artifacts, grey matter to white matter perfusion contrast"; and "4 = clear perfusion image with good grey matter to white matter perfusion contrast, the least image artifacts such as rings or bright strips," respectively. We randomly selected CBF images from 30 subjects (15 AD and 15 NC) processed by the three different methods: NonDL, DTF, and DLASLFT. The CBF images were shuffled before they were presented to the three blinded independent reviewers.

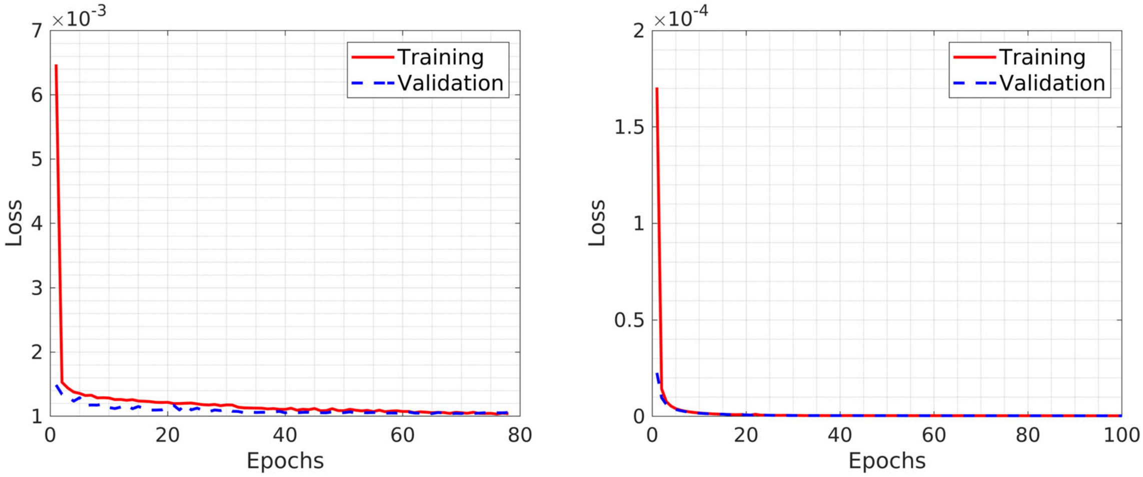

FIGURE 2: Left, training and validation loss of DLASL on Dataset 1. Right, training and validation loss of DLASL on Dataset 3.

Statistical Analyses

After obtaining the CNRs of the three denoising methods NonDL, DTF, and DLASLFT, a group of paired *t*-tests and two-sample *t*-tests were performed. On Dataset 2, we performed the paired *t*-test between the methods, NonDL and DTF, on the AD group and NC group. We then performed paired *t*-tests between methods, DTF vs. the NonDL, and methods, DLASLFT vs. NonDL, on both AD group and NC group of Dataset 3. On both Dataset 2 and Dataset 3, to assess the hypoperfusion in AD, a voxel-wise two-sample *t*-test was performed in SPM12 using the CBF maps (with or without DL denoising) spatially normalized into the MNI space. After the three pre-selected reviewers evaluated the denoising results generated by the NonDL, DTF, and DLASLFT on the selected 30 subjects, we performed the one-way analysis of variance (ANOVA) on the average evaluation scores of each method. In addition, we performed the Tukey honestly significant difference post-hoc tests for all pairwise group comparisons using 0.05 as the familywise error rate. At last, a linear mixed-effects model incorporating fixed effects (three denoising methods) and random effects (three reviewers) was used to assess the evaluation scores which quantify the image qualities and the transfer capability of different methods. Moreover, the interaction between methods and reviewers was considered and to be tested for its statistical significance. We intend to investigate how significant each level of effects (methods and reviewers) affects the evaluation scores (and eventually the transfer capability). In this study, $P < 0.05$ was considered statistically significant.

Results

Qualitative Evaluation

Figure 3 showed CBF images from a representative AD subject and NC subject from Dataset 2. Figure 3a,b represented the AD patient's CBF maps processed by the conventional processing method and DTF, respectively. Compared with Fig. 3a, Fig. 3b showed better image quality in terms of fewer artifacts, especially on the boundary of the CBF map. Figure 3c,d was the NC's CBF maps produced by the NonDL and DTF. Figure 3d showed better image quality than Fig. 3c. Figure 3 indicated that DTF produced better image quality than the NonDL for both AD and NC subjects.

Figure 4 showed the denoised PASL CBF images from a representative AD subject and a NC subject from Dataset 3. Figure 4a–c was the CBF images of an AD patient

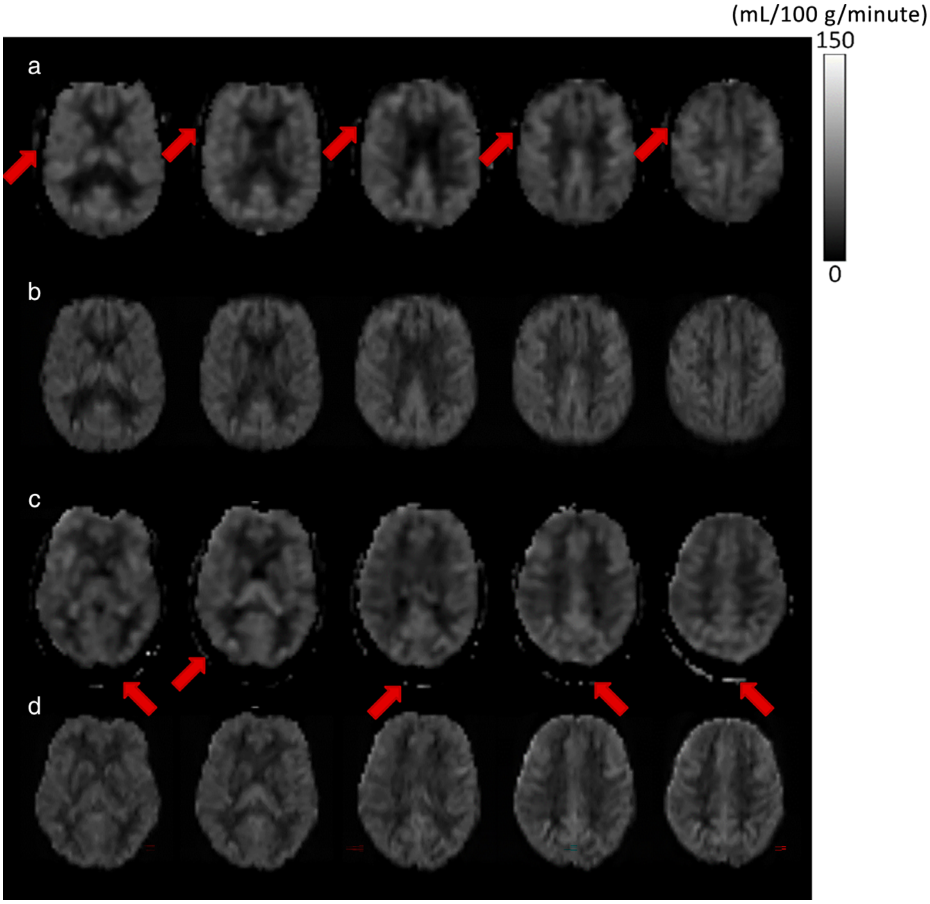

FIGURE 3: Mean PCASL CBF images of a representative AD subject and a NC subject from dataset 2. The rows from top to bottom are: (a) CBF map of an AD patient generated by the NonDL (pseudo ground truth)[30]; (b) DTF denoised CBF map from the same AD patient; (c) CBF map of a NC by the NonDL (pseudo ground truth); (d) DTF denoised version of (c). The noisy structures can only be removed by DTF were illustrated with red arrows. PCASL = pseudo-continuous arterial spin labeling; CBF = cerebral blood flow; AD = Alzheimer's disease; NC = normal controls.

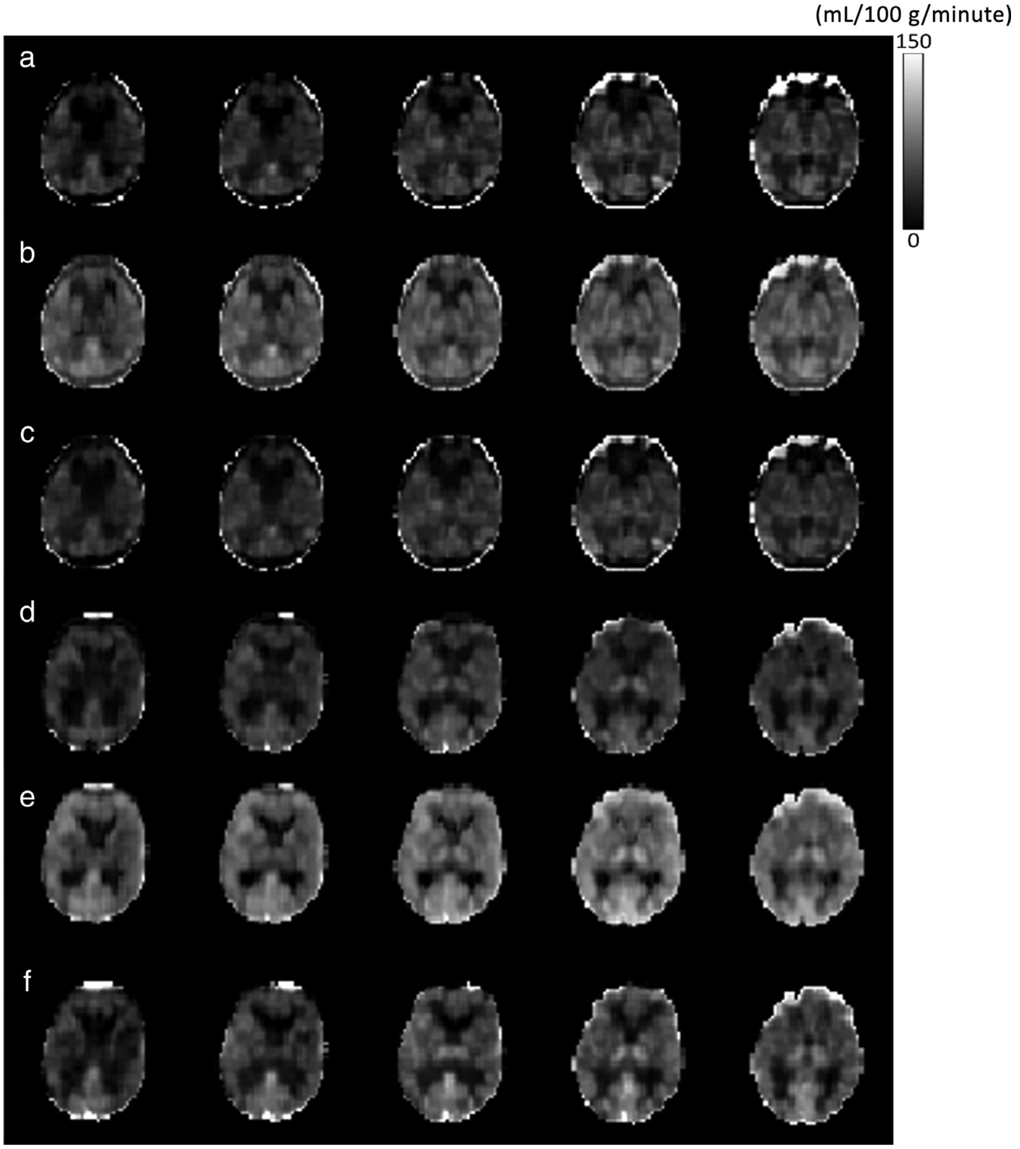

FIGURE 4: Mean CBF images of a representative AD subject and a NC subject produced by different methods on PASL sequences from Dataset 3. The rows from top to bottom are: (a) Output of the NonDL (pseudo ground truth) in Ref. 30 (input is an AD subject); (b) output of the DTF with the same AD subject as the input; (c) output of DLASLFT with the same AD subject as the inputs; (d) output of the NonDL (pseudo ground truth) in Ref. 30 (input is a NC subject); (e) output of DTF with the same NC subject as the input; (f) output of DLASLFT with the same NC subject as the input. CBF = cerebral blood flow; AD = Alzheimer's disease; NC = normal controls; PASL = pulsed arterial spin labeling.

processed by the NonDL, DTF, and DLASLFT. Directly transferring the DLASL trained with the young adults' PCASL data to the ADNI PASL data produced higher CBF values as reflected by the image intensity (Fig. 4b) than the NonDL (Fig. 4a). Fine-tuning the model yielded similar CBF values (Fig. 4c) to those by the NonDL (Fig. 4a). Both Fig. 4b,c showed improved image qualities as compared with Fig. 4a. For the NC subject, both DTF (Fig. 4e) and DLASLFT (Fig. 4f) produced CBF images with improved quality, compared with the NonDL (Fig. 4d). DLASLFT yielded similar CBF values to those generated by the NonDL, while DTF produced CBF images with higher image intensity than the NonDL did.

Quantitative Evaluation

Table 1 showed the CNRs of three different denoising methods on Dataset 2 and Dataset 3. Figures 5 and 6 showed the box plots of CNRs on both the PCASL sequences from Dataset 2 and PASL sequences from Dataset 3. Figure 5 revealed that DTF consistently achieved higher CNRs on both AD group and NC group of Dataset 2 than the NonDL.[30] The image quality improvement on both AD and NC groups of Dataset 2 by DTF was statistically significant as confirmed by the paired *t*-test. As shown in Fig. 6, both DTF and DLASLFT outperformed the NonDL. The paired *t*-test of DTF vs. the NonDL and DLASLFT vs. the NonDL was significant on both AD group and NC group of Dataset

3. Hence, the image quality improvement of the proposed two DL-based methods was statistically significant on both AD group and NC group of Dataset 3.

Figures 7 and 8 showed the resulting T-maps of the AD vs. NC CBF two-sample *t*-test using data from Dataset 2 (PCASL data) and Dataset 3 (PASL data), respectively. The statistical significance level was defined by the same threshold of $P < 0.001$ for all T-maps associated with each of the many two-sample *t*-tests. Both figures showed that DL methods yielded spatially more extended hypoperfusion patterns in the well-characterized frontal–parietal regions than the NonDL method. In Fig. 8, DTF only exhibited minor hypoperfusion detection sensitivity improvement. With model refining, DLASLFT (the third row of Fig. 8) showed substantially enlarged suprathreshold hypoperfusion clusters in the precuneus and parietal cortex. Mean CBF values, peak *t*–values, and the locations in the MNI space of the NC vs. AD CBF two-sample *t*-test were listed in Tables 2 and 3. Suprathreshold clusters with a size >100 were retained. In Table 2, DTF showed overall larger cluster sizes than that of the NonDL. In Table 3, DLASLFT generated clusters with larger sizes than the NonDL did.

TABLE 1. Contrast-to-Noise Ratio (CNR, Mean ± SD) of NonDL, DTF, and DLASLFT on Test Datasets

Data	Method	Group	CNR
Dataset 2			
	NonDL	AD	2.64 ± 0.62
	DTF	AD	3.38 ± 0.40
	NonDL	NC	3.36 ± 0.65
	DTF	NC	3.80 ± 0.41
Dataset 3			
	NonDL	AD	1.87 ± 0.70
	DTF	AD	3.82 ± 0.57
	DLASLFT	AD	2.45 ± 0.49
	NonDL	NC	2.17 ± 0.69
	DTF	NC	4.10 ± 0.64
	DLASLFT	NC	2.54 ± 0.49

AD = Alzheimer's disease; NC = normal controls.

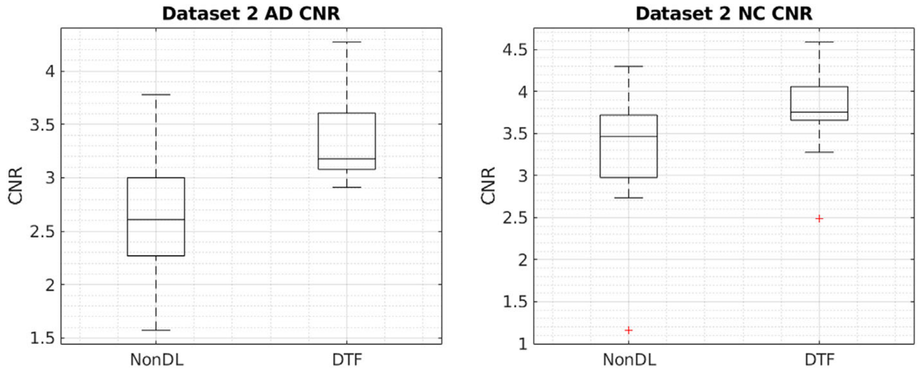

FIGURE 5: The box plot of the CNR from Dataset 2 (i.e., 21 AD subjects' CBF maps and 24 NC subjects' CBF maps) with different processing methods. CNR = contrast-to-noise ratio; AD = Alzheimer's disease; CBF = cerebral blood flow; NC = normal controls.

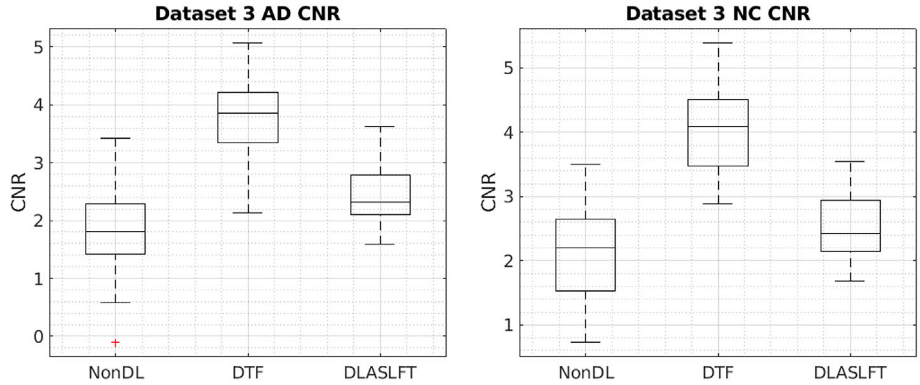

FIGURE 6: The box plot of the CNR from Dataset 3 (i.e., 53 AD subjects' CBF maps and 50 NC subjects' CBF maps) with different processing methods. CNR = contrast-to-noise ratio; AD = Alzheimer's disease; CBF = cerebral blood flow; NC = normal controls.

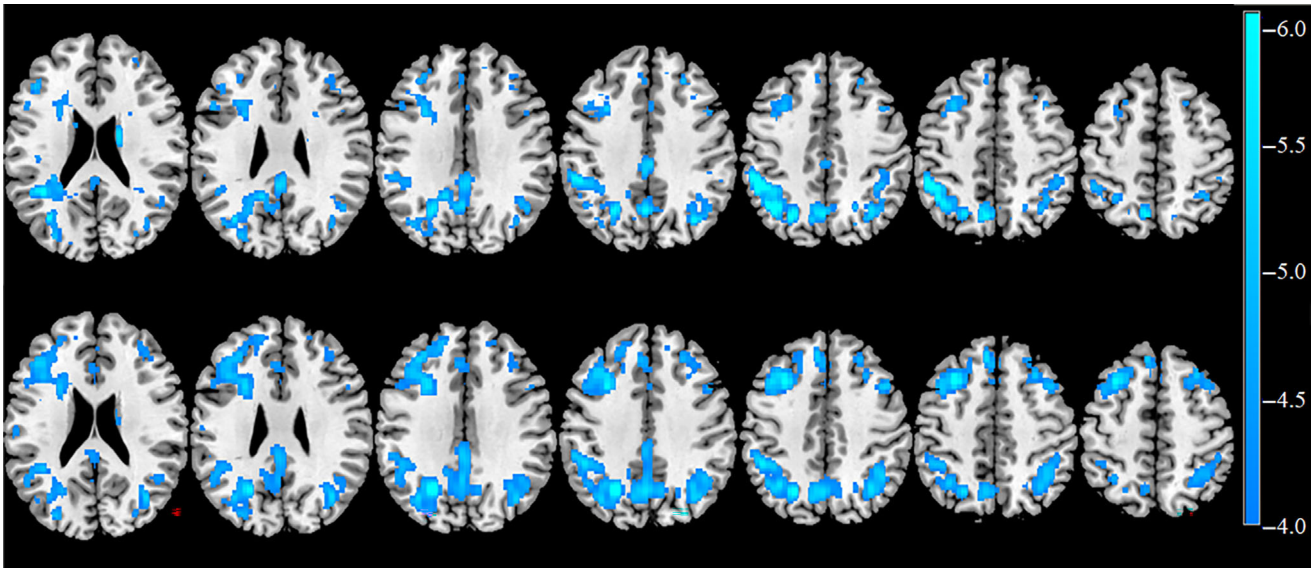

FIGURE 7: The resulting T map of two-sample t-test (AD vs. control) on Dataset 2. The top row shows the results obtained by the NonDL.[30] The bottom row shows the results obtained by the DTF. From left to right: slices 95, 100, 105, 110, 115, 120, and 125 in MNI space. Display window: [−4, −6]. P-value threshold is 0.001. AD = Alzheimer's disease; MNI = Montreal Neurological Institute.

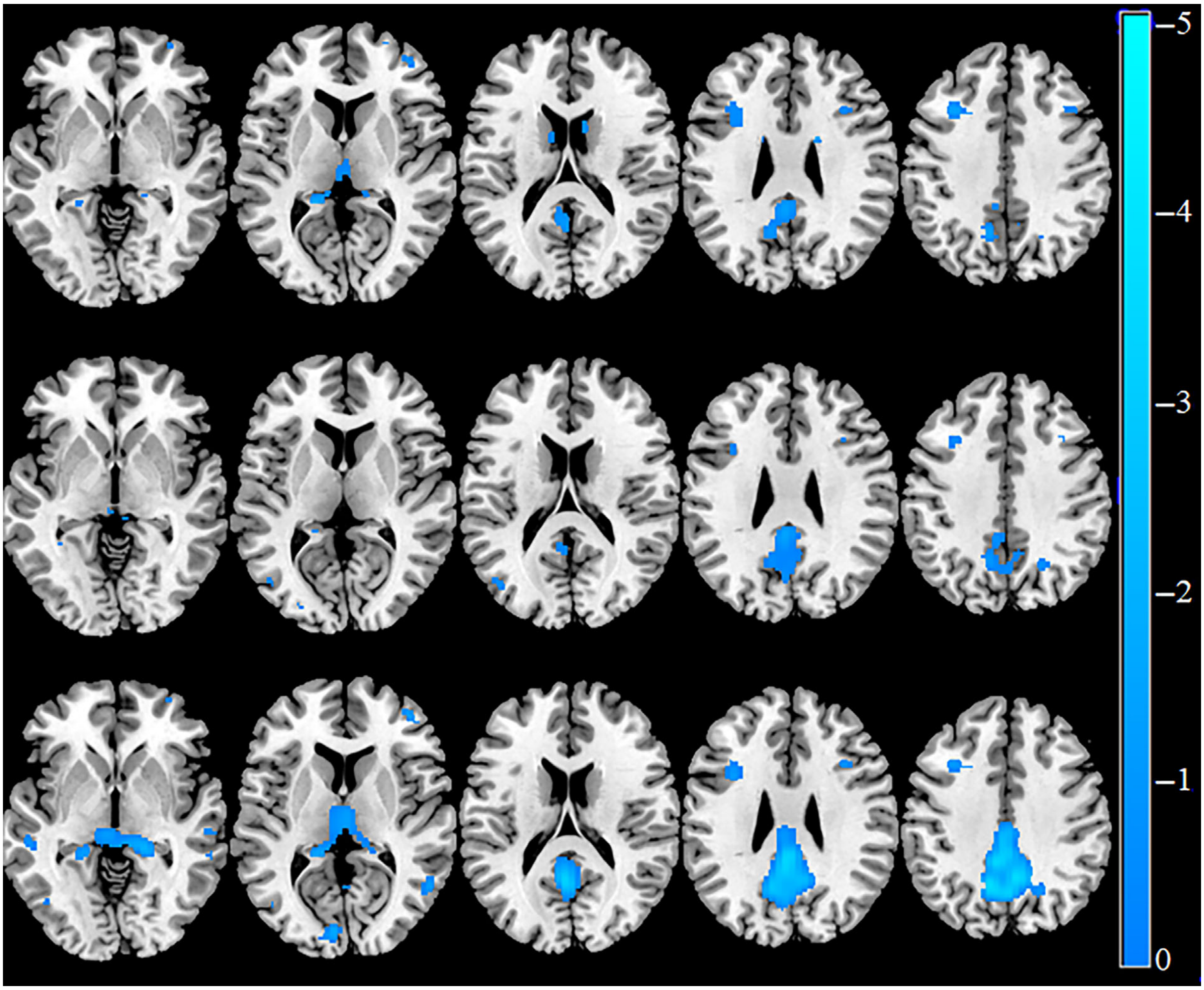

FIGURE 8: The resulting T map of two-sample t-test (AD vs. control) on Dataset 3. The top row, the middle row, and the bottom row show the results obtained by NonDL,[30] DTF, and DLASLFT, respectively. From left to right: slices 70, 80, 90, 100, and 110 in MNI space. Display window: [−3, −5]. P-value threshold is 0.001. AD = Alzheimer's disease; MNI = Montreal Neurological Institute.

TABLE 2. Cluster Size, Peak *t*-Value, Mean CBF (Mean ± SD) of Dataset 2

Method	Cluster Size	Peak *t*-Value	Peak *t*-Locations	AD Mean CBF	NC Mean CBF
NonDL	4274	−7.99	−50, −40, 44	24.44 ± 8.93	36.90 ± 11.12
	1463	−5.71	−16, 12, 18	20.11 ± 8.08	31.45 ± 9.71
	913	−5.82	−2, 50, −4	28.04 ± 5.46	41.55 ± 4.94
	795	−6.07	34, −68, 40	29.12 ± 2.41	43.19 ± 2.63
	470	−6.72	16, −12, 18	13.40 ± 5.03	26.19 ± 6.05
	422	−5.24	44, −12, 2	31.94 ± 3.73	44.37 ± 3.97
	253	−5.55	36, −40, 16	11.16 ± 3.82	19.64 ± 4.73
	239	−5.69	−38, 18, −2	33.85 ± 2.56	46.15 ± 2.58
	136	−4.91	36, 8, −42	21.93 ± 4.13	36.52 ± 4.00
	113	−5.10	26, 34, 42	28.85 ± 1.90	40.24 ± 1.30
	110	−4.71	60, −34, −22	16.88 ± 8.30	39.63 ± 10.58
	104	−5.62	50, −56, −14	24.71 ± 2.94	43.09 ± 1.93
DTF	4470	−6.38	−30, 18, 48	30.11 ± 4.67	39.80 ± 5.79
	2726	−7.36	−24, −62, 36	28.78 ± 5.08	38.06 ± 6.26
	1577	−5.56	34, −62, 44	30.48 ± 2.64	41.50 ± 3.48
	1487	−5.64	−2, −64, 40	33.92 ± 3.78	45.18 ± 5.00
	1079	−5.36	44, 14, 42	31.62 ± 2.55	42.34 ± 2.51
	495	−5.54	−6, 36, 48	31.25 ± 2.19	42.51 ± 2.25
	383	−4.72	−4, 52, −2	29.85 ± 1.67	40.24 ± 1.22
	313	−5.79	12, 8, 14	21.03 ± 2.10	28.62 ± 2.44
	130	−4.86	62, −36, −10	23.02 ± 6.65	33.85 ± 6.25
	116	−4.68	−60, −24, 18	34.23 ± 0.64	42.05 ± 0.79
	111	−5.48	−10, −92, 4	26.68 ± 2.04	34.11 ± 1.64

AD = Alzheimer's disease; CBF = cerebral blood flow; NC = normal controls.

Subject Image Quality Assessment

In Table 4, the average evaluation scores of 30 subjects provided by three reviewers for NonDL, DTF, and DLASLFT were 2.44, 2.02, and 2.86, respectively. Figure 9 showed the 100% stacked bar charts of three reviewers' reading scores of PASL CBF images from Dataset 3 processed by different methods.

The resulting *F*-statistic and *P*-values of ANOVA indicated the statistically significant differences between the compared methods. The corrected *P*-values of comparison between DLASLFT and DTF, comparison between DLASLFT and the NonDL, and comparison between DTF and NonDL suggested statistically significant differences in all pairwise group comparisons.

In the linear mixed-effects model, the *P*-value of the interaction between reviewers and denoising methods was 0.2215, indicating no significant interaction effect. We, therefore, refitted the model by removing the interaction term and the updated *P*-values for the two effect terms. The *P*-values indicated reviewers' effects were not statistically significant (*P*-value = 0.2) but denoising methods significantly affected the evaluation scores, i.e., the underlying transferability of different methods as a consequence.

Discussion

We have recently proposed a two-pathway-based DL network, the DLASL for ASL MRI denoising.[4] In the original

TABLE 3. Cluster Size, Peak *t*-Value, Mean CBF (Mean ± SD) of Dataset 3

Method	Cluster Size	Peak *t*-Value	Peak *t*-Locations	AD Mean CBF	NC Mean CBF
NonDL	993	−4.3021	−20, 12, 28	13.46 ± 11.01	22.01 ± 11.93
	674	−3.7147	−12, −58, 34	34.73 ± 6.00	46.12 ± 6.83
	530	−3.77	−18, −42, 6	19.71 ± 6.86	30.19 ± 6.45
	268	−3.9791	16, 4, 26	3.49 ± 2.39	10.82 ± 2.37
	138	−3.5764	36, 20, 44	24.69 ± 3.25	37.08 ± 2.49
DTF	1135	−3.6118	−2, −42, 30	32.99 ± 3.26	38.52 ± 3.44
	101	−3.5874	−56, −24, −14	34.35 ± 1.26	38.96 ± 1.30
DLASLFT	3520	−4.6833	0, −40, 32	34.39 ± 5.86	46.08 ± 7.01
	1331	−4.0005	6, −36, 2	26.78 ± 6.58	36.49 ± 6.43
	406	−4.559	−56, −26, −10	38.02 ± 3.16	48.82 ± 3.76
	326	−3.4955	−34, 10, 26	27.37 ± 6.40	35.97 ± 7.35
	121	−3.6676	42, 48, 6	27.79 ± 3.48	41.20 ± 2.77
	112	−3.8074	−6, −88, 4	37.55 ± 7.42	48.29 ± 7.37

AD = Alzheimer's disease; CBF = cerebral blood flow; NC = normal controls.

paper, DLASL was trained on CBF images calculated from PCASL data from a group of young healthy subjects. Although DLASL showed promising denoising results in young healthy subjects, its benefit for older subjects including patients with AD has not been examined. Also, since ASL MRI data have been acquired using different sequences, another important question is whether DLASL can be generalizable to ASL CBF images acquired with different ASL MRI sequences. The purpose of this study was to address these two questions. Our strategy was to use direct transfer learning or transfer learning with moderate model fine-tunning. Two major types of ASL (i.e., PCASL and PASL) data from normal healthy elderlies and AD patients were tested. Our results showed that the NonDL which improved the ASL image quality by removing outlier slices performed the worst on both Dataset 2 and Dataset 3. Priors played an important role in the algorithm design. If the datasets did not match the priors, the outlier slices may not be removed efficiently. While DL-based methods were data-driven. We showed that DLASL can be directly transferred from young adults' data to data from older subjects as well as AD patients. DLASL showed increased SNR and improved sensitivity for using the corresponding CBF images for detecting the well-characterized fronto-parietal brain hypoperfusion patterns in AD as compared with NC. For different types of ASL data, our results showed that model fine-tuning should be performed since the direct transfer learning only showed minor sensitivity improvement.

DLASL transfer learning yielded better CNR for the NC and AD subjects' CBF map than with NonDL. It also better preserved image resolution than NonDL[30] in terms of less blurring and better preservation of tissue boundaries. This performance gain is caused by the convolutional feature extraction in DLASL and the two pathways-based DL

TABLE 4. Three Radiologists' Scores (Mean ± SD) of 30 Subjects Selected From Dataset 3

	NonDL	DTF	DLASLFT
Reviewer #1	2.27 ± 0.91	2.23 ± 0.63	2.50 ± 0.73
Reviewer #2	2.27 ± 0.87	1.83 ± 1.05	2.67 ± 0.84
Reviewer #3	2.80 ± 0.89	2.00 ± 0	3.40 ± 0.56
Average of 3 reviewers	2.44 ± 0.91	2.02 ± 0.72	2.86 ± 0.82

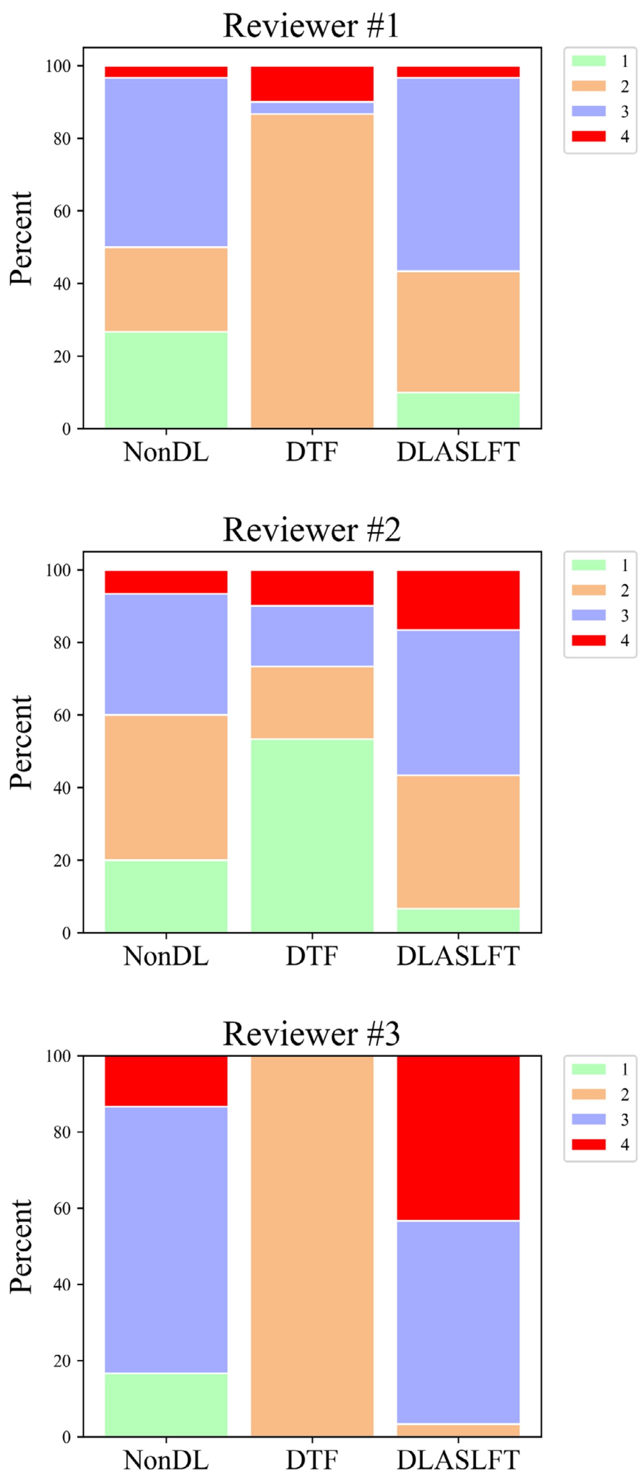

FIGURE 9: Comparison of reading score between different methods over 30 subjects' CBF maps using 100% stacked bar chart. Reading scores are displayed with four different colors. CBF = cerebral blood flow.

network structure as well as the dilated wide activation residual block.[4] The convolution-based feature extraction can suppress random noise. The dilated wide activation residual block can extract data features from different aspects and from non-local places.[4] In this study, the global pathway was used to capture global CBF patterns, and the local pathway was to preserve local patterns. The former contributes to higher SNR, and the latter preserves local tissue structure. Together, they contribute to improved CNR and better image details.[4]

The CBF map quality improvement by DLASL subsequently resulted in higher sensitivity for detecting the typical hypoperfusion patterns in AD patients as compared with NC. Hypoperfusion was statistically defined by the two-sample *t*-test on CBF images of NC and AD. Both PCASL and PASL data proved the direct transferability of DLASL. This is because PCASL and PASL share the theoretical background and are both designed for measuring the quantitative CBF though they differ by ASL schemes and SNR.[31] The common data features of the two types of data also explained why the training loss and validation loss during fine-tuning decreased faster than those during training from scratch. Direct transfer and transfer learning with fine-tuning showed a big denoising performance difference. The former showed better CNR in the resultant CBF images than the latter, but the latter presented higher hypoperfusion detection sensitivity. This difference can be attributed to the large SNR difference between PCASL and PASL.[37] PCASL has much higher SNR than PASL. Young healthy subjects have higher CBF values than old subjects and AD patients. Likely, the old subjects and AD patients' PASL data cannot be fully modeled by the data features learned from the young subjects' PCASL data, resulting in a superficial over-denoising. Fine-tuning the model with old subjects' PASL data helped refining the learned data distribution, which explained why the CNR dropped after fine-tuning but the AD vs. NC CBF difference (hypoperfusion) detection sensitivity increased. Because the sample size of the PASL data was much smaller than the young adults' PCASL data (Dataset 1), the risk of overfitting to PASL data is low even when all PASL data were included as new training data during fine-tuning. In fact, if PASL data overfitting occurred, the DLASL output would be the same as the input PASL data, and the hypoperfusion patterns would be the same as the NonDL processed CBF data. The early stopping strategy we used also reduced the risk of data overfitting.

After the three reviewers completed the image quality evaluation, we noticed that reviewer #3 assigned the same score of 2 to DTF results across all 30 subjects (both AD and NC). We conjectured that the images generated by DTF were with contrasts that were different from the other two methods. The DTF model achieved the best CNR in Dataset 3, but it performed the worst on Dataset 3 in terms of radiologic score. Our explanation was that the DTF model derived the knowledge from the PCASL sequence of Dataset 2. When DTF was used to denoise the PASL sequence images in Dataset 3, it would generate denoised results as the PCASL sequence. The best CNR of the DTF model was attributed to the training PCASL data in Dataset 1 were from younger healthy subjects (age: 23–47) with stronger blood flow signal than subjects in Dataset 3 (age: 59–87). In addition, the higher SNR of PCASL compared with PASL[31] also contributed to the higher CNR of DTF than the NonDL and DLASLFT. The worst radiologic score of DTF was because that DTF model can only generate results like healthy subjects of PCASL sequence which were visually different from the older subjects of PASL sequence. Since the DTF model was trained using PCASL images instead of PASL images, DTF model cannot perform well on unseen PASL images. So, it was necessary to do the model fine-tuning on PASL images, i.e., DLASLFT. And DLASLFT (i.e., DLASL with fine-tuning) outperformed NonDL method on PASL images, which is consistent with the previous study[4] that showed DLASL outperformed NonDL on PCASL images in terms of the radiologic score.

Limitations

First, we only used 2D slices instead of the 3D ASL MRI scans due to the complexity involved in the 3D DL model.[38] We did not take advantage of the correlation between slices. Second, the AD and NC subjects included in this study may still not be sufficient to cover the entire population. We have planned to evaluate the proposed method on a larger cohort in the foreseeable future. Finally, we need further rigorous validation on a larger study cohort of the transfer learning of the DL ASL denoising, which is another direction of our future research.

Conclusion

DLASL trained in healthy subjects' PCASL data can be directly applied to older healthy subjects' and AD patients' PCASL data. Model fine-tuning is recommended for using the model trained from young healthy subjects' data to NC and AD's PASL acquired at different sites. DLASL is clinically valuable for improving ASL CBF sensitivity for detecting the NC vs. AD CBF difference.

Acknowledgments

The research effort involved in this study was supported by NIH grants: R01AG060054, R01 AG070227, R01EB031080-01A1, P41EB029460-01A1. Data collection and sharing for the ADNI PASL data were 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 Neuroimaging at the University of Southern California. The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

1. Detre JA, Leigh JS, Williams DS, Koretsky AP. Perfusion imaging. Magn Reson Med 1992;23(1):37-45.
2. Detre JA, Rao H, Wang DJJ, Chen YF, Wang Z. Applications of arterial spin labeled MRI in the brain. J Magn Reson Imaging 2012;35(5):1026-1037.
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 2015;73(1):102-116.
4. 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.
5. Wang Z. Improving cerebral blood flow quantification for arterial spin labeled perfusion MRI by removing residual motion artifacts and global signal fluctuations. Magn Reson Imaging 2012;30(10):1409-1415.
6. 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.
7. Liang X, Connelly A, Calamante F. Voxel-wise functional connectomics using arterial spin labeling functional magnetic resonance imaging: The role of denoising. Brain Connect 2015;5(9):543-553.
8. Hu WT, Wang Z, Lee VM-Y, Trojanowski JQ, Detre JA, Grossman M. Distinct cerebral perfusion patterns in FTLD and AD. Neurology 2010; 75(10):881-888.
9. Dominic C, Harston GWJ, Garrard J, et al. ICA-based denoising for ASL perfusion imaging. Neuroimage 2019;200:363-372.
10. Wang Z. Support vector machine learning-based cerebral blood flow quantification for arterial spin labeling MRI. 2014.
11. 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.
12. Gong E, Pauly J, Zaharchuk G. Boosting SNR and/or resolution of arterial spin label (ASL) imaging using multi-contrast approaches with multi-lateral guided filter and deep networks. In: *Proceedings of the*

Annual Meeting of the International Society for Magnetic Resonance in Medicine, Honolulu, Hawaii; 2017.

13. Kim B, Schär M, Park HW, Heo H-Y. A deep learning approach for magnetization transfer contrast MR fingerprinting and chemical exchange saturation transfer imaging. Neuroimage 2020;221:117165.

14. Zaharchuk G, Gong E, Wintermark M, Rubin D, Langlotz CP. Deep learning in neuroradiology. Am J Neuroradiol 2018;39(10):1776-1784.

15. Xie D, Li Y, Yang H, et al. Estimating cerebral blood flow from BOLD signal using deep dilated wide activation networks. ISMRM 2020; 2020.

16. Xie D, Li Y, Yang H, Bai L, Wang Z. A learning-from-noise dilated wide activation network for denoising arterial spin labeling (ASL) perfusion images. ISMRM 2020; 2020.

17. Xie D, Li Y, Yang H, Wang Z. SuperASL: Improving SNR and temporal resolution of ASL MRI using deep learning. In: *ISMRM Workshop on Machine Learning*; 2018.

18. Liu Q, Shi J, Wang Z. Increasing arterial spin labeling perfusion image resolution using convolutional neural networks with residual-learning. ISMRM 2018; 2018.

19. 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: *International Conference on Medical Image Computing and Computer-Assisted Intervention*; 2018, p 30–38.

20. Li Z, Liu Q, Li Y, et al. A two-stage multi-loss super-resolution network for arterial spin labeling magnetic resonance imaging. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*; 2019, p 12–20.

21. Alsop DC, Dai W, Grossman M, Detre JA. Arterial spin labeling blood flow MRI: Its role in the early characterization of Alzheimer's disease. J Alzheimers Dis 2010;20(3):871-880.

22. Wang Z. Characterizing early Alzheimer's disease and disease progression using hippocampal volume and arterial spin labeling perfusion MRI. J Alzheimers Dis 2014;42(s4):S495-S502.

23. Wu W-C, Fernández-Seara M, Detre JA, Wehrli FW, Wang J. A theoretical and experimental investigation of the tagging efficiency of pseudocontinuous arterial spin labeling. Magn Reson Med 2007;58(5): 1020-1027.

24. Luh WM, Wong EC, Bandettini PA, Hyde JS. QUIPSS II with thin-slice TI1 periodic saturation: a method for improving accuracy of quantitative perfusion imaging using pulsed arterial spin labeling. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine 1999;41(6):1246-1254.

25. Wang Z, Aguirre GK, Rao H, et al. Empirical optimization of ASL data analysis using an ASL data processing toolbox: ASLtbx. Magn Reson Imaging 2008;26(2):261-269.

26. Fonov VS, Evans AC, McKinstry RC, Almli CR, Collins DL. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. Neuroimage 2009;47:S102.

27. Fonov V, Evans AC, Botteron K, et al. Unbiased average age-appropriate atlases for pediatric studies. Neuroimage 2011;54(1): 313-327.

28. Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. In: *International Conference on Learning Representations*; 2016.

29. Huber PJ. Robust estimation of a location parameter. *Breakthroughs in statistics*. New York: Springer; 1992. p 492-518.

30. Li Y, Dolui S, Xie D-F, Wang Z, Alzheimer's Disease Neuroimaging Initiative. Priors-guided slice-wise adaptive outlier cleaning for arterial spin labeling perfusion MRI. J Neurosci Methods 2018;307:248-253.

31. Chen Y, Wang DJJ, Detre JA. Test–retest reliability of arterial spin labeling with common labeling strategies. J Magn Reson Imaging 2011;33(4):940-949.

32. Chollet F, et al. Keras. GitHub; 2015. Available from: https://github.com/fchollet/keras

33. Kingma DP, Jimmy BA. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. 2014.

34. Prechelt L. Early stopping-but when? *Neural networks: Tricks of the trade*. Berlin: Springer; 1998. p 55-69.

35. Abadi M, Barham P, Chen J, et al. Tensorflow: A system for large-scale machine learning. In: *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*; 2016, p 265–283.

36. Welvaert M, Rosseel Y. On the definition of signal-to-noise ratio and contrast-to-noise ratio for fMRI data. PLoS One 2013;8(11):e77089.

37. Wong EC. Potential and pitfalls of arterial spin labeling based perfusion imaging techniques for MRI. In: Moonen CTW, Bandettini PA, editors. *Functional MRI*; Heidelberg: Springer-Verlag; 1999. p 63-69.

38. Dou Q, Chen H, Yu L, et al. Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks. IEEE Trans Med Imaging 2016;35(5):1182-1195.