Published in final edited form as:

IEEE Trans Med Imaging. 2020 November ; 39(11): 3691–3702. doi:10.1109/TMI.2020.3002708.

Hierarchical Nonlocal Residual Networks for Image Quality Assessment of Pediatric Diffusion MRI with Limited and Noisy Annotations

Siyuan Liu [Member, IEEE],

Department of Radiology and Biomedical Research Imaging Center (BRIC), University of North Carolina at Chapel Hill, NC, U.S.A.

Kim-Han Thung,

Department of Radiology and Biomedical Research Imaging Center (BRIC), University of North Carolina at Chapel Hill, NC, U.S.A.

Weili Lin,

Department of Radiology and Biomedical Research Imaging Center (BRIC), University of North Carolina at Chapel Hill, NC, U.S.A.

Dinggang Shen [Fellow, IEEE],

Department of Radiology and Biomedical Research Imaging Center (BRIC), University of North Carolina at Chapel Hill, NC, U.S.A.

Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea.

Pew-Thian Yap [Senior Member, IEEE]

Department of Radiology and Biomedical Research Imaging Center (BRIC), University of North Carolina at Chapel Hill, NC, U.S.A.

Abstract

Fast and automated image quality assessment (IQA) for diffusion MR images is a crucial step for swiftly making a rescan decision during or after the scanning session. However, learning a model for this task is challenging as the number of annotated data is limited and the annotation labels might not always be correct. As a remedy, we will introduce in this paper an automatic image quality assessment (IQA) method based on hierarchical non-local residual networks for pediatric diffusion MR images. Our IQA is performed in three sequential stages, i.e., 1) *slice-wise IQA*, where a nonlocal residual network is first pre-trained to annotate each slice with an initial quality rating (i.e., pass/questionable/fail), which is subsequently refined via iterative semi-supervised learning and slice self-training; 2) *volume-wise IQA*, which agglomerates the features extracted from slices of a volume, and uses a nonlocal network to annotate the quality rating for each volume via iterative volume self-training; and 3) *subject-wise IQA*, which ensembles the volumetric IQA results to determine the overall image quality pertaining to a subject. Experimental results demonstrate that our method, trained using only samples of modest size,

exhibits great generalizability, and is capable of conducting rapid hierarchical IQA with near-perfect accuracy.

Index Terms—

Image quality assessment; hierarchical nonlocal residual networks; semi-supervised learning; self-training

I. Introduction

Diffusion magnetic resonance imaging (dMRI) is a powerful tool for probing tissue microstructure and investigating white matter pathways [1–3]. However, it is susceptible to imaging artifacts caused by motion, which is especially common for pediatric subjects. Artifacts could bias analyses, invalidate conclusions, and weaken interpretability [4, 5]. Therefore, extra care must be taken in ensuring data quality prior to subsequent processing and analysis.

Image quality assessment (IQA) is key in determining whether an acquired image is usable and whether a re-scan is necessary. Further decisions can be made based on IQA to exclude affected slices, volumes, and subjects or to correct for artifacts. IQA can be carried out either subjectively via visual inspection by a human rater, or objectively by a computer algorithm [6, 7]. Subjective IQA can be time-consuming, labor-intensive, costly, error-prone, and inconsistent [8]. IQA is especially challenging in dMRI as typically a series of volumes are acquired for each subject. In time-sensitive studies involving for example newborns, scheduling rescans is often not an option since the brain changes literally on a daily basis. Therefore, a fast, reliable, accurate, and fully-automated objective IQA for dMRI is highly desirable.

Generally, objective IQA can be grouped into three categories based on the availability of reference images, i.e., 1) full-reference IQA (FR-IQA), which requires a pristine image as reference; 2) reduced-reference IQA (RR-IQA), which requires partial information from a reference image; and 3) no-reference IQA (NR-IQA), which requires no information from the reference image. Specifically, FR-IQA methods measure the quality of an image by comparing it with a reference using evaluation metrics [9–12]. RR-IQA methods utilize only a limited number of features extracted from a reference [13] to provide near FR-IQA performance. However, FR-IQA and RR-IQA have limited practical applications as full or partial information from a pristine reference image is not always available. In contrast, NR-IQA methods [14–17] do not require any information from a reference image and is well-suited for practical applications. Currently, most NR-IQA methods are designed for natural 2D images, while diffusion MR images are 3D with intensity distributions and artifacts that differ greatly from natural images.

Recently, deep neural networks (DNNs), particularly convolutional neural networks (CNNs), have been demonstrated to have great potential for IQA [18]. Instead of hand-crafting features always developed in-house [19], CNNs automatically learn image features that are pertinent to IQA. However, the performance of these deep learning methods generally

depend on the number of labeled training samples and the correctness of their corresponding labels. Medical images, particularly diffusion MR images, are typically limited on the number of annotated images, as annotating them requires a huge amount of efforts from experts. For example, it is labor-intensive and time-consuming to annotate quality scores for all the image volumes in dMRI of all the subjects to train a volume-wise IQA network. To address the issue of limited annotated data, Graham et al. [20] proposed to use simulated images to augment training data. However, simulated images may not fully capture the characteristics of real acquired images. Another less labor-insensitive alternative is to annotate each volume (or scan session) with a single quality label. This is, however, inaccurate and will introduce label noise as each volume (or scan session) might inevitably contain a mixture of good or bad image slices (or volumes). Furthermore, the label noises might also be introduced by 1) mislabeling or unreliable labeling due to human error; and 2) inconsistent labeling among different experts.

In this paper, on the basis of our preliminary version [21], we address the above issues by proposing a deep learning based hierarchical IQA method that requires only a small amount of annotated volumes for training and is robust to annotation errors. To the best of our knowledge, this is the first deep learning based dMRI IQA method that caters to limited annotated images and label noise. The key features of our method are summarized as follows:

1. Our method consists of three sequential stages: slice-wise IQA using a slice-wise QA network (SQA-Net), volume-wise IQA by agglomerating the extracted slice features using a volume-wise QA network (VQA-Net), and subject-wise IQA by ensembling the volume-wise IQA results using a decision rule. Unlike previous deep learning based dMRI IQA methods that only provide volume assessment, our method simultaneously provides assessment results at different levels, i.e., slice, volume and subject.
2. For better computational efficiency and feature expression capability, we construct SQA-Net by incorporating depthwise separable convolutions (DSConv) and nonlocal mean operation into residual blocks. Similarly, VQA-Net consists of a nonlocal block and a classifier block. The computation-reduction property of DSConv and the information fusion property of nonlocal block enable SQA-Net and VQA-Net to be trained efficiently using only a small number of data, making the networks suitable for real-time IQA.
3. We utilize semi-supervised learning to deal with the scenario where we have a small amount of labeled data but a large amount of unlabeled data. Specifically, using a small number of labeled samples, we pre-train SQA-Net, and then use it to select confident slices from unlabeled slices, which are then used together with the labeled slices to re-train SQA-Net. This labeling and re-training process is iterated until the annotation of unlabeled slices stabilizes.
4. We employ an iterative self-training mechanism to remove or relabel unreliable labels to improve assessment accuracy of SQA-Net and VQA-Net.

The rest of this paper is organized as follows. In Section II, we briefly review related work on learning with limited labeled data affected by label noise. Section III describes the architectures of slice-/volume-wise quality assessment networks, semi-supervised learning, and slice/volume self-training. We report the experimental results in Section IV and conclude in Section VI.

II. Related Work

A. Learning with Limited Labeled Data

When only a small quantity of labeled data are available, the parameters of CNNs may not be learned properly. This problem can be mitigated by either 1) data augmentation, 2) pre-training the network using an alternative but related large dataset and then fine-tuning it using the original small dataset, or 3) label the unlabeled data.

Data augmentation is a popular approach to significantly increase the number of samples to train an effective deep learning network. This strategy has been used for dMRI IQA in [20] with data simulated using noise and motion artifacts. However, the trained model is only sensitive to the simulated artifacts, and cannot detect other types of artifacts in dMRI data.

Transfer learning methods adopt the second idea [22, 23]. Network weights are first initialized by pre-training the network using a large related dataset, and are then fine-tuned using a relatively smaller dataset related to the target task. However, the effectiveness of transfer learning might be deteriorated when the two datasets are only loosely related. This is especially true for dMRI IQA, where a larger alternative dataset with matching characteristics is seldom available.

Semi-supervised learning [24] adopt the third idea and trains networks/models with a small amount of labeled data and a large amount of unlabeled data. Strategies to label unlabeled data can be divided into three classes: generative methods [25], graph methods [26, 27], pseudo-labeling methods [28, 29]. *Generative methods* assume a label-parameterized model for labeled or unlabeled data. However, performance degrades with model inaccuracy. *Graph methods* represent each labeled or unlabeled data sample as a node in a graph and define the edge between two nodes as sample similarity computed with a Gaussian kernel. Fixing the labels of labeled data, the labels of unlabeled data are iteratively updated until convergence. However, graph methods have large storage requirements and the graph has to be reconstructed and updated for new data. *Pseudo-labeling methods* assign initial pseudo-labels to unlabeled data via an initial model that is trained using a small-scale labeled dataset. Then, the pseudo-labels of the unlabeled data are updated via an optimized model that is obtained by retraining the initial model using both labeled and pseudo-labeled datasets.

B. Learning with Noisy Labels

Noisy labels can result in performance deterioration and increase in model complexity [30]. Machine learning methods proposed to address this issue can be grouped into two categories: direct and indirect learning with noisy labels.

Direct learning with noisy labels: Some researches proposed to use reweighted or rectified loss functions that are robust to label noises [31–33], such as ramp loss [34] and unhinged loss [35]. However, it is shown in [36] that these loss functions are not completely robust to label noise.

An alternative is to utilize classifier algorithms that are robust to label noise, such as SVM, and random forest. Nevertheless, the robustness of these methods is limited, as shown empirically in [37]. Label noise models can also be embedded into the classification model to alleviate the influence of label noise [38–43]. However, this approach is largely dependent on the availability and quality of prior label noise information. Furthermore, this approach increases the complexity of learning algorithms and may result in overfitting due to the additional model parameters.

Indirect learning with noisy labels: Some label noise cleansing methods are proposed to remove or relabel problematic samples before training [44, 45]. However, the removal operation may lead to overcleansing in data-imbalance cases [46]. That is, minority classes might be entirely removed, causing degradation in classification performance. In comparison, relabeling mislabeled data maintains the sample size, but incorrect relabeling may also lead to performance degradation. In this context, our method adopts the label noise cleansing method to simultaneously maintain the sample size and remove unreliable data.

III. Method Architecture

Fig. 1 shows an overview of our dMRI IQA method, which consists of three stages: slice-wise, volume-wise, and subject-wise IQA stages. The slice-wise IQA stage is designed to predict the quality rating of each slice, which involves two iterative processes, i.e., semi-supervised learning and slice self-training. The volume-wise IQA stage is introduced to evaluate the quality of each volume based on the quality of the constituent slices. Volume-wise IQA is iterative refined via a self-training algorithm. The subject-wise IQA stage agglomerates the quality ratings of volumes belonging to this subject. The details on each component of our dMRI IQA method will be described next.

A. Slice Quality Assessment Network (SQA-Net)

Our slice quality assessment network (SQA-Net) is designed with both accuracy and speed in mind. Fig. 2 shows the proposed network, consisting of four types of network blocks, i.e., depthwise separable residual (DSR) block (in light pink), nonlocal depthwise separable residual (NLDR) block (in light blue color), nonlocal residual (NLR) block (in orange), and classifier (CLF) block (in green). The DSR blocks and NLDR blocks extract low- and high-level features, respectively, whereas the NLR block [47] computes the response function at each position as a weighted summation of all features in the feature maps. The CLF block outputs three probability values indicating whether a slice is “pass”, “questionable”, or “fail”. The number of channels for each network block from the first DSR block to CLF block is 32, 64, 28, 256, 256, and 3, respectively.

Our IQA framework is holistic in that it relies on both local and global information. That is, it considers relationship not only among local features, but also among non-local features.

While a typical convolutional layer can extract local information easily, we need multiple layers of convolution operations to extract non-local information. This is inefficient, computationally expensive, and unable to directly capture dependencies. Thus, most of our network blocks adopt two major network architectures, i.e., the non-local network architecture, where the response function at each position is computed as the weighted summation of features at both short-range (local) and long-range (global) positions, and the residual network architecture, which is devised to avoid the vanishing/exploding gradient issue caused by the increase of network layers and connections. In this way, our network can assess the quality of an image holistically to output a more accurate rating.

The DSR block is constructed by integrating 3×3 depthwise separable convolutions (DSCConv) layers [48] with unit stride into a residual block, where each DSCConv layer is followed by batch normalization (BN) and rectified linear unit (ReLU) activations, and a 2×2 max-pooling layer is adopted to down-sample the intermediate features. The shortcut branch of DSR block is implemented by a 1×1 standard convolution with a 2×2 stride.

The NLDR block is based on DSR block, where a nonlocal network block is added before max-pooling layer. The non-local network block computes the output at the i -th location $r_i \in \mathbb{R}^c$ of the nonlocal network block as the weighted sum of all input features for a c -channel $h \times w$ input feature map x , given as

$$r_i = \frac{1}{C_i(x)} \sum_{\forall j} f(x_i, x_j) g(x_j), \quad (1)$$

where the weight function $f(\cdot)$ computes the pairwise relationship (e.g., similarity) between feature vectors at two locations, $g(\cdot)$ computes a representation of the input feature, and $C_i(x) = \sum_{\forall j} f(x_i, x_j)$ is a normalization factor. In this work, the weight function $f(x_i, x_j)$ is defined in the following embedded Gaussian form:

$$f(x_i, x_j) = \exp \phi^T(x_i) \psi(x_j), \quad (2)$$

where $\phi(\cdot)$ and $\psi(\cdot)$ are the unary kernel functions that are implemented with 1×1 convolution kernels with unit stride, thus making $\frac{1}{C_i(x)} f(x_i, x_j)$ a softmax function in implementation.

The NLR block is a combination of nonlocal and residual network architectures. The CLF block is realized with a 3×3 convolution layer with unit stride, global average pooling (GAP), and softmax activation function.

B. Volume Quality Assessment Network (VQA-Net)

Using the output of the NLR block in SQA-Net (i.e., some kind of high-level features) for slices belonging to the same volume as input, we design a volume quality assessment network (VQA-Net) to determine the quality for this volume. Specifically, we first employ a 1×1 convolution to reduce the channel dimension of the feature representation for each slice from 256 to 16. Then, using the concatenated dimension-reduced features as inputs, we

subsequently employ a NLR block and a CLF block to output the quality label (i.e., “pass”, “questionable” or “fail”) for the current volume based on the quality rating with the highest probability.

C. Semi-Supervised Learning in SQA-Net

As mentioned previously, we have limited number of labeled samples. Thus, we employ semi-supervised learning when training SQA-Net (orange box in Fig. 1) to make full use of plentiful unlabeled data by progressively selecting confident slices from the unlabeled dataset to be included in labeled dataset for retraining the SQA-Net.

Specifically, we first utilize SQA-Net that has been pre-trained with the labeled dataset to predict the quality label for each slice of unlabeled volumes. We then pseudo-label each slice with the predicted quality label. Slices with high confident pseudo-labels (maximal probability greater than a predefined threshold, i.e., 0.9) are then merged with the labeled dataset to retrain SQA-Net. This is done iteratively until the network converges, i.e., when performance improvement is minimal.

D. Slice Self-Training

We only have volumetric label and assuming the label for each slice in a volume to be the same, which is inaccurate. To deal with inaccurate slice labels, we propose a slice self-training mechanism that involves iterative slice relabeling/removal and SQA-Net retraining (green box in Fig. 1). Specifically, we first predict the quality of each slice using the pre-trained SQA-Net and then keep slices that meet one of the following conditions: 1) the predicted label is identical to the prediction in previous iteration; 2) the label is predicted with high-confidence, i.e., its prediction maximal probability beyonds a threshold. In addition to removing slices that do not meet the above criteria, we also relabel the selected slices with high-confidence predicted labels. We then retrain SQA-Net for the next iteration until accuracy improvement is minimal.

E. Volume Self-Training

VQA-Net is employed to predict the volumetric quality based on the slice features extracted by SQA-Net. Both labeled and unlabeled volumes are utilized to train VQA-Net, where the initial quality ratings of unlabeled volumes are determined based on the following rules: 1) “Pass” if more than 60 percent of the slices in the volume are labeled as “pass”; 2) “Fail” if more slices are labeled as “fail” than “pass” and “questionable”; 3) “Questionable” if otherwise. However, the initial volumetric quality determined by the empirical rules is inaccurate due to the indistinct bounds of each class for different subjects. To cope with inaccurate volumetric labels, similar to slice self-training, volume self-training involves iterative volume relabeling/removal, i.e., VQA-Net retraining (blue box in Fig. 1). Specifically, we retrain our VQA-Net with volumetric data where 1) the predicted labels are identical to the prediction in previous iteration; or 2) the labels are predicted with high-confidence, i.e., maximal probabilities beyond a threshold. We also replace the labels of the selected volumes with the predicted labels, and remove volumetric data that do not meet the above criteria from the training dataset. We then retrain VQA-Net and repeat the whole process until the accuracy improvement is minimal.

F. Subject Assessment

To ensure that image quality is sufficient for subsequent processing [49], we assign quality label for each subject by using the following rules: 1) “Pass” if more than 80 percent of the volumes in a subject is labeled as “pass”; 2) “Fail” if more than 20 percent of the volumes in a subject is labeled as “fail”; 3) “Questionable” if otherwise, where the volumetric quality ratings are annotated by the trained VQA-Net.

G. Loss Function

We also use L_2 regularization for the Conv and DSConv layers to avoid overfitting. In addition, to deal with data imbalance, we use a multi-class balanced focal loss [50] with L_2 regularization for both SQA- and VQA-Net, i.e.,

$$\mathcal{L}(p_t) = -\alpha_t(1-p_t)^\kappa \log(p_t) + \frac{\lambda}{2n_w} \sum_w \|\mathbf{W}_w\|_2^2 \quad (3)$$

where p_t ($t=1, 2, 3$) are the predicted probabilities for “pass”, “questionable”, and “fail”, respectively. $\kappa > 0$ is a focusing parameter, $\mathbf{W}_w$ is the w -th weight matrices of SQA-/VQA-Net, $\lambda = 0.01$ is a tuning parameter for L_2 regularization, and n_w is the number of weight matrices. Here, the class weights $\alpha_t = \max(N_1, N_2, N_3)/N_t$ are used for balancing the contributions of different classes in imbalanced datasets. N_t is the number of slices associated with the t -th class.

IV. Experiments

A. Dataset

The experiments were conducted using dMRI data of pediatric subjects from one month to six years of age. The diffusion MR images were acquired using Siemens Prisma MRI scanners using a 32-channel head coil [51] and echo planar imaging sequence with TR/TE = 2640/88.6 ms, voxel size 1.5mm×1.5mm×1.5mm, and $b=0, 500, 1000, 1500, 2000, 2500, 3000$ s/mm² with a total of 144 non-collinear gradient directions. As shown in Table I, the images were separated into three datasets: 1) Training dataset; 2) Testing dataset; and 3) Unlabeled dataset.

A total of 151 volumes were acquired for each subject. 11778 unlabeled volumes and 3624 testing volumes were, respectively, extracted from 78 unlabeled subjects and 24 testing subjects. Since artifacts are more apparent when viewed using sagittal slices, 9000, 706680, and 217440 sagittal slices were extracted respectively from the 150 labeled volumes, 11778 unlabeled volumes, and 3624 testing volumes. Each slice was zero-padded to 144×144, min-max intensity normalized, and labeled according to the volume it belongs to. For the training data, subject-wise annotation was performed. For the testing data, more accurate image quality annotation was performed volume-wise. Fig. 3 shows examples of slices labeled as “pass” (no/minor artifacts), “questionable” (moderate artifacts), and “fail” (heavy artifacts).

B. Implementation Details

The proposed method was implemented using Keras with Tensorflow backend. Evaluation was based on a machine with a CPU (i.e., Intel i7-8700K) and a GPU (i.e., NVIDIA GeForce GTX 1080Ti 11GB). The RMSprop optimizer was utilized for training with initial learning rate set to 1×10^{-5} and decay rate set to 5×10^{-8} . To avoid overfitting, the data were augmented via rotation from 0 to 30 degrees and horizontal flipping. Semi-supervised learning, and slice and volume self-training were repeated twice with the same probability threshold described in Section IV–E.

C. Evaluation Metrics

We evaluated our method by using three performance metrics, including classification accuracy (ACC), sensitivity (SEN), and specificity (SPE), which are defined as

$$ACC = \frac{TP+TN}{TP+TN+FP+FN}, SEN = \frac{TP}{TP+FN}, SPE = \frac{TN}{FP+TN}, \quad (4)$$

where TP, TN, FP and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. In addition, we also evaluated the computational efficiency of our method by using six metrics, i.e., the number of parameters (NoP), the maximal dimension (MD), and GPU and CPU time cost (TC) for slice-wise, volume-wise and subject-wise quality assessment.

D. Compared Methods

To verify the effectiveness of the proposed IQA method, we compared it with several learning based methods:

- Supervised learning with fine tuning [20]: InceptionV3 [52], ResNet [53], VGG [54];
- Supervised learning with hand-crafted features: random forest with features extracted via Gabor filters;
- Unsupervised learning: k -means clustering [55].

For subject-wise IQA, our subject assessment strategy is applied to the volume-wise quality ratings given by these methods.

To verify the effectiveness and superiority of NLDR block and NLR block on constructing SQA-Net, we also compared with 3 ablated versions of SQA-Net:

- Depthwise separable residual (DSR) network : The NLDR block and NLR block are substituted with a DSR block with 128, 256, and 512 channels;
- DSR+NLR network: The NLDR block is substituted with a DSR block with 128 and 256 channels;
- NLDR+NLR network: The DSR block is substituted with a NLDR block with 32 and 64 channels.

Note that SQA-Net is a DSR+NLDR+NLR network. These comparison networks were trained in similar fashion as our proposed SQA-Net.

E. Determination of Thresholds

The thresholds used for semi-supervised learning, and slice/volume relabeling and removal in slice/volume self-training are core factors that affect the overall performance. We determined the thresholds based on the following considerations.

The thresholds in slice/volume self-training (i.e., TH-S and TH-V, respectively) are utilized to relabel or remove slices/volumes with noisy labels, and to retain highly confident slices/volumes for network retraining. Similarly, the threshold in iterative semi-supervised learning is utilized to select confidently annotated pseudo-labeled slices, they share the same threshold parameter (i.e., TH-S). Besides, as these thresholds are used to determine highly confident slices/volumes, we set these threshold values according to the training accuracy. From our experiments, the slice and volume training accuracies are both greater than 0.85. Thus, we consider 3 possible threshold values for TH-S and TH-V, i.e., 0.7, 0.8 and 0.9.

Fig. 4 shows the comparison results of volume- and subject-wise IQA using different combinations of these three threshold values. From the figure, we can see that the performance of volume- and subject-wise IQA are improved whenever a larger TH-S or TH-V value is used, and the best performance is achieved when both thresholds are set to 0.9. Thus, in subsequent experiments, we set both TH-S and TH-V to 0.9. Note that increasing this threshold further (e.g., 0.99) will significantly reduce the amount of confident slices/volumes that can be used for further training to improve performance.

F. Performance Comparison

Fig. 5 shows the evaluation results in terms of sensitivity, specificity, and accuracy for volume- and subject-wise IQA of the testing dMRI data, using different IQA methods, i.e., InceptionV3 ResNet, VGG, k -means, random forest, and our method. From Fig. 5, our method significantly outperforms all compared methods for all evaluation metrics except sensitivity to “pass” volumes (i.e., SEN-V-P). The highest performance for SEN-V-P is achieved by k -means. However, k -means is actually overfitting, as it yields zero sensitivity to “questionable” volumes (i.e., SEN-V-Q) and “fail” volumes (i.e., SEN-V-F). This might be caused by class imbalance in the training data.

G. Ablation Study

1) Computational Efficiency: Computational efficiency was evaluated based on six metrics, i.e., number of parameters (NoP), maximal dimension (MD), GPU and CPU time cost (TC) of slice-, volume- and subject-wise IQA, as summarized in Table II.

For comparison, we summarized in Table II the TCs of manual assessment based on two experienced radiologists: slice, volume and subject TCs, respectively, are ~1s, ~22s, and ~3000s. These are much longer than automatic IQA.

Compared with DSR and DSR+NLR, the NLR block reduces the maximal dimension from 512 to 256 and reduces the NoP by over two times. This reduces the GPU and CPU slice,

volume, and subject TCs by over 3.5%, and both the GPU slice and subject TCs by over 11.4%.

We can see that DSR+NLR with NLDR+NLR have the same MD, whereas the NoP of NLDR+NLR is over 50% larger than that of DSR+NLR due to the nonlocal network block integrated in the NLDR block. This causes the GPU and CPU slice, volume and subject TCs to be increased by at least a factor of 2.

NLDR+NLR and DSR+NLDR+NLR have the same maximal dimension and similar NoPs. The GPU and CPU slice, volume and subject TCs are reduced by over 26%, particularly the CPU slice and volume TCs, which are reduced by over almost four times due to the substitution of the first two NLDR blocks with DSR blocks.

The analysis above shows that the DSR and NLR blocks improve the computational efficiency of the overall network. Although our method is not the most computationally efficient, the GPU and CPU TCs differ only slightly than DSR and DSR+NLR. The proposed method is hence suitable for fast dMRI IQA.

2) Network Efficiency: Tables III and IV show respectively the confusion matrix, sensitivity, specificity, and accuracy of the compared methods for volume- and subject-wise IQA of the testing dMRI data. As shown in Table IV, the proposed method (DSR+NLDR+NLR network) yields the best volume- and subject-wise IQA performance when compared with other methods in terms of sensitivity, specificity, and accuracy.

Compared with DSR and DSR+NLR, the NLR block improves the volume sensitivity for “pass” and “questionable” by over 10.2%, the volume specificity for all quality ratings by over 4.5%, and the accuracy of volume- and subject-wise IQA by over 16.3%. Compared with DSR+NLR and NLDR+NLR, replacing DSRs with NLDRs greatly improves sensitivity, specificity, and accuracy in both volume- and subject-wise IQA, e.g., over 24.0% improvements in sensitivity to “pass” and “fail” volumes, specificity to “questionable” volumes and subjects, and accuracy of volume- and subject-wise IQA. Compared with NLDR+NLR and DSR+NLDR+NLR (i.e., the proposed method), where the DSR block instead of the NLDR block is used to extract low-level features, the performance of volume- and subject-wise IQA is further improved.

It follows from the observations above that DSR, NLDR and NLR blocks enhance the performance of the overall network and make it more suitable for hierarchical IQA.

H. Discordant Case Analysis

Fig. 6 shows the predicted quality rating for each volume of the testing subjects using our volume-wise IQA method. From Table III(d), Table IV(d), and Fig. 6, we can observe that the actual quality ratings disagree with our predictions only for “pass” and “questionable” volumes. That is, our method may occasionally mistaken “pass” and “questionable” volumes.

Inspecting retrospectively all “pass” and “questionable” training volumes, subtle degradation such as contrast reduction and local fuzziness, as shown in Figs. 7 and 8, can confuse the IQA network and cause mispredictions.

I. Effectiveness of Semi-Supervised Learning

We evaluated the effectiveness of semi-supervised learning by comparing SQA-Net with and without semi-supervised learning. The subsequent volume assessment was implemented using VQA-Net without volume self-training.

From Fig. 9, we can observe that semi-supervised learning improves volume- and subject-wise IQA performance. The greatest improvement is the sensitivity to “fail” volumes or subjects. Although there is performance drop for some performance metrics (e.g., sensitivity to “pass” volumes, and specificity to “questionable” and “fail” volumes), the degree of degradation is minimal. Thus, in overall, the pseudo-labeled slices that were selected via semi-supervised learning to retrain SQA-Net are beneficial for improving the performance of volume- and subject-wise IQA.

J. Effectiveness of Slice Self-Training

We evaluated the effectiveness of slice self-training by comparing the performance of SQA-Net with and without slice self-training. In the latter, slice self-training is removed from the training of SQA-Net. The subsequent volume assessment was implemented using VQA-Net without volume self-training.

From Fig. 10, we can observe that slice self-training improves the volume- and subject-wise IQA performance, especially sensitivity to “questionable” and “fail” volumes, specificity to “pass” volumes, and accuracy of volume-wise IQA. This indicates that retaining confident slices and relabeling/removing noisy slices via the self-training strategy improve the performance of subsequent volume- and subject-wise IQA.

K. Effectiveness of Volume Self-Training

We evaluated the effectiveness of the volume self-training by comparing the performance of VQA-Net with and without volume self-training.

From Fig. 11, we can observe that volume self-training improves volume- and subject-wise IQA performance, especially sensitivity to “pass” volumes, specificity to “questionable” volumes and accuracy of volume-wise IQA. This indicates that retaining volumes with reliable labels and relabeling volumes via self-training strategy can improve the performance of subsequent subject-wise IQA. Although there is performance drop for some performance metrics (e.g., sensitivity to “questionable” volumes and specificity to “pass” volumes), the degree of degradation is minimal.

V. Discussion

A. Comparison With Previous Work

In general, no-reference image quality assessment (NR-IQA) methods can be divided into metric-based methods and learning-based methods. Metric-based methods aim to devise a metric to measure severity of an artifact, and thus the quality of an image. The major drawback of metric-based methods is that they are often artifact-specific, i.e., they often use different metrics to measure different artifacts. For example, the image blurriness can be assessed based on image sharpness via edge analysis [56–59], domain transform [60, 61], or pixel-wise statistics [62, 63]. On the other hand, the image noisiness can be estimated by measuring the noise variance via wavelet transform [64], block variance estimation [65–67] and filtering [68–70]. Currently, most metric-based methods focus on measuring blurriness and noisiness of an image, which might not necessarily cover all types of artifacts in diffusion imaging. In contrast, learning-based methods, such as SVM based [17, 71–73] and deep learning based [20, 74–79], aim to replicate the assessments of human raters using labeled (and unlabeled) data. The main difference between SVM-based and deep-learning based methods is that the former relies on hand-crafted features, whereas the latter can automatically extract features that are pertinent to IQA.

For deep learning-based methods, we discuss the differences between our method and previous deep learning based methods in two aspects, i.e., hierarchical organization and learning algorithm.

1) Hierarchical Organization: Compared with the existing learning-based IQA methods that assess 2D [74–76] and 3D anatomical [77, 78], or 4D functional [20, 79] neuroimages, our method adopts 2D slices as inputs to develop a hierarchical IQA model. Specifically, we hierarchically construct multi-scale (i.e., slice-, volume-, and subject-wise) quality assessment networks, where the outputs of the preceding networks are used as inputs to subsequent network. Furthermore, our hierarchical method can be adjusted flexibly according to data availability, i.e., when a large labeled dataset is available, the semi-supervised learning process can be omitted; when a small labeled dataset and a large unlabeled dataset are available, the whole proposed framework can be employed.

2) Learning Algorithm: The learning algorithm of our method differs from that of existing learning-based IQA methods [74–77, 79]. In contrast to existing learning-based methods that require sufficient labeled data for training, our method only require a small number of labeled training data, which are progressively expanded by iteratively including confident slices into the training dataset using semi-supervised learning. Moreover, unlike most previous methods that assume “clean” labels, our method considers multi-scale (i.e., slice/volume) noisy labels during network training, and iteratively retains high-confident data and relabel noisy data via our self-training strategy. In addition, the hierarchical training mechanism of our method also implies that less training data are required to optimize network parameters, compared with end-to-end networks.

B. Limitations and Future Work

While our method achieved near-perfect performance for dMRI IQA, its generalizability can be further improved in the future by addressing the following limitations and challenges.

In our current implementation, the performance refinement strategies, i.e., semi-supervised learning and slice/volume self-training, were carried out individually for time efficiency.

Alternatively, these refinement processes can be trained end-to-end.

Furthermore, currently whole 2D slices are used for training, ignoring the fact that the background region may have little information about motion artifacts. To focus more on the brain, a brain detection model can be incorporated into our framework to determine a brain mask as additional input to the SQA-Net.

VI. Conclusion

In this paper, we have proposed a hierarchical deep learning method for QA of pediatric diffusion MR images. Our framework performs IQA slice-, volume-, and subject-wise. Our method requires only a small amount of quality-annotated images for pretraining the SQA-Net, which is then used to pseudo-label unannotated image slices to expand the training dataset with confidently pseudo-labeled images. Our IQA network is then further refined via subsequent slice and volume self-training. Our method can cope with label noise effectively, affording tolerance to inevitable rater errors, minimizing the requirement of labeled data and reducing the time of manual labeling. Experimental results verify that the proposed method yields near perfect IQA accuracy at a low computational cost.

VII. Acknowledgement

This work was supported in part by NIH grants (NS093842, AG053867, EB006733, MH117943, and MH110274), and the efforts of the UNC/UMN Baby Connectome Project Consortium. We thank Martin A. Styner for providing the quality labels.

References

- [1]. Basser PJ and Pierpaoli C, "Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI," *Journal of Magnetic Resonance, Series B*, vol. 111, no. 3, pp. 209–219, 6 1996. [PubMed: 8661285]
- [2]. Basser PJ and Jones DK, "Diffusion-tensor MRI: theory, experimental design and data analysis - a technical review," *NMR in Biomedicine*, vol. 15, no. 7–8, pp. 456–467, 11 2002. [PubMed: 12489095]
- [3]. Bihan DL, Poupon C, Amadon A, and Lethimonnier F, "Artifacts and pitfalls in diffusion MRI," *Journal of Magnetic Resonance Imaging*, vol. 24, no. 3, pp. 478–488, 8 2006. [PubMed: 16897692]
- [4]. Reuter M, Tisdall MD, Qureshi A, Buckner RL, van der Kouwe AJ, and Fischl B, "Head motion during MRI acquisition reduces gray matter volume and thickness estimates," *NeuroImage*, vol. 107, pp. 107–115, 2 2015. [PubMed: 25498430]
- [5]. Liu S, Thung K-H, Lin W, Yap P-T, and Shen D, "Real-time quality assessment of pediatric mri via semi-supervised deep nonlocal residual neural networks," *IEEE Transactions on Image Processing*, 2020.
- [6]. Sheikh H and Bovik A, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2 2006. [PubMed: 16479813]

- [7]. Chow LS and Paramesran R, "Review of medical image quality assessment," *Biomedical Signal Processing and Control*, vol. 27, pp. 145–154, 5 2016.
- [8]. Gedamu EL, Collins D, and Arnold DL, "Automated quality control of brain MR images," *Journal of Magnetic Resonance Imaging*, vol. 28, no. 2, pp. 308–319, 8 2008. [PubMed: 18666143]
- [9]. Wang Z, Bovik A, Sheikh H, and Simoncelli E, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 4 2004. [PubMed: 15376593]
- [10]. Chandler DM, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, p. 011006, 1 2010.
- [11]. Zhang L, Zhang L, Mou X, and Zhang D, "FSIM: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 8 2011. [PubMed: 21292594]
- [12]. Nafchi HZ, Shahkolaei A, Hedjam R, and Cheriet M, "Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator," *IEEE Access*, vol. 4, pp. 5579–5590, 8 2016.
- [13]. Abdul Rehman ZW, "Reduced-reference image quality assessment by structural similarity estimation," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3378–3389, 5 2012. [PubMed: 22562759]
- [14]. Manap RA and Shao L, "Non-distortion-specific no-reference image quality assessment: A survey," *Information Sciences*, vol. 301, pp. 141–160, 4 2015.
- [15]. Ye P and Doermann D, "No-reference image quality assessment using visual codebooks," *IEEE Transactions on Image Processing*, vol. 21, no. 7, pp. 3129–3138, 7 2012. [PubMed: 22410336]
- [16]. Tang H, Joshi N, and Kapoor A, "Learning a blind measure of perceptual image quality," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, USA, 6 2011, pp. 305–312.
- [17]. Saad MA, Bovik AC, and Charrier C, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, 8 2012. [PubMed: 22453635]
- [18]. Kim J, Zeng H, Ghadiyaram D, Lee S, Zhang L, and Bovik AC, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 130–141, 11 2017.
- [19]. Pizarro RA, Cheng X, Barnett A, Lemaitre H, Verchinski BA, Goldman AL, Xiao E, Luo Q, Berman KF, Callicott JH, Weinberger DR, and Mattay VS, "Automated quality assessment of structural magnetic resonance brain images based on a supervised machine learning algorithm," *Frontiers in Neuroinformatics*, vol. 10, 12 2016.
- [20]. Graham MS, Drobnyak I, and Zhang H, "A supervised learning approach for diffusion MRI quality control with minimal training data," *NeuroImage*, vol. 178, pp. 668–676, 9 2018. [PubMed: 29883734]
- [21]. Liu S, Thung K-H, Lin W, Yap P-T, and Shen D, "Multi-stage image quality assessment of diffusion mri via semi-supervised nonlocal residual networks," in *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Shenzhen, China, 10 2019.
- [22]. Pan SJ and Yang Q, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 10 2010.
- [23]. Tan C, Sun F, Kong T, Zhang W, Yang C, and Liu C, "A survey on deep transfer learning," in *ICANN 2018: Artificial Neural Networks and Machine Learning*, Rhodes, Greece, 9 2018, pp. 270–279.
- [24]. Zhu X, "Semi-supervised learning literature survey," *Computer Sciences*, University of Wisconsin-Madison, Tech. Rep 1530, 2005.
- [25]. Nigam K, McCallum AK, Thrun S, and Mitchell T, "Text classification from labeled and unlabeled documents using em," *Machine Learning*, vol. 39, no. 2/3, pp. 103–134, 5 2000.
- [26]. Avrim Blum SC, "Learning from labeled and unlabeled data using graph mincuts," in *Proceedings of International Conference on Machine Learning (ICML)*, 6 2001, pp. 19–26.
- [27]. Blum A, Lafferty J, Rwebangira MR, and Reddy R, "Semi-supervised learning using randomized mincuts," in *Proceedings of International Conference on Machine Learning (ICML)*. ACM Press, 6 2004.

- [28]. Lee D-H, "Pseudo-label: The simple and efficient semisupervised learning method for deep neural networks," in Proceedings of International Conference on Machine Learning, Atlanta, USA, 6 2013.
- [29]. Wu H and Prasad S, "Semi-supervised deep learning using pseudo labels for hyperspectral image classification," IEEE Transactions on Image Processing, vol. 27, no. 3, pp. 1259–1270, 3 2018. [PubMed: 29990156]
- [30]. Frenay B and Verleysen M, "Classification in the presence of label noise: A survey," IEEE Transactions on Neural Networks and Learning Systems, vol. 25, no. 5, pp. 845–869, 5 2014. [PubMed: 24808033]
- [31]. Reed S, Lee H, Anguelov D, Szegedy C, Erhan D, and Rabinovich A, "Training deep neural networks on noisy labels with bootstrapping," CoRR, vol. abs/1412.6596, 2014.
- [32]. Li Y, Yang J, Song Y, Cao L, Luo J, and Li L-J, "Learning from noisy labels with distillation," in Proceedings of IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 10 2017, pp. 1910–1918.
- [33]. Patrini G, Rozza A, Menon AK, Nock R, and Qu L, "Making deep neural networks robust to label noise: A loss correction approach," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, 7 2017, pp. 1944–1952.
- [34]. Brooks JP, "Support vector machines with the ramp loss and the hard margin loss," Operations Research, vol. 59, no. 2, pp. 467–479, 4 2011.
- [35]. van Rooyen B, Menon A, and Williamson RC, "Learning with symmetric label noise: The importance of being unhinged," in Proceedings of Advances in Neural Information Processing Systems, 2015, pp. 10–18.
- [36]. Bartlett PL, Jordan MI, and McAuliffe JD, "Convexity, classification, and risk bounds," Journal of the American Statistical Association, vol. 101, no. 473, pp. 138–156, 3 2006.
- [37]. Teng C-M, "A comparison of noise handling techniques," in Proceedings of International Florida Artificial Intelligence Research Society Conference, Key West, USA, 5 2001, pp. 269–273.
- [38]. Mnih V and Hinton GE, "Learning to label aerial images from noisy data," in Proceedings of International Conference on Machine Learning (ICML), Edinburgh, Scotland, 6 2012, pp. 203–210.
- [39]. Sukhbaatar S, Bruna J, Paluri M, Bourdev L, and Fergus R, "Training convolutional networks with noisy labels," CoRR, vol. abs/1406.2080, 2014.
- [40]. Xiao T, Xia T, Yang Y, Huang C, and Wang X, "Learning from massive noisy labeled data for image classification," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA, 6 2015, pp. 2691–2699.
- [41]. Jindal I, Nokleby M, and Chen X, "Learning deep networks from noisy labels with dropout regularization," in Proceedings of IEEE International Conference on Data Mining (ICDM), Barcelona, Spain, 12 2016, pp. 967–972.
- [42]. Misra I, Zitnick CL, Mitchell M, and Girshick R, "Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 6 2016, pp. 2930–2939.
- [43]. Yao J, Wang J, Tsang IW, Zhang Y, Sun J, Zhang C, and Zhang R, "Deep learning from noisy image labels with quality embedding," IEEE Transactions on Image Processing, vol. 28, no. 4, pp. 1909–1922, 4 2019.
- [44]. Cuendet S, Hakkani-Tür D, and Shriberg E, "Automatic labeling inconsistencies detection and correction for sentence unit segmentation in conversational speech," in Proceedings of MLMI on Machine Learning for Multimodal Interaction, Brno, Czech Republic, 6 2007, pp. 144–155.
- [45]. Miranda ALB, Garcia LPF, Carvalho ACPLF, and Lorena AC, "Use of classification algorithms in noise detection and elimination," in Proceedings of International Conference on Hybrid Artificial Intelligence Systems, Salamanca, Spain, 2009, pp. 417–424.
- [46]. Seiffert C, Khoshgoftaar TM, Hulse JV, and Folleco A, "An empirical study of the classification performance of learners on imbalanced and noisy software quality data," Information Sciences, vol. 259, pp. 571–595, 2 2014.

- [47]. Wang X, Girshick R, Gupta A, and He K, "Non-local neural networks," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, USA, 6 2018, pp. 7794–7803.
- [48]. Chollet F, "Xception: Deep learning with depthwise separable convolutions," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, 7 2017, pp. 1800–1807.
- [49]. O'Donnell LJ, Daducci A, Wassermann D, and Lenglet C, "Advances in computational and statistical diffusion MRI," *NMR in Biomedicine*, vol. 32, no. 4, p. e3805, 11 2017. [PubMed: 29134716]
- [50]. Lin T-Y, Goyal P, Girshick R, He K, and Dollar P, "Focal loss for dense object detection," in Proceedings of IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 10 2017, pp. 2980–2988.
- [51]. Howell BR, Styner MA, Gao W, Yap P-T, Wang L, Baluyot K, Yacoub E, Chen G, Potts T, Salzwedel A, Li G, Gilmore JH, Piven J, Smith JK, Shen D, Ugurbil K, Zhu H, Lin W, and Elison JT, "The UNC/UMN baby connectome project (BCP): An overview of the study design and protocol development," *NeuroImage*, vol. 185, pp. 891–905, 1 2019. [PubMed: 29578031]
- [52]. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, and Wojna Z, "Rethinking the inception architecture for computer vision," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6 2016.
- [53]. He K, Zhang X, Ren S, and Sun J, "Deep residual learning for image recognition," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6 2016.
- [54]. Simonyan K and Zisserman A, "Very deep convolutional networks for large-scale image recognition," in International Conference on Learning Representations, 2015.
- [55]. Sculley D, "Web-scale k-means clustering," in Proceedings of the 19th International Conference on World Wide Web (WWW), 4 2010, pp. 1177–1178.
- [56]. Chung Y-C, Wang J-M, Bailey R, Chen S-W, and Chang S-L, "A non-parametric blur measure flased on edge analysis for image processing applications," in Proceedings of IEEE Conference on Cybernetics and Intelligent Systems (ICCIS), Singapore, Singapore, 7 2005.
- [57]. Wu S, Lin W, Jiang L, Xiong W, Chen L, and Ong SH, "An objective out-of-focus blur measurement," in Proceedings of International Conference on Information Communications & Signal Processing, Bangkok, Thailand, 9 2006.
- [58]. Liang L, Chen J, Ma S, Zhao D, and Gao W, "A no-reference perceptual blur metric using histogram of gradient profile sharpness," in Proceedings of IEEE International Conference on Image Processing (ICIP), Cairo, Egypt, 11 2010.
- [59]. Yan Q, Xu Y, and Yang X, "No-reference image blur assessment based on gradient profile sharpness," in Proceedings of IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), London, UK, 10 2013.
- [60]. Ciancio A, da Costa A, da Silva E, Said A, Samadani R, and Obrador P, "Objective no-reference image blur metric based on local phase coherence," *Electronics Letters*, vol. 45, no. 23, p. 1162, 11 2009.
- [61]. Hassen R, Wang Z, and Salama MMA, "Image sharpness assessment based on local phase coherence," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2798–2810, 7 2013. [PubMed: 23481852]
- [62]. Crete F, Dolmiere T, Ladret P, and Nicolas M, "The blur effect: perception and estimation with a new no-reference perceptual blur metric," in Proceedings of Human Vision and Electronic Imaging, 2 2007.
- [63]. Liu D, Chen Z, Ma H, Xu F, and Gu X, "No reference block based blur detection," in Proceedings of International Workshop on Quality of Multimedia Experience, San Diego, CA, USA, 7 2009.
- [64]. Donoho DL and Johnstone IM, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 9 1994.
- [65]. Meer P, Jolion J-M, and Rosenfeld A, "A fast parallel algorithm for blind estimation of noise variance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 2, pp. 216–223, 2 1990.

- [66]. Tian J and Chen L, "Image noise estimation using a variation-adaptive evolutionary approach," *IEEE Signal Processing Letters*, vol. 19, no. 7, pp. 395–398, 7 2012.
- [67]. Huang X, Chen L, Tian J, Zhang X, and Fu X, "Blind noisy image quality assessment using block homogeneity," *Computers & Electrical Engineering*, vol. 40, no. 3, pp. 796–807, 4 2014.
- [68]. Rank K, Lendl M, and Unbehauen R, "Estimation of image noise variance," *IEE Proceedings - Vision, Image, and Signal Processing*, vol. 146, no. 2, p. 80, 8 1999.
- [69]. Shin D-H, Park R-H, Yang S, and Jung J-H, "Block-based noise estimation using adaptive gaussian filtering," in *Proceedings of International Conference on Consumer Electronics (ICCE)*, Las Vegas, NV, USA, 5 2005.
- [70]. Tai S-C and Yang S-M, "A fast method for image noise estimation using laplacian operator and adaptive edge detection," in *Proceedings of International Symposium on Communications, Control and Signal Processing (ISCCSP)*, St Julians, Malta, 3 2008.
- [71]. Moorthy A and Bovik A, "A two-step framework for constructing blind image quality indices," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 513–516, 5 2010.
- [72]. Mittal A, Moorthy AK, and Bovik AC, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 12 2012. [PubMed: 22910118]
- [73]. Zhang Y and Chandler DM, "No-reference image quality assessment based on log-derivative statistics of natural scenes," *Journal of Electronic Imaging*, vol. 22, no. 4, p. 043025, 12 2013.
- [74]. Ding Y, Suffren S, Bellec P, and Lodygensky GA, "Supervised machine learning quality control for magnetic resonance artifacts in neonatal data sets," *Human Brain Mapping*, vol. 40, no. 4, pp. 1290–1297, 11 2018. [PubMed: 30467922]
- [75]. Esses SJ, Lu X, Zhao T, Shanbhogue K, Dane B, Bruno M, and Chandarana H, "Automated image quality evaluation of T2 -weighted liver MRI utilizing deep learning architecture," *Journal of Magnetic Resonance Imaging*, vol. 47, no. 3, pp. 723–728, 6 2017. [PubMed: 28577329]
- [76]. Küstner T, Gatidis S, Liebgott A, Schwartz M, Mauch L, Martirosian P, Schmidt H, Schwenzer N, Nikolaou K, Bamberg F, Yang B, and Schick F, "A machine-learning framework for automatic reference-free quality assessment in MRI," *Magnetic Resonance Imaging*, vol. 53, pp. 134–147, 11 2018. [PubMed: 30036653]
- [77]. Küstner T, Liebgott A, Mauch L, Martirosian P, Bamberg F, Nikolaou K, Yang B, Schick F, and Gatidis S, "Automated reference-free detection of motion artifacts in magnetic resonance images," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 31, no. 2, pp. 243–256, 9 2017.
- [78]. Esteban O, Birman D, Schaer M, Koyejo OO, Poldrack RA, and Gorgolewski KJ, "MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites," *PLOS ONE*, vol. 12, no. 9, p. e0184661, 9 2017. [PubMed: 28945803]
- [79]. Kelly C, "Transfer learning and convolutional neural net fusion for motion artefact detection," in *Proceedings of the Annual Meeting of the International Society for Magnetic Resonance in Medicine (ISMRM)*, Honolulu, Hawaii, 4 2017, p. 3523.

Fig. 1.

Overview of our IQA method, which includes slice IQA, Volume IQA, and subject IQA. n_s and n_v denote the number of slices in one volume and the number of volumes in one subject, respectively.

IEEE Trans Med Imaging. Author manuscript; available in PMC 2021 November 01.

Fig. 3.

Example of image slices labeled as “pass” (no/minor artifacts), “questionable” (moderate artifacts), and “fail” (heavy artifacts).

Fig. 4.

The accuracy of VQA and SQA with three different threshold values in slice and volume self-training. TH-S: threshold used in slice self-training, TH-V: threshold used in volume self-training.

Fig. 5.

Comparison of volume- and subject-wise IQA. Evaluation metrics are denoted in the following form: metric (SEN/SPE/ACC) - stage (V/S) - quality (P/Q/F/-).

Fig. 6.

Volume- and subject-wise QA using our method. All volumes are sorted in the sequence of “fail”, “questionable” and “pass” from top to bottom.

Fig. 7.
Slices belonging to a discordant “pass” volume, which is mis-predicted as “questionable”. (b
= 2000 s/mm²)

Fig. 8.

Slices belonging to a discordant “questionable” volume, which is mis-predicted as “pass”. ($b = 1500 \text{ s/mm}^2$)

(a) Volume-wise IQA with or without semi-supervised learning

(b) Subject-wise IQA with or without semi-supervised learning

Fig. 9.

Volume- and subject-wise IQA using our method trained with and without semi-supervised learning. Self-training is removed for both cases in this experiment. Evaluation metrics are denoted in the following form: metric (SEN/SPE/ACC)-stage (V/S)-quality (P/Q/F/-).

(a) Volume-wise IQA with or without slice self-learning

(b) Subject-wise IQA with or without slice self-learning

Fig. 10.

Volume- and subject-wise IQA using our method trained with and without slice self-training.

Evaluation metrics are denoted in the following form: metric (SEN/SPE/ACC)-stage (V/S)-quality (P/Q/F/-).

(a) Volume-wise IQA with or without volume self-learning

(b) Subject-wise IQA with or without volume self-learning

Fig. 11.

Volume- and subject-wise IQA using our model that was trained with and without volume self-training. SQA-Net was pre-trained via semi-supervised learning and subsequent VQA-Net did not use volume self-training. Evaluation metrics are denoted in the following form: metric (SEN/SPE/ACC)-stage (V/S)-quality (P/Q/F/-).

TABLE I

Datasets used for evaluation

Training Dataset (Unit: Volume)			Unlabeled Dataset (Unit: Volume)	Testing Dataset (Unit: Volume)		
Pass	Ques	Fail		Pass	Ques	Fail
85	25	40	11778	2148	1170	306

TABLE II

Computation efficiency of different methods

Method	NoP		MD	Slice TC		Volume TC		Subject TC	
	SQA-Net	VQA-Net		GPU	CPU	GPU	CPU	GPU	CPU
DSR	0.74M	2.11M	512	0.149s	0.152s	7.652s	8.611s	16.773s	258.679s
DSR+NLR	0.32M	2.11M	256	0.132s	0.151s	7.141s	8.310s	14.421s	237.999s
NLDR+NLR	0.50M	2.11M	256	0.275s	0.834s	15.903s	45.692s	35.312s	571.352s
DSR+NLDR+NLR (Proposed)	0.49M	2.11M	256	0.202s	0.210s	9.321s	11.635s	18.786s	263.062s
Manual Assessment	-	-	-	~1s		~22s		~3000s	

TABLE III

Confusion matrices

(a) DSR							
Image Quality		Predicted					
		Pass		Ques		Fail	
		Volume	Subject	Volume	Subject	Volume	Subject
Actual	Pass	1254	4	892	6	2	0
	Ques	132	1	1000	9	38	0
	Fail	10	0	62	1	234	3
(b) DSR+NLR							
Image Quality		Predicted					
		Pass		Ques		Fail	
		Volume	Subject	Volume	Subject	Volume	Subject
Actual	Pass	1569	7	579	3	0	0
	Ques	63	1	1102	9	5	0
	Fail	18	0	66	1	222	3
(c) NLDR+NLR							
Image Quality		Predicted					
		Pass		Ques		Fail	
		Volume	Subject	Volume	Subject	Volume	Subject
Actual	Pass	1952	9	196	1	0	0
	Ques	33	0	1135	10	2	0
	Fail	0	0	15	0	291	4
(d) DSR+NLDR+NLR							
Image Quality		Predicted					
		Pass		Ques		Fail	
		Volume	Subject	Volume	Subject	Volume	Subject
Actual	Pass	2134	10	14	0	0	0
	Ques	23	0	1147	10	0	0
	Fail	0	0	0	0	306	4

TABLE IV

Sensitivity, specificity, and accuracy

(a) DSK						
Image Quality	SEN		SPE		ACC	
	Volume	Subject	Volume	Subject	Volume	Subject
Pass	0.5838	0.4000	0.9038	0.9286	0.6865	0.6667
Ques	0.8547	0.7000	0.6112	0.5000		
Fail	0.7647	0.7500	0.9879	1.0000		
(b) DSR+NLR						
Image Quality	SEN		SPE		ACC	
	Volume	Subject	Volume	Subject	Volume	Subject
Pass	0.7304	0.7000	0.9451	0.9286	0.7983	0.7917
Ques	0.9419	0.9000	0.7372	0.7143		
Fail	0.7255	0.7500	0.9985	1.0000		
(c) NLDR+NLR						
Image Quality	SEN		SPE		ACC	
	Volume	Subject	Volume	Subject	Volume	Subject
Pass	0.9088	0.9000	0.9776	1.0000	0.9321	0.9583
Ques	0.9701	1.0000	0.9140	0.9286		
Fail	0.9510	1.0000	0.9994	1.0000		
(d) DSR+NLDR+NLR						
Image Quality	SEN		SPE		ACC	
	Volume	Subject	Volume	Subject	Volume	Subject
Pass	0.9935	1.0000	0.9844	1.0000	0.9898	1.0000
Ques	0.9803	1.0000	0.9943	1.0000		
Fail	1.0000	1.0000	1.0000	1.0000		