Incorporating Radiologist Knowledge Into MRI Quality Metrics for Machine Learning Using Rank-Based Ratings

Chenwei Tang, MS,[1] Laura B. Eisenmenger, MD,[2] Leonardo Rivera-Rivera, PhD,[1,3] Eugene Huo, MD,[4] Jacqueline C. Junn, MD,[5] Anthony D. Kuner, MD,[2] Thekla H. Oechtering, MD,[2,6] Anthony Peret, MD,[2] Jitka Starekova, MD,[2] and Kevin M. Johnson, PhD[1,2*]

Background: Deep learning (DL) often requires an image quality metric; however, widely used metrics are not designed for medical images.
Purpose: To develop an image quality metric that is specific to MRI using radiologists image rankings and DL models.
Study Type: Retrospective.
Population: A total of 19,344 rankings on 2916 unique image pairs from the NYU fastMRI Initiative neuro database was used for the neural network-based image quality metrics training with an 80%/20% training/validation split and fivefold cross-validation.
Field Strength/Sequence: 1.5 T and 3 T T1, T1 postcontrast, T2, and FLuid Attenuated Inversion Recovery (FLAIR).
Assessment: Synthetically corrupted image pairs were ranked by radiologists ($N = 7$), with a subset also scoring images using a Likert scale ($N = 2$). DL models were trained to match rankings using two architectures (EfficientNet and IQ-Net) with and without reference image subtraction and compared to ranking based on mean squared error (MSE) and structural similarity (SSIM). Image quality assessing DL models were evaluated as alternatives to MSE and SSIM as optimization targets for DL denoising and reconstruction.
Statistical Tests: Radiologists' agreement was assessed by a percentage metric and quadratic weighted Cohen's kappa. Ranking accuracies were compared using repeated measurements analysis of variance. Reconstruction models trained with IQ-Net score, MSE and SSIM were compared by paired *t* test. $P < 0.05$ was considered significant.
Results: Compared to direct Likert scoring, ranking produced a higher level of agreement between radiologists (70.4% vs. 25%). Image ranking was subjective with a high level of intraobserver agreement ($94.9\% \pm 2.4\%$) and lower interobserver agreement ($61.47\% \pm 5.51\%$). IQ-Net and EfficientNet accurately predicted rankings with a reference image ($75.2\% \pm 1.3\%$ and $79.2\% \pm 1.7\%$). However, EfficientNet resulted in images with artifacts and high MSE when used in denoising tasks while IQ-Net optimized networks performed well for both denoising and reconstruction tasks.
Data Conclusion: Image quality networks can be trained from image ranking and used to optimize DL tasks.
Level of Evidence: 3
Technical Efficacy: Stage 1

J. MAGN. RESON. IMAGING 2025;61:2572–2584.

View this article online at wileyonlinelibrary.com. DOI: 10.1002/jmri.29672

Received Jul 14, 2024, Accepted for publication Nov 19, 2024.

*Address reprint requests to: K.M.J., 1133 Wisconsin Institutes for Medical Research (WIMR), 1111 Highland Ave, Madison, WI 53705, USA.
E-mail: kmjohnson3@wisc.edu

From the [1]Department of Medical Physics, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, USA; [2]Department of Radiology, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, USA; [3]Department of Medicine, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, USA; [4]Department of Radiology, University of California, San Francisco, California, USA; [5]Department of Radiology, Icahn School of Medicine at Mount Sinai, New York, New York, USA; and [6]Department of Radiology and Nuclear Medicine, Universität zu Lübeck, Lübeck, Germany

Additional supporting information may be found in the online version of this article

Many studies have shown the application of deep learning (DL) to improvement of medical images.[1–9] Such DL applications commonly use supervised learning, an approach where a network is trained to match reference images based on an image quality metric.[7,8] Despite its critical role, the image quality metric is often heuristically selected from metrics developed for natural images.[7,8] A common metric is the pixel-wise mean squared error (MSE); however, this can be severely affected by local variations and often leads to blurry images and feature loss.[10–12] To address these limitations, structural similarity (SSIM)[13,14] and perceptual loss[10] have been applied to DL reconstruction.[15] However, these methods have been designed for images outside of the medical imaging domain, where the artifacts and texture are substantially different. There have also been efforts to provide image quality metrics without reference images based on distortion (perception-based image quality evaluator[16]) and sharpness (Tenengrad metrics[17]), but these may not be suitable for assessing medical images. Recently, adversarial loss[18,19] and unsupervised feature loss[11] have been proposed as MRI specific quality metrics, yet these metrics are challenging to train and do not consider radiologists and their interpretation of the images and thus remain disconnected from the clinical use of the images.

Incorporating the knowledge of radiologists and other human observers into image quality metrics has potential to improve DL applications. In previous studies in natural images and PET imaging,[20,21] image quality assessing neural networks (NNs) have been derived using training based on Likert scoring, providing a method to incorporate human observers into automated measures of image quality. In these studies, a large set of images were scored by observers and an image quality assessing NN trained to predict the observer scores. The image quality assessing NN was then used autonomously to perform quality assurance or to be incorporated into other DL tasks (eg, image reconstruction). Likert scoring has limitations which leads to challenges for its use in DL applications.[22–24] Likert scoring inherently uses discrete binning and thus will often be unable to capture subtle differences between images. It is also sensitive to baseline shifts between observations. This includes the shift in scoring between readers and the shifts for individual readers over time. This creates an effective noise and bias in the scoring. Given these limitations, image quality networks trained on Likert scoring may be best suited for image quality assurance tasks but may not be suited for directly training image improving NNs.

Previous studies in nonmedical applications[25,26] have developed image quality metrics based on image ranking, rather than Likert scoring. Prospective ranking, rather than scoring, captures subtle preferences as two images are directly compared. Ranking is also insensitive to baseline shifts and thus has the potential to enable collecting rankings from many radiologists.

PLAIN LANGUAGE SUMMARY

Deep learning (DL) can enhance MR images using past data, neural networks, and training to optimize based on an image quality metric. Despite its essential role, MR specific image quality metrics that can be easily used in such tasks are yet to be developed. This study used radiologist perception rankings to train an MRI-specific image quality neural network (NN). The learned metric more accurately represented radiologist preferences compared to mean squared error (MSE) and structural similarity (SSIM). The network was also used to train a DL denoiser and a model-based DL reconstruction, showing advantages over MSE and SSIM.

Thus, the aims of this study were 1) to develop a robust method to directly incorporate radiologist expertise into image quality metrics based on image ranking, 2) to incorporate it into an image quality assessing NN, and 3) to demonstrate the utility of this network as a loss function in image denoising and reconstruction tasks. A further aim was to compare the image ranking-based image quality metric with a Likert scoring approach.

Materials and Methods

Data used for the ranking, denoising and reconstruction tasks was from the NYU fastMRI Initiative neuro database (fastmri.med.nyu.edu). Curation of these data is part of an institutional review board-approved study, and informed consent was obtained prior to data acquisition. Data were deidentified by NYU researchers.[27,28]

Figure 1 provides an overview of the perception studies and their use in training an image quality assessing convolutional neural network (CNN). First, synthetically corrupted pairs of images were generated. These images were created from 2916 slices of raw k-space data which were fully sampled and thus had a ground truth reference image. The corrupted images had arbitrary types and levels of corruption designed to mimic the expected range of degradations seen in clinical settings (detailed descriptions in the Images for Ranking section). Pairs of corrupted images, both generated from the same raw data, were then shown to radiologists simultaneously in a web interface. Within the web interface, radiologists were asked to indicate if one of the corrupted images was of higher diagnostic quality than the other, or if both had similar quality. Once rankings were collected from all radiologists across the range of images, the corrupted image pairs and radiologist rankings of those images were used to train an image quality scoring CNN. The image quality scoring CNN was designed to take an input and produce a single, continuous image quality score. Once trained based on rankings, the produced score will ideally be lower for higher quality, radiologist preferred images and higher for lower quality images. To facilitate this training and accommodate the radiologists indicating images to be similar, a small classifier is simultaneously trained to discriminate the three possible outcomes (i.e., image 1 is preferred, image 2 is

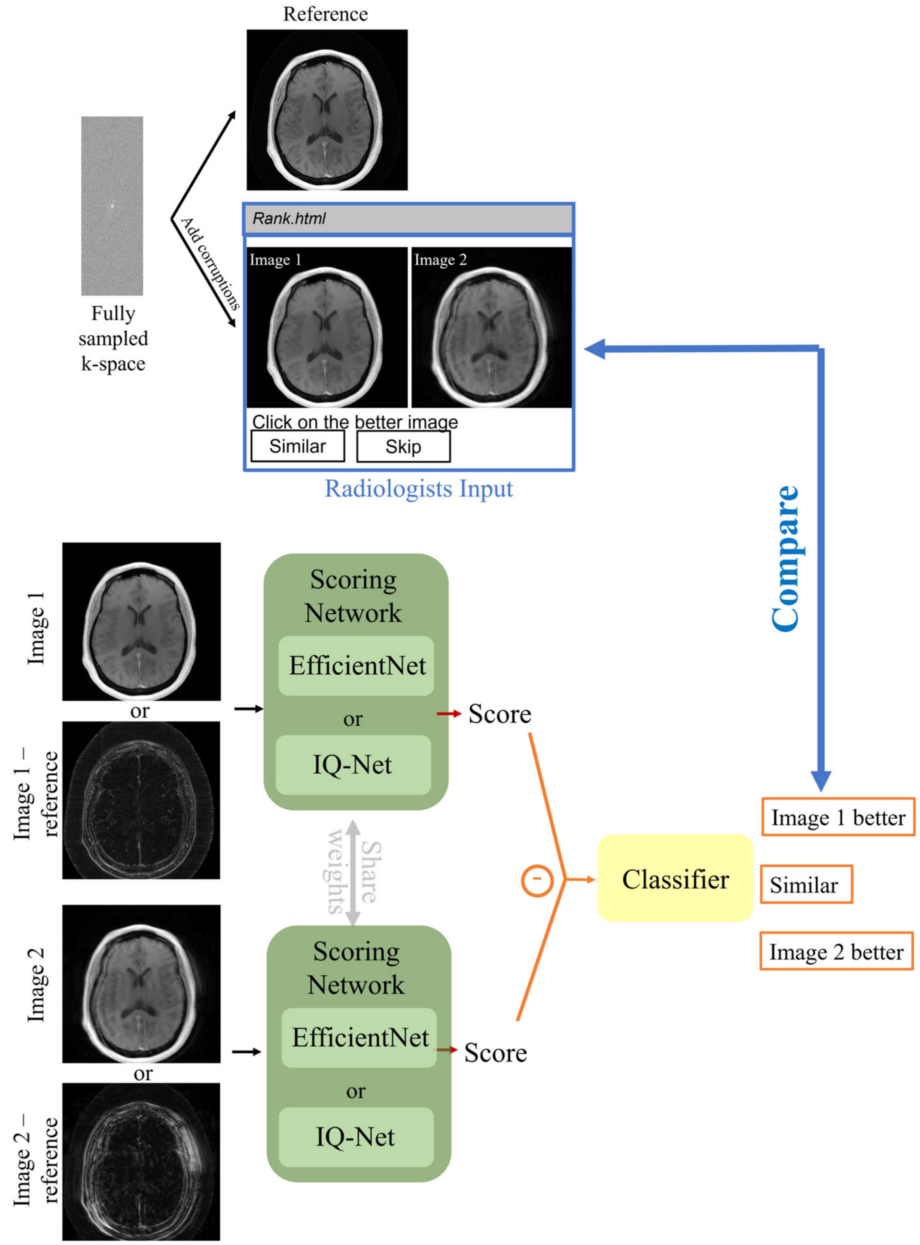

Figure 1: A convolutional neural network (IQ-Net or EfficientNet) and a classifier were trained based on radiologist rankings of images. Radiologists were shown two images and asked to rank the images or indicate them to be of similar quality. In training, each image from the pair was passed through the same network, producing an image quality score. These scores were then passed to a small classifier to produce a rank, which was compared to the radiologists' rankings using cross-entropy loss. Details on the architecture of IQ-Net and the Classifier can be found in Supplemental Materials.

preferred, the images are similar) from the two scores produced from the images by the CNN. Training of the CNN and classifier is performed using a standard cross entropy loss function which compares the radiologists' rankings to those produced by the combination of the scoring CNN and the classifier. Training based on pair-wise comparisons enabled the NN to implicitly learn the correlation between a human preference score and different image textures and artifacts. As detailed below, we tested the effects of using different input types (corrupted images and corrupted images with the reference images subtracted) and architectures of the scoring CNNs. The trained scoring network could then be utilized to train other DL networks requiring a continuous quality metric. All DL networks were implemented in PyTorch[29] with code available: https://github.com/uwmri/LearnedImagingMetrics.

Images for Ranking

In the ranking part of this study we included axial T1, T1 post-contrast, T2, and T2FLAIR 2D images ($N = 378$, 428, 1894, and 216, respectively) from the fastMRI Initiative neuro database.[27,28] Raw k-space data were randomly corrupted and reconstructed with different methods to create image pairs with heterogenous artifacts. A total of 2916 image pairs were generated. Added corruption included:

1. Gaussian low pass filter with a random $\sigma \in (0,15)$ applied to k-space for blurring.
2. Undersampling performed in 1D by randomly removing 0%–30% of the phase encoding (PE) lines in the outer region of k-space. The central region (20%–40%) was unaffected.
3. Undersampling performed in 2D by randomly removing 0%–20% of the data points in the outer region of k-space. The central square region (20%–40%) was unaffected.
4. Additional Gaussian complex noise added to k-space data with random levels $n\sigma, (n \in (0,12))$. σ is the noise level of the truth image estimated by taking the median of the edge of the fully sampled k-space.
5. Translational bulk motion in one or two directions by modulating k-space. We simulated the motion as a single abrupt event during acquisition. The amount of motion is random with the maximum set to be 20 pixels. The start of motion-corrupted phase encode lines were random numbers following Gaussian distribution. No motion will be added if the start is within the central 10% of the k-space.

Each image contained a random number (0–5) of these types of corruption and was reconstructed by one of the following methods chosen at random: Fast Fourier Transform (FFT) followed by sum-of-squares, FFT followed by coil combined with sensitivity maps, sensitivity encoding (SENSE),[30] and compressed sensing.[31] All reconstructions were implemented in SigPy.[32] Different reconstruction methods were included in the study as they tend to produce unique image textures that may affect radiologists' perception. To ensure some degree of similarity in each image pair, a hard constraint was placed on the SSIM between the reference and the two corrupted images ($|\,\mathrm{SSIM}_{\text{image 1}-\text{reference}} - \mathrm{SSIM}_{\text{image 2}-\text{reference}}\,| < 0.09$). If the image pair did not satisfy this constraint, images were regenerated.

Radiologist Ranking and Agreement Evaluation

A total of 2916 corrupted image pairs were shown to radiologists with random repeats and random left/right ordering on a HTML page. They were instructed to do one of the following: 1) select the image with perceived higher diagnostic quality, 2) select the diagnostic image quality to be similar, or 3) skip the image pair if it did not include relevant anatomy. The time for ranking each image pair was recorded. No specific details were provided to the radiologists to specify imaging features of interest to reduce bias. Radiologists were recruited from the imaging service of the three participating institutions and visiting research radiologists ($N = 7$). Reader 1–7 (LBE, ADK, JCJ, EH, AP, THO, JS) has 3, 8, 2, 11, 0, 5, and 5 years of experience as board-certified radiologists, respectively, and reader 5 has 2 years of experience as a radiology resident at the time of ranking.

To determine inter and intra observer agreement from repeated rankings, the overlap in the selection probability between two readers was summed over the number of common rankings. This methodology is detailed in Supplemental Material A. The overall agreement among all reviewers was evaluated as the ratio of majority votes versus all reported results. When one reader reported different rankings for the same image pair, the ranking with the most occurrences was regarded as the final vote of that reader.

Comparison to Likert Scores

In a small subset of cases, Likert scores were also collected. One hundred and four corrupted images (52 image pairs from the total pool of 2916 image pairs) were shown individually in a random order to two radiologists (Reader 5 and 7). The readers were prompted to give each image a score based on a 5-point scale Likert scale for diagnostic quality. Criteria for scoring is detailed in Table S4 in the Supplemental Material. The time for scoring each image was recorded. The confusion matrices between the two readers for Likert scoring, rankings based on Likert scoring and direct rankings were determined. The quadratic weighted Cohen's kappa κ_w for Likert scoring and percentage agreement was calculated.

Scoring Network and Classifier

For the scoring network, two different architectures, EfficientNet[33] and a simplified homemade CNN (IQ-Net) were investigated. Training was supervised by radiologists' direct rankings. As the scoring networks were to be used as a reconstruction loss function, the aim was to have minimal size EfficientNet-b0 (43,306 trainable parameters) was initially adopted. IQ-Net, a simple downsampling encoder with 3 × 3 convolution kernels (Fig. S1a in the Supplemental Material) was also developed and investigated. To maintain independence of the input image scale and other statistics, normalization layers were not included in IQ-Net. (5562 trainable parameters). Another noticeable difference between IQ-Net and EfficientNet is that IQ-Net combines output from each layer, pooling information from large and detailed feature maps.

In the classifier network (Fig. 1 and Fig. S1b in the Supplemental Material), the input is the difference of the scores and the output is a three-value vector containing the likelihood of the radiologist selecting either image or indicating that the images are similar. Thus, the goal of the network was to estimate a simple function describing how the difference in scores relates to radiologists identifying the images as similar. To achieve this, a symmetric design was adopted consisting of two modules, f and g, both consisting of a fully connected layer, a sigmoid activation, and another fully connected layer. The output of each module activated by Softmax was treated as the probability of a certain result. The symmetric design ensured that if image pairs are swapped, the classifier should also swap to maintain the same preference prediction. The classifier added 196 parameters to training.

The ranking network was trained on image pairs of dimensions 396×396, batch size $= 48$, and learning rate $= 10^{-4}$ for the scoring network and 10^{-3} for the classifier with an Adam optimizer (weight decay rate $= 0.999$). Dropout with a rate of 0.5 was used to reduce overfitting for training. Data augmentation was also performed including random rotations, translations, and phase modulation by first-order polynomials. The network was trained using an 80%/20% training/validation split of the consensus ranking from all radiologists. A fivefold cross-validation was performed for both networks and for both input types and the average accuracies were reported. Both IQ-Net and EfficientNet were trained with and without the reference images subtracted to explore their capability as full-reference and no-reference metrics.

Table 1. Estimated Agreement in the Ranking of Image Pairs Within and Between Radiologist Reviewers Based on Repeated Rankings

Readers							
Agreement (%)	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	88.938	63.597	69.293	71.231	74.274	69.577	68.336
2		95.723	53.92	63.324	62.436	69.756	57.619
3			95.678	70.596	73.035	61.327	70.548
4				96.519	77.797	69.097	65.986
5					95.644	69.312	71.23
6						96.011	64.534
7							95.629
Overall	81.37						

Evaluation and Interpretation of EfficientNet and IQ-Net

As the aim was to design a MR specific metric that can be used as a loss function, an image denoising task and an image reconstruction task were performed. A separate 24,002 slices (18,518/5484 slices for training and validation) not seen during the training of the ranking networks were used. For the denoising task training, during each epoch, the denoiser network went through 512/50 random slices in the training/validation set. For the reconstruction task, each epoch went through 2048 random training slices and 200 validation slices. All inference were performed on test dataset not seen by either the ranking or the denoising/reconstruction training.

For the denoising task, the loss function was $|\text{Score}(I) - \text{Score}(I_{\text{reference}})|^2$, where I is the denoised image. Input noisy images were generated by adding complex Gaussian noise to fully sampled k-space data. These images were fed to a ResUNet with three levels of convolutional downsampling and upsampling (kernel size $= 3 \times 3$, initial kernel $= 64$, layer growth $= 2$) with ReLU activation applied to real and imaginary parts independently. Three different trainings were performed: 1) using trained IQ-Net with reference subtraction, 2) using trained EfficientNet with reference subtraction, and 3) using trained EfficientNet without a reference image. Considering the poor ranking performance of IQ-Net without reference image subtraction (see Results section), it was omitted for the denoising task. To investigate which image features were perceived to be important to the two scoring networks, saliency maps[34] were computed for IQ-Net and EfficientNet trained with and without reference subtraction.

For the image reconstruction task, a model-based DL (MBDL) reconstruction was performed, using IQ-Net trained with reference subtraction as a loss function and compared with MBDL trained with SSIM and MSE as losses. We undersampled the multichannel k-space data by Poisson disc with an acceleration factor of ~16X. The coil sensitivity maps were estimated using ESPIRiT.[35] In MBDL, we alternated between a Landweber gradient descent step based on data consistency and a NN-based image domain denoiser acting as a regularizer, as illustrated in Fig. S2 in the Supplemental Material. For the denoiser, a two-layer complex ResUNet (kernel size $= 3 \times 3$, initial kernel $= 64$, layer growth $= 2$) was used. The optimizer used was Adam, learning rate $= 10^{-4}$, decay rate $= 0.99$. The trained IQ-Net was fixed and used as the loss function. The same network was trained using MSE and SSIM losses. For inference, we tested undersampling rates of 8X and 16X on images not seen during training from the fastMRI dataset.

To probe the cause of the differences in the reconstructed images, the behavior of IQ-Net, SSIM and MSE with different types of image corruption was investigated. Corruptions detailed in a previous section were added to an image not included in the training and validation data.

Statistical Analysis

To compare the time required for ranking and Likert scoring, paired t-test was performed on the time collected from reader 5 and reader 7 ranking and scoring the same 52 pairs of images. We compared the accuracies from the five-fold cross validation of classification based on scores by IQ-Net and EfficientNet against classification based on MSE and SSIM using repeated measurements analysis of variance (ANOVA). For the reconstruction task, we reconstructed 500 images not included in the training and validation using a reconstruction network trained by IQ-Net scores and performed paired t-test compared to MSE and SSIM trained reconstruction networks.

Results

Ranking

Among the total of 19,344 rankings, radiologists 1–7 ranked 6797, 2108, 2108, 2102, 3585, 1722, and 2644 image pairs, respectively. The overall agreement among all radiologists was 81.4%. The intra- and inter-reviewer agreement is shown in Table 1 and Table S2 in the Supplemental Material. Overall,

Table 2. Agreement Between Readers 5 and 7 in Likert Scores (a), Rankings Converted From Likert Scores (b), and Direct Rankings (c)

(a) Agreement: Likert Scores

$k_w = \mathbf{0.47}, a = \mathbf{25}\%$

		Reader 7					Total
		1	2	3	4	5	
Reader 5	1	0	4	2	3	1	10
	2	0	2	13	10	0	25
	3	0	0	3	14	4	21
	4	0	0	1	11	24	36
	5	0	0	1	1	10	12
Total		0	6	20	39	39	104

(b) Agreement: Convert Likert Score to Ranking

$a = \mathbf{65.38}\%$

		Reader 7			Total
		Image 1	Image 2	Same	
Reader 5	Image 1	8	0	4	12
	Image 2	1	12	6	19
	Same	5	2	14	21
Total		14	14	24	52

(c) Agreement: Direct Rankings

$a = \mathbf{70.37}\%$

		Reader 7			Total
		Image 1	Image 2	Similar	
Reader 5	Image 1	23	5	0	28
	Image 2	2	13	2	17
	Similar	0	5	2	7
Total		25	23	4	52

a is an agreement metric described in Materials and Methods section and k_w is quadratic weighted Cohen's kappa.

the reviewers showed a high level of self-agreement ($94.9\% \pm 2.4\%$) while the inter-reviewer agreement was substantially lower ($61.47\% \pm 5.51\%$) and more variable, ranging from 54% to 78%. Reader agreement differentiated by image contrast (T1, T2, postcontrast T1, and FLAIR) can be found in Table S2 in the Supplemental Material.

Comparing Rankings and Likert Scores

Table 2 shows the confusion matrices, percentage agreement, and the quadratic weighted Cohen's kappa coefficient κ_w when comparing the rankings and Likert scoring between readers 5 and 7. Of the 52 pairs of images scored, 21 and 24 of the pairs were given the same Likert scores by the two readers, respectively (Table 2, panel b). In contrast, when the images were shown to the same radiologists simultaneously and ranked, only 7 and 4, respectively, of the pairs were reported to be of similar quality (Table 2, panel c). For Likert scoring individual images, the two radiologists had a low agreement of 25% (Table 2, panel a). When the scores were used to rank image pairs, agreement increased to 65%, while agreement for direct ranking of the same image pairs was 70%. Further, as shown in Table 3, the agreement between Likert ranking and direct ranking was 60% and 52% for readers 5 and 7, respectively. This was driven by the high prevalence of image pairs with the same Likert scores but with different direct ranks. The average time required for Likert scoring was 4.3 seconds per image (8.6 seconds per pair). For the same image pairs, direct ranking required an average of 3.95 seconds per pair, and based on the paired t-test, it is significantly faster.

Neural Network Ranking Performance

The scoring network and the classifier were trained with consensus ranking, i.e., with the majority vote of all reviewers. The accuracies of IQ-Net trained with and without reference standard images subtracted were $75.2\% \pm 1.3\%$ and $49.8\% \pm 10.0\%$, respectively. The accuracies of EfficientNet trained with and without reference standard images subtracted were $79.2\% \pm 1.7\%$ and $79.4\% \pm 1.2\%$, respectively. Classifiers based on MSE and SSIM achieved accuracies of $70.3\% \pm 2.3\%$ and $66.7\% \pm 2.3\%$, respectively. ANOVA tests showed that classification by scores by the IQ-Net with reference subtraction, EfficientNet with and without reference subtraction performed significantly better than classification by MSE and SSIM ($P < 0.001$).

Image Denoising Task and Saliency Maps

Figure 2 shows representative test images from denoisers trained using the different learned image quality networks. Image quality was subjectively poor when using EfficientNet with reference image subtraction where the resulting images were blocky and blurred. Without the reference image subtraction, the EfficientNet images still show substantial noise. For the IQ-Net-based training, a visually apparent reduction in noise levels was achieved with reference image subtraction with much loser MSE and higher SSIM than the images supervised by EfficientNet.

Table 3. Agreement Between Likert Scores and Direct Rankings for Readers 5 and 7

Reader 5		Percentage Agreement, $a = 59.62\%$			
		Image With Higher Likert Score			
		Image 1	Image 2	Same	Total
Image with higher ranking	Image 1	12	0	13	25
	Image 2	0	17	6	23
	Similar	0	2	2	4
Total		12	19	21	52
Reader 7		**Percentage Agreement, $a = 51.92\%$**			
		Image With Higher Likert Score			
		Image 1	**Image 2**	**Same**	**Total**
Image with higher ranking	Image 1	13	1	14	28
	Image 2	1	10	6	17
	Similar	0	3	4	7
Total		14	14	24	52

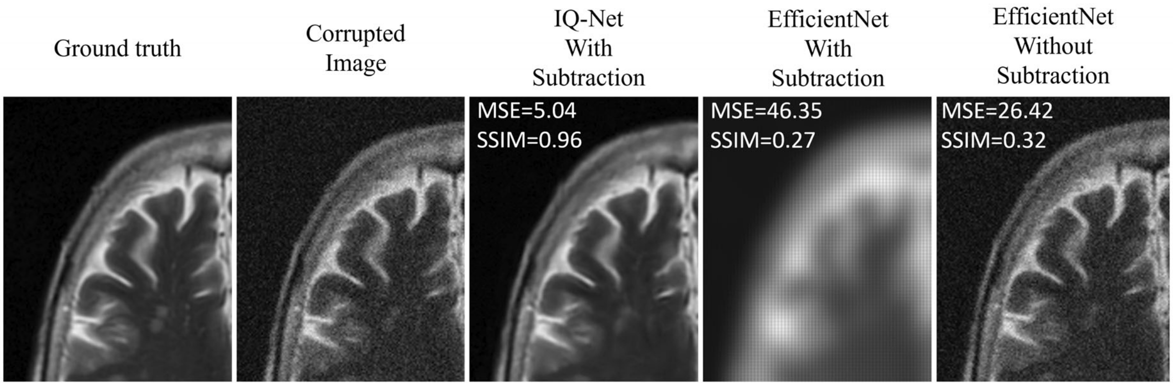

Figure 2: When training a neural network denoiser to remove complex Gaussian noise, the appearance of resulting images depended on the loss function, i.e. scores from different scoring networks.

Saliency maps of the three different scoring networks with respect to the input corrupted images are shown in Fig. 3 with three types of corruption as examples. These maps effectively show areas driving the output scores, and for known and simple corruption, should identify the features negatively impacting image quality. IQ-Net trained with reference image subtraction produced saliency maps that largely replicated the added noise, blurred edges, and aliasing caused by undersampling. Without reference image subtraction, IQ-Net failed to identify any features. On the contrary, despite good accuracy in classification, EfficientNet produced saliency maps with more random patterns. For example, in the case of blurring, the locations driving the score in EfficientNet were not localized around edges, as expected, but rather in central and flat regions of the image.

Image Reconstruction Task

Figure 4 shows 16X undersampled T1-weighted images reconstructed with a DL model trained with MSE, SSIM, and the IQ-Net with reference standard subtraction. Overall, SSIM-optimized reconstructions were noisier and showed artifactual patterns, as indicated by the solid boxes and arrows in the last column. MSE optimized and IQ-Net optimized reconstructions were comparable for T1-weighted postcontrast and T2-weighted images, with MSE loss being slightly more blurred. For T1-weighted and T2-FLAIR images, reconstructions trained with MSE were noisier in the white matter and the noise appeared to have patterns, especially around the white matter-CSF boundaries (dashed arrows and boxes in Fig. 4).

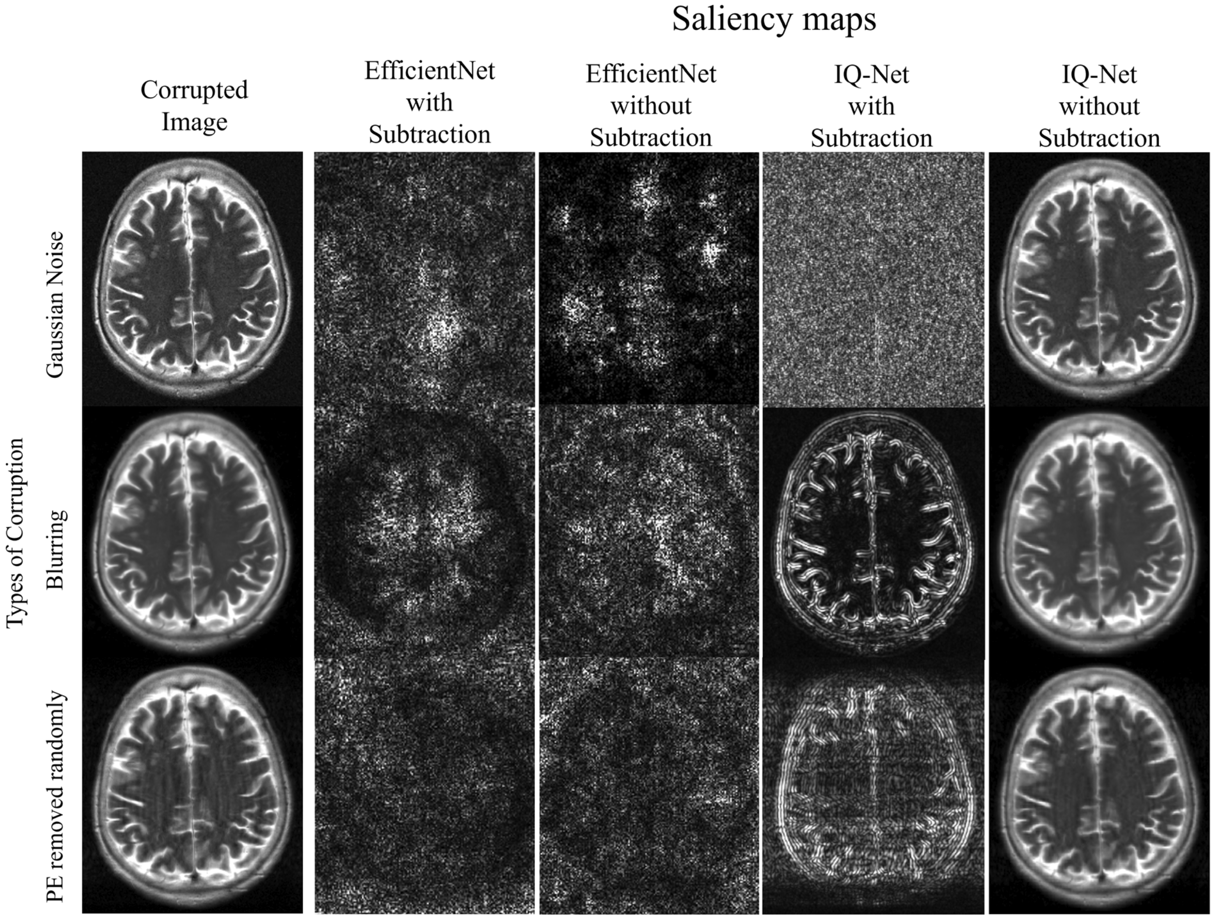

Figure 3: Saliency maps of the four different scoring networks (EfficientNet and IQ-Net, with and without reference standard image subtraction), showing the gradient with respect to the input corrupted images. Corruptions shown are Gaussian noise (top row), blurring (middle row) and undersampling (bottom row). In EfficientNet, the gradients are random and often not aligned with relevant brain structures. IQ-Net with a reference does produce images representing the corruption but only when a reference image was utilized.

Figure 5 shows the MSE, SSIM, and IQ-Net score (y-axis) of 500 images reconstructed with MBDL trained with MSE (blue), SSIM (yellow), and IQ-Net score (green) as loss functions at 8X (top row) and 16X (bottom row) acceleration. When using IQ-Net as a loss function for training the reconstruction network, the resulting images not only had the best (lowest) score, but lower MSE as well. The difference in IQ-Net scores and MSE between any two reconstruction networks were found to be statistically significant by t test ($P < 0.05$). This was true for all image sequences and both accelerations. Although the SSIM-trained reconstruction network consistently provided images with high SSIM, the artifacts highlighted in Fig. 4 suggest SSIM may not be a good representation of perceived MR image quality.

Comparing IQ-Net, SSIM, and MSE Sensitivities to Image Corruption

Figure 6a shows the correlations of IQ-Net scores, MSE, and SSIM over a set of corrupted images. Figure 6b shows the IQ-Net, MSE, and SSIM response to different image corruptions on one image (see Fig. S4 in the Supplemental Material for other corruption types). Overall, the IQ-Net is most similar in behavior to MSE, which is also reflected in MSE more accurately predicting the radiologist rankings. One factor that appears to drive this is a higher sensitivity to motion corruption, blurring, and undersampling compared to SSIM.

Discussion

In this study, a method to include radiologist preference in image quality assessment was developed based on image ranking. The evaluation of rankings identified higher intrareader agreement compared to interobserver agreement, suggesting personal preference in image quality even among radiologists. Compared to Likert scoring, ranking produced a higher level of agreement between radiologists, significantly reduced the image review time, and identified differences in image quality between images that were not captured when Likert scoring individual images. The scoring network IQ-Net achieved high accuracy in predicting consensus radiologist preferences,

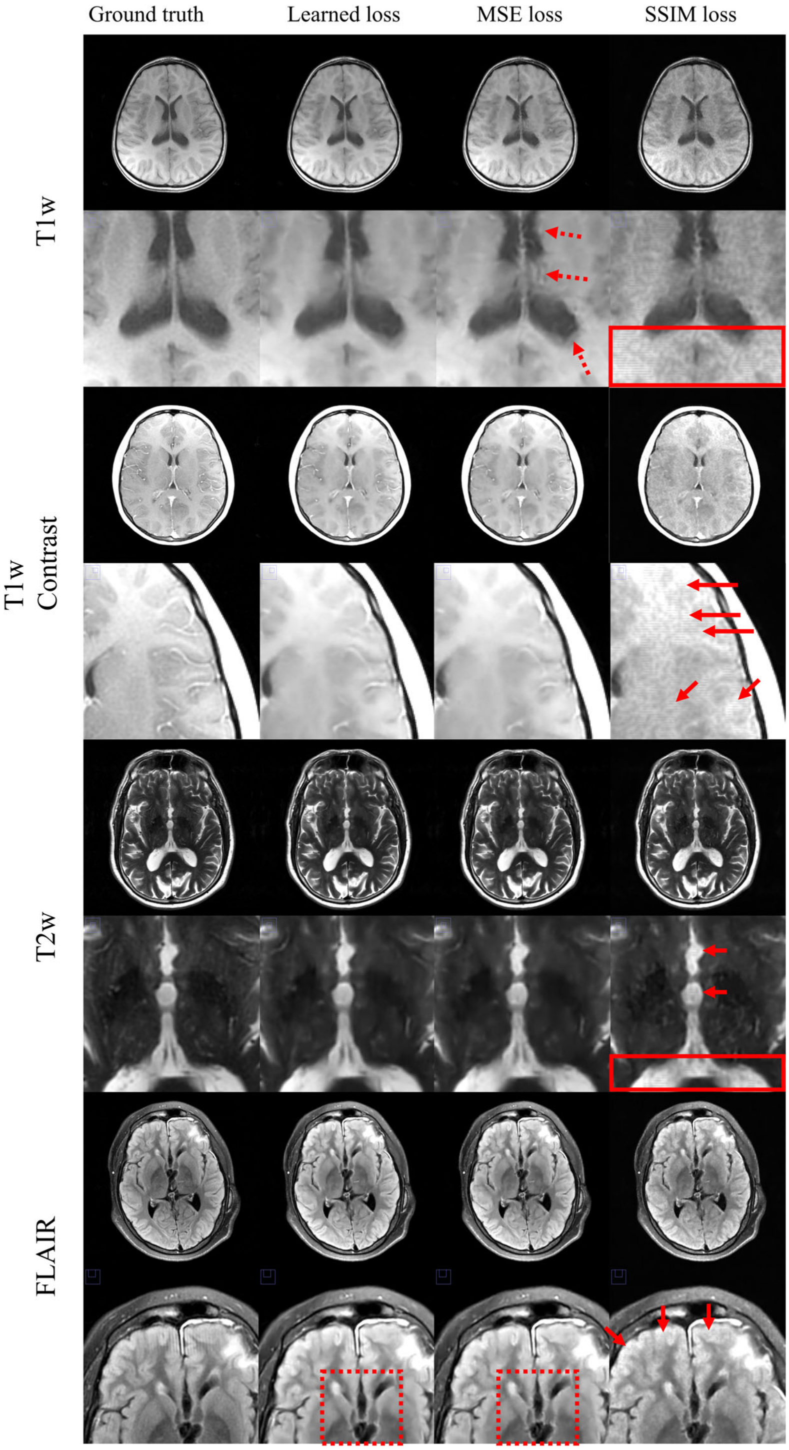

Figure 4: Reconstruction of T1-weighted, T1-weighted with contrast, T2-weighted and FLAIR images with deep learning networks trained with the learned IQ-Net, MSE, and SSIM loss (acceleration ~16X). At this high acceleration, some loss of detail is seen in all images; however, MSE images show greater artifacts (arrows) and SSIM images exhibit higher noise and a banding artifact (box).

with greater accuracy than both MSE and SSIM metrics. When used as a loss for training DL denoising and reconstruction, IQ-Net provided high-quality images.

EfficientNet, proved to be highly accurate in natural image classification,[33] was also used for scoring. While achieving high accuracy in rank training, EfficientNet-based score performed poorly when used as a loss function for image denoising. Effective gradient flow with respect to the denoiser parameters was observed, therefore, the poor performance was attributed to the failure of EfficientNet to capture MRI features relevant to human readers, which was confirmed in saliency maps. This likely arises from an out of distribution overfitting, as the training data set still contained a small number of rankings and corruptions relative to the number

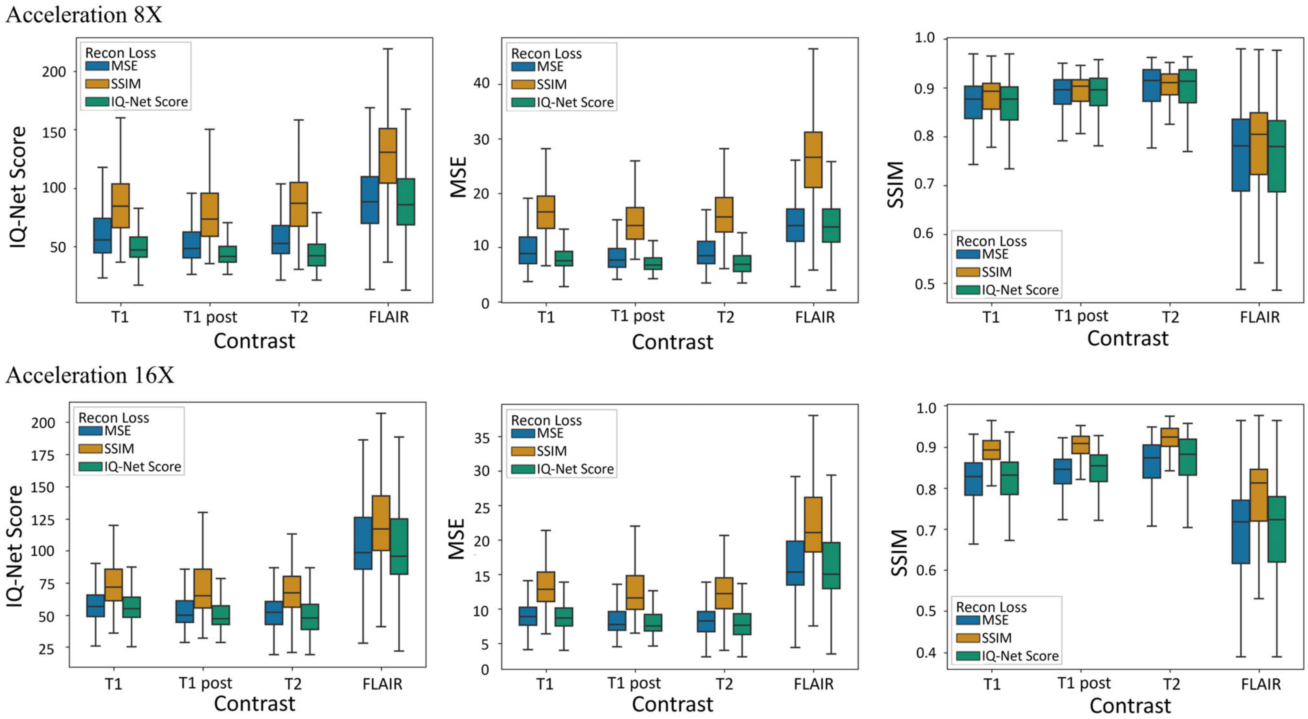

Figure 5: Evaluation of images reconstructed by networks trained with mean square error (MSE, blue), structural similarity (SSIM, yellow), and IQ-net loss (green) for 8X (top) and 16X (bottom) acceleration for T1-weighted, T1-weighted postcontrast, T2-weighted, and FLAIR images. Lower MSE, lower score, and higher SSIM indicate better image quality per that metric.

of parameters in the network. This type of overfitting is not captured in cross-validation since all data are generated in a similar way. In contrast, IQ-Net captured the added corruptions accurately when trained with reference images subtracted. When the input image was Gaussian blurred, the edges of the brain were highlighted in saliency maps, i.e., they were the features that drove the gradient. Similarly, with undersampling and Gaussian noise, IQ-Net was able to identify aliasing and the added noise. This ability to identify image corruption is the basis of understanding of what makes an MR image good to humans and thus informs downstream networks to perform well in tasks such as image denoising and reconstruction. This ability is likely enhanced by the small number of trainable parameters, which is less susceptible to overfitting. The use of more complicated architecture with more parameters may be achievable with a higher number of cases included and with an increase in the diversity of the artifacts, which could potentially further improve the ranking performance of the NN.

One of the novelties of this study is the usage of pairwise ranking. Likert scoring has been popular for the evaluation of image quality in non-AI and AI related reader studies[36,37] but is challenging to use in many applications. Major issues include baseline shifts between readers and readings, the dependence on subjective rubrics, and poor distinction between the scores. Baseline shifts in Likert scoring can be partially addressed using detailed rubrics and having the readers discuss cases to provide consensus scores. The rubric used in the current study, however, failed to compensate for the baseline shifts in scores. Consensus ranking is extremely time-consuming and can therefore only be performed by a small number of readers, opposing the aim of creating a general metric. When converting reader Likert scores to rankings, the inter-reader agreement improved substantially but was not as high as those for direct ranking, indicating pairwise ranking to be less sensitive to baseline shifts in scores than evaluating single images. In the current study, the readers frequently gave images the same score when the images were shown individually but had a preference for one of the images if they were shown simultaneously in pair wise ranking. The lack of delineation between images in Likert scoring reduces the effective information in the labels. As a whole, this supports pairwise ranking as a method to crowdsource image quality assessment, allowing for a high number of observers to contribute to a common cause and to aid in solving the need for a larger number of cases. Although this study itself is much smaller scale than would be desirable for universal metrics, such quality metrics have potential to be useful in developing commercial tools and methods that may cater to the wider community.

In this study, IQ-Net was also compared with MSE and SSIM which are commonly used engineering metrics. A single image was corrupted with typical corruptions that occur clinically and IQ-Net and MSE were found to behave similarly, although IQ-Net was more sensitive to motion. Meanwhile, SSIM was insensitive to both motion and blurring and

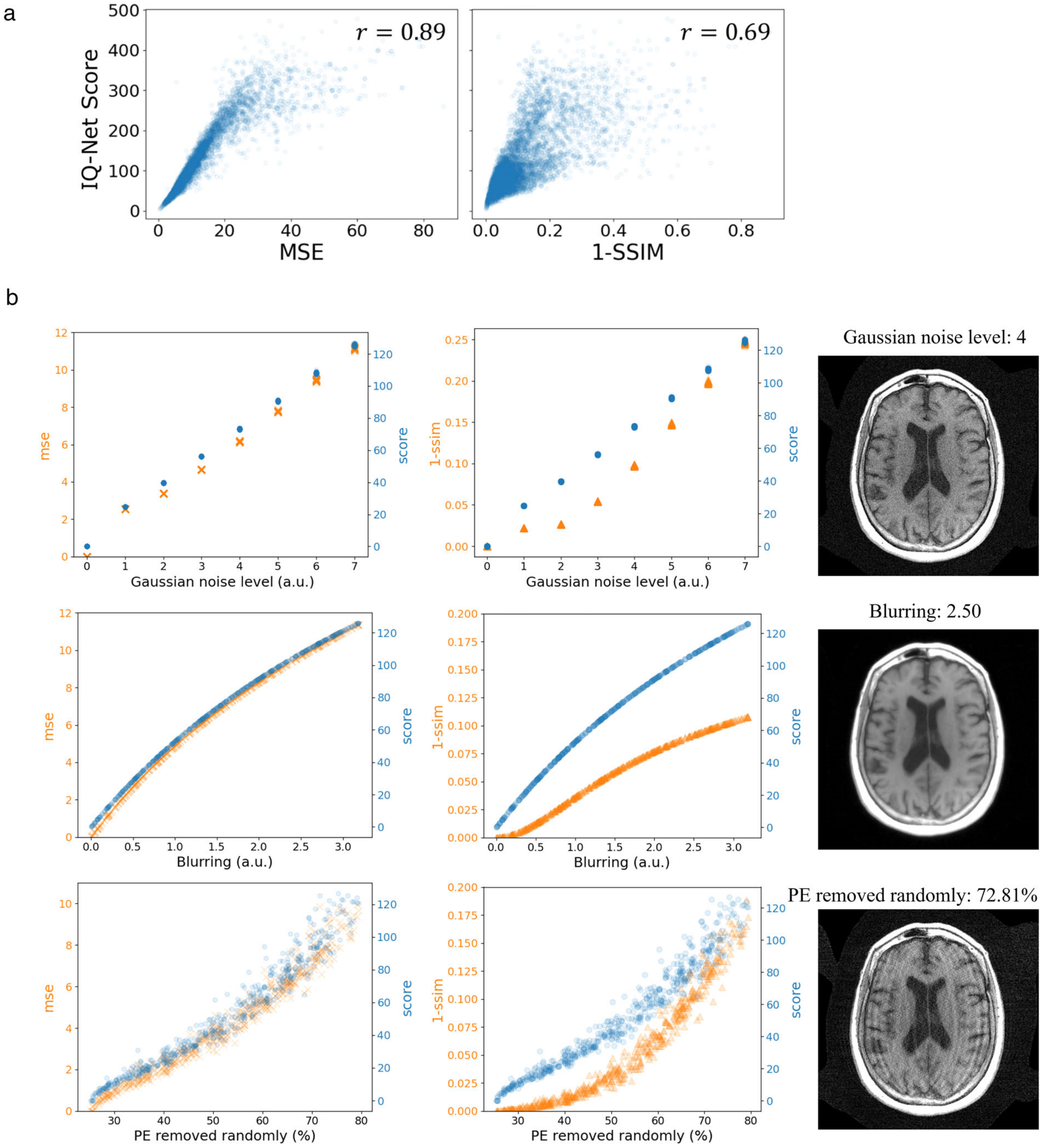

Figure 6: (a) Scatter plots and Pearson correlation coefficients comparing IQ-Net score with mean square error (MSE) and structural similarity (SSIM) for a set of corrupted brain images. Lower IQ-Net scores, MSE, and 1-SSIM indicate higher image quality. (b) MSE, SSIM and IQ-Net score for a single image with different types of corruption (left) and example images of each type of corruption (right).

showed a more nonlinear behavior for noise and undersampling. The different responses of the IQ-Net score may indicate how artifacts affect diagnostic quality. From this evaluation, motion artifacts are suggested to be the most destructive in radiologists' perception.

When used as a loss function for DL image reconstruction, the IQNet, SSIM, and MSE resulted in different textures in the reconstructed images. For example, SSIM loss resulted in higher apparent noise while MSE loss produced images that were more blurred in T1 postcontrast and T2-weighted images and showed structured artifacts. Images with radiologists preferred texture may help improve diagnostic confidence, reduce the changes of pathology being dismissed or obscured by artifact, and potentially improve image interpretation time. Interestingly, the reconstruction network trained with IQ-Net score loss also produced images with lower MSE than the network specifically trained with MSE. This was true at both 8X and 16X acceleration. It may suggest that IQ-Net could be a better choice than traditional metrics. SSIM provided the poorest image quality with a high level of noise and artifacts. Further work is needed to extend the results of this study to other applications and imaging modalities, and to improve our understanding of image preference.

Limitations

As discussed, it is a possibility to develop a tool for collecting pairwise rankings from a wider audience for various applications. However, one arising problem would be to determine the inclusion criteria for the rankings, for example, the number of rankings from each reader, the training background and experience level of the readers. In this pilot study, we included readers from a large range of professional backgrounds, which may have contributed to some disagreement. Also, considering the difference in their time commitment, the numbers of rankings from readers varied, which may introduce bias in the majority vote.

One other limitation worth mentioning is that IQ-Net requires reference standard images, which limits its application to unsupervised or self-supervised learning. The optimization of metrics for self-supervised tasks is still an open problem and this is not addressed by our current approach. Further investigation is also needed to compare IQ-Net with other types of specific feature losses such as the adversarial loss[18,19] and UFLoss.[11] These methods can be challenging to train and thus this study was limited to the commonly used SSIM and MSE metrics.

Lastly, although a large number of rankings from a group of radiologists were collected, we did not perform comparisons regarding readers' variation in training specialization (e.g., cardiac vs. neuro radiologists), or differences in institutions (e.g., academic vs. community). The study is retrospective only and the dataset only consisted of axial brain scans. Further evaluation would be required for other plane orientations, body locations, and modalities.

Conclusions

In this study, the viability of training a NN imitating radiologists' perception of image quality was demonstrated. The network was also used as a loss function for training networks to perform image-related tasks such as denoising and reconstruction. IQ-Net was superior to EfficientNet in image denoising and to MSE and SSIM in image reconstruction.

Acknowledgments

We gratefully acknowledge GE Healthcare for assistance and support. We received funding from NIH NIA R01AG075788, R21AG077337, NIH NINDS R21NS125094, and Oechtering T. H. received funding from the German Research Foundation (OE 746/1-1).

Data Availability Statement

Data used for the ranking task were obtained from the NYU fastMRI Initiative neuro database (fastmri.med.nyu.edu). Code and models used for preparation of this manuscript can be found at https://github.com/uwmri/LearnedImagingMetrics.

References

1. Chaudhari AS, Mittra E, Davidzon GA, et al. Low-count whole-body PET with deep learning in a multicenter and externally validated study. npj Digit Med 2021;4:1-11.
2. Sanaat A, Shiri I, Arabi H, Mainta I, Nkoulou R, Zaidi H. Deep learning-assisted ultra-fast/low-dose whole-body PET/CT imaging. Eur J Nucl Med Mol Imaging 2021;48:2405-2415.
3. Kulathilake KASH, Abdullah NA, Sabri AQM, Lai KW. A review on deep learning approaches for low-dose computed tomography restoration. Complex Intell Syst 2021;9:2713-2745.
4. Wu W, Hu D, Niu C, et al. Deep learning based spectral CT imaging. Neural Netw 2021;144:342-358.
5. Wang T, Lei Y, Fu Y, et al. A review on medical imaging synthesis using deep learning and its clinical applications. J Appl Clin Med Phys 2021; 22:11-36.
6. Chaithya GR, Ramzi Z, Ciuciu P. Hybrid learning of non-Cartesian k-space trajectory and MR image reconstruction networks. arXiv: 211012691 [eess, math]. 2021.
7. Aggarwal HK, Mani MP, Jacob M. MoDL: Model-based deep learning architecture for inverse problems. IEEE Trans Med Imaging 2019;38: 394-405.
8. Sriram A, Zbontar J, Murrell T, et al. End-to-end variational networks for accelerated MRI reconstruction. arXiv:200406688 [cs, eess]. 2020.
9. Yaman B, Hosseini SAH, Moeller S, Ellermann J, Uğurbil K, Akçakaya M. Self-supervised learning of physics-guided reconstruction neural networks without fully sampled reference data. Magn Reson Med 2020;84:3172-3191.
10. Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution. arXiv:160308155 [cs]. 2016.
11. Wang K, Tamir JI, De Goyeneche A, et al. High fidelity deep learning-based MRI reconstruction with instance-wise discriminative feature matching loss. arXiv:210812460 [cs, eess]. 2021.
12. Zhao H, Gallo O, Frosio I, Kautz J. Loss functions for image restoration with neural networks. IEEE Trans Comput Imaging 2017;3:47-57.
13. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. IEEE Trans Image Process 2004;13:600-612.
14. Wang Z, Simoncelli EP, Bovik AC. Multiscale structural similarity for image quality assessment. *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Pacific Grove, CA: IEEE; 2003. p 1398-1402.
15. Ghodrati V, Shao J, Bydder M, et al. MR image reconstruction using deep learning: Evaluation of network structure and loss functions. Quant Imaging Med Surg 2019;9:1516-1527.
16. Venkatanath N, Praneeth D, Bh MC, Channappayya SS, Medasani SS. Blind image quality evaluation using perception based features. *2015 Twenty First National Conference on Communications (NCC). Mumbai, India: IEEE*; 2015. p 1-6.
17. Tenenbaum JM. *Accommodation in computer vision*. Stanford, CA, USA: Stanford University;1971.
18. ISMRM. Unsupervised Image Reconstruction using Deep Generative Adversarial Networks. 2020. Available from: https://archive.ismrm.org/2020/0685.html.

19. Sandino CM, Cheng JY, Chen F, Mardani M, Pauly JM, Vasanawala SS. Compressed sensing: From research to clinical practice with deep neural networks: Shortening scan times for magnetic resonance imaging. IEEE Signal Process Mag 2020;37:117-127.

20. Qi C, Wang S, Yu H, et al. An artificial intelligence-driven image quality assessment system for whole-body [18F]FDG PET/CT. Eur J Nucl Med Mol Imaging 2023;50:1318-1328.

21. Bosse S, Maniry D, Wiegand T, Samek W. A deep neural network for image quality assessment. *2016 IEEE International Conference on Image Processing (ICIP)*. Phoenix, Az, USA: IEEE; 2016. p 3773-3777.

22. Krupinski EA. Current perspectives in medical image perception. Atten Percept Psychophys 2010;72:1205-1217. https://doi.org/10.3758/APP.72.5.1205.

23. Norman G. Likert scales, levels of measurement and the "laws" of statistics. Adv Health Sci Educ Theory Pract 2010;15:625-632.

24. Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian J Intern Med 2013;4: 627-635.

25. Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol 1. San Diego, CA: IEEE; 2005. p 539-546.

26. Liu X, Van De Weijer J, Bagdanov AD. RankIQA: Learning from rankings for no-reference image quality assessment. *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: IEEE; 2017. p 1040-1049.

27. Knoll F, Zbontar J, Sriram A, et al. fastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning. Radiol Artif Intell 2020;2: e190007.

28. Zbontar J, Knoll F, Sriram A, et al. fastMRI: An open dataset and benchmarks for accelerated MRI. arXiv:181108839 [physics, stat]. 2019.

29. Paszke A, Gross S, Massa F, et al. PyTorch: An imperative style, high-performance deep learning library. Adv Neural Inf Process Syst 2019; 32:8026-8037.

30. Pruessmann KP, Weiger M, Scheidegger MB, Boesiger P. SENSE: Sensitivity encoding for fast MRI. Magn Reson Med 1999;42:952-962.

31. Lustig M, Donoho DL, Santos JM, Pauly JM. Compressed sensing MRI. IEEE Signal Process Mag 2008;25:72-82.

32. ISMRM. SigPy: A Python Package for High Performance Iterative Reconstruction. 2019. Available from: http://archive.ismrm.org/2019/4819.html.

33. Tan M, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. arXiv:190511946 [cs, stat]. 2019.

34. Niebur E, Koch C. Control of selective visual attention: Modeling the "Where" pathway. In: Touretzky D, Mozer MC, Hasselmo M, editors. *Advances in neural information processing systems*, Vol 8.Cambridge, MA: MIT Press; 1995.

35. Uecker M, Lai P, Murphy MJ, et al. ESPIRiT—an eigenvalue approach to autocalibrating parallel MRI: Where SENSE meets GRAPPA. Magn Reson Med 2014;71:990-1001.

36. Esteban O, Blair RW, Nielson DM, et al. Crowdsourced MRI quality metrics and expert quality annotations for training of humans and machines. Sci Data 2019;6:30.

37. Piccini D, Demesmaeker R, Heerfordt J, et al. Deep learning to automate reference-free image quality assessment of whole-heart MR images. Radiol Artif Intell 2020;2:e190123.