IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 11 January 2023, accepted 2 February 2023, date of publication 8 February 2023, date of current version 15 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3243466

SURVEY

Image Quality Assessment for Magnetic Resonance Imaging

SERGEY KASTRYULIN[1,2], JAMIL ZAKIROV[2], NICOLA PEZZOTTI[3,4], AND DMITRY V. DYLOV[2], (Member, IEEE)

[1]Philips Research, 127051 Moscow, Russia
[2]Skolkovo Institute of Science and Technology, 121205 Moscow, Russia
[3]Philips Research, 5656 AE Eindhoven, The Netherlands
[4]Eindhoven University of Technology, 5612 AZ Eindhoven, The Netherlands

Corresponding author: Dmitry V. Dylov (d.dylov@skoltech.ru)

ABSTRACT Image quality assessment (IQA) algorithms aim to reproduce the human's perception of the image quality. The growing popularity of image enhancement, generation, and recovery models instigated the development of many methods to assess their performance. However, most IQA solutions are designed to predict image quality in the general domain, with the applicability to specific areas, such as medical imaging, remaining questionable. Moreover, the selection of these IQA metrics for a specific task typically involves intentionally induced distortions, such as manually added noise or artificial blurring; yet, the chosen metrics are then used to judge the output of real-life computer vision models. In this work, we aspire to fill these gaps by carrying out the most extensive IQA evaluation study for Magnetic Resonance Imaging (MRI) to date (14,700 subjective scores). We use outputs of neural network models trained to solve problems relevant to MRI, including image reconstruction in the scan acceleration, motion correction, and denoising. Our emphasis is on reflecting the radiologist's perception of the reconstructed images, gauging the most diagnostically influential criteria for the quality of MRI scans: signal-to-noise ratio, contrast-to-noise ratio, and the presence of artefacts. Seven trained radiologists assess these distorted images, with their verdicts then correlated with 35 different image quality metrics (full-reference, no-reference, and distribution-based metrics considered). The top performers – DISTS, HaarPSI, VSI, and $\text{FID}_{\text{VGG16}}$ – are found to be efficient across three proposed quality criteria, for all considered anatomies and the target tasks.

INDEX TERMS Image quality, deep learning, metrics, reconstruction quality, MRI.

I. INTRODUCTION

Image quality assessment (IQA) is a research area occupied with constructing accurate computational models to predict the perception of image quality by human subjects, the ultimate consumers of most image processing applications [1].

The growing popularity of image enhancement and image generation algorithms increases the need for a quality assessment of their performance. The demand has led to the abundance of IQA methods emerging over the last decades. The well-known *full-reference* (FR) metrics, such as MSE, PSNR, and SSIM [2], [3], became a de-facto standard in many computer vision applications. The more recent *no-reference* (NR) metrics, such as BRISQUE [4], have also found their use, especially when the ground truth images are absent or hard to access. Yet, another class of *distribution-based* metrics (DB) earned the community's attention, thanks to the advent of generative adversarial networks (GANs), enabling the quality assessment using distributions of thousands of images instead of gauging them individually. The popular new DB IQA methods include such metrics as Inception Score [5], FID [6], KID [7], MSID [8], and many others. Despite being widely used, the DB metrics were neither included in the recent large scale general domain reviews [9], [10], nor in the medical ones [11].

IQA measures are applied to estimate the quality of image processing algorithms and systems. For example, when several image denoising and restoration algorithms are

The associate editor coordinating the review of this manuscript and approving it for publication was Jiachen Yang.

available to recover images distorted by blur and noise contamination, a perceptual objective IQA could help pick the one that generates the best perceptual image quality after the restoration. To do that reliably, Image Quality Metrics (IQMs) need to show a high correlation with the perceptual estimates of the quality reported by human subjects for a given image processing algorithm. However, IQA algorithms are often *evaluated on non-realistic distortions*, such as added noise or artificial blurring [9], [10], [11]. Such a discrepancy between the synthetic evaluation and the practical use may cause misleading results.

While most metrics are designed to predict image quality in the general domain, Magnetic Resonance Imaging (MRI) provides gray-scale data with the content and style noticeably different from the natural images. Hence, the applicability of the IQMs in the MRI domain must be validated.

Moreover, IQMs trained on natural images attempt to describe the overall perception of the quality of an entire scene. On the contrary, an MRI scan can be perceived as high-quality when specific characteristics, responsible for the scan's value, are deemed adequate. Those are the characteristics that are deemed important components of the radiographic image quality [12], including perceived level of noise (signal-to-noise ratio, SNR), perceived soft tissue contrast (contrast-to-noise ratio, CNR), and the presence of artefacts. Unfortunately, none of the previous IQM studies considered them. Besides, these specific quality criteria are coupled. For example, some denoising algorithms tend to introduce additional blurring (lowering of CNR) in exchange for increased SNR, and some motion correction approaches tend to introduce noticeable artefacts. Therefore, a more detailed evaluation of IQM's ability to express separate MRI quality criteria is required.

The remainder of this paper is structured as follows. After discussing the related work, we describe how we generate an image library that consists of disrupted and reference MRI image pairs. In Section IV-A, we provide a detailed description of data selection, corruption, and restoration processes that populate the image library with realistic yet diverse data. We then use the image library to survey expert radiologists and collect a set of labels to be then correlated with IQM values in Section IV-C. Finally, we report and discuss the results in Sections V and VI, where we indicate the top-performing metrics, and provide insights about their performance for different distortions, robustness to the domain shift, anatomies, and quality criteria. Section VII concludes the work by proposing the best IQA approaches for MRI. Appendices include a list of abbreviations (A), reconstruction examples (B), and a screenshot of the labeling user interface (C).

The main contributions of this paper are:

- The most extensive study of IQA in medical imaging, in general, and in MRI, in particular (14,700 subjective scores collected). Unlike previous metric evaluation studies, we avoid artificially added distortions and assess the outputs of *popular image restoration models* instead. The assessment is based on *proposed three criteria* and allows us to make profound conclusions on what modern metrics can capture and when exactly they should be used.
- To the best of our knowledge, we provide the first thorough study of the application of DB metrics for objective IQA of both natural and medical images. We evaluate their performance and show when they give advantage over the common FR and NR metrics. We study the robustness of metrics' performance across these two vastly different domains and show that the best performing IQMs produce valuable results even when the data distribution drastically changes.

II. RELATED WORK

The evaluation of metrics for IQA in the domain of natural images started from the early task-specific works that considered FR methods to characterize color displays and half-toning optimization methods [13].

More recent task-specific studies explored IQA for the images of scanned documents [14] and screen content [15]. Likewise, fused images [16], smartphone photographs [17], remote sensing data [18], and climate patterns [19] demanded the development of targeted IQA approaches. Historically, many of these works have been focusing on the quality degradation caused by the compression algorithms [19], [20], [21], [22], with relatively small datasets appearing publicly for the IQ evaluation. However, the small dataset size and the excessive re-use of the same test sets have led to the promotion of the IQMs poorly generalizable to the unseen distortions.

This was recognized as a major problem, stimulating the emergence of large-scale studies [9], [23]. Among the large-scale evaluations, the majority compared multiple FR metrics, ranging from just a handful [24], [25], [26], [27] to several dozens [9], [28] of IQMs analyzed on popular datasets.

The medical domain stands out from the others by a special sense of what is deemed informative and acceptable in the images [29]. Resulting from years of training and practice, the perception of medical scan quality by adept radiologists relies on a meticulous list of anatomy-specific requirements, on their familiarity with particular imaging hardware, and even on their intuition.

Given the majority of IQMs were not designed for the healthcare domain, some recent works were dedicated to the niche. One small-scale study considered a connection of IQA of natural and medical images via SNR estimation [30]. Others assessed common FR IQMs using non-expert raters [31], [32]. Sufficient for the general audience, these methods proved incapable of reflecting the fine-tuned perception of the radiologists [33].

Expert raters were then engaged in [11] and [34]. The former studied only IQMs from the SSIM family and the latter assessed 10 FR IQMs, reporting that VIF [35], FSIM [36],

and NQM [37] yield the highest correlation with the radiologists' opinions.

On the other hand, Crow et al. argue that NR IQA are preferable for assessing medical images because there may be no perfect reference image in the real-world medical imaging [38]. To address this issue, several recent studies also propose new NR IQMs for MRI image quality assessment [39], [40], [41], [42], [43], [44], [45], [46], [47], [48]. The recent survey [49] overviews MRI-specific IQMs and concludes that the number of available metrics is relatively low and that their development is hindered by the lack of publicly available datasets. Also, none of these new metrics have an open-source implementation, making verification of the claimed results problematic.

III. IMAGE QUALITY METRICS CONSIDERED

In this work, we evaluate the most widely used and publicly available general-purpose FR, NR, and DB IQMs to find the best algorithms for the quality assessment on arguably the most important MRI-related image-to-image tasks: *scan acceleration*, *motion correction*, and *denoising*. Instead of modeling the disrupted images, we use outputs of trained neural networks and compare them with the clean reference images from the fastMRI dataset.

Our study includes the following 35 metrics: **17 Full-Reference IQMs** (PSNR, SSIM [2], MS-SSIM [50], IW-SSIM [51], VIF [35], GMSD [52], MS-GMSD [53], FSIM [36], VSI [54], MDSI [55], HaarPSI [56], Content and Style Perceptual Scores [57], LPIPS [58], DISTS [59], PieAPP [60], DSS [61]), **3 No-Reference IQMs** (BRISQUE [4], PaQ-2-PiQ [62], MetaIQA [63]), and **15 Distribution-Based IQMs** (KID [7], FID [6], GS [64], Inception Score (IS) [5], MSID [8], all implemented with three different feature extractors: Inception Net [65], VGG16, and VGG19 [66]). For brevity of the presentation, we will showcase only the analysis of the best performing four metrics, in the order of their ranking: VSI [54], HaarPSI [56], DISTS [59], and $\text{FID}_{\text{VGG16}}$ [6]. All metrics were re-implemented in Python to enable a fair comparison, with the PyTorch Image Quality (PIQ) [67] chosen as the base library for implementing all metrics. The resulting implementations were verified to be consistent with the original implementations proposed by the authors of each metric.

Noteworthy, in our survey, we dismissed some recent results reported for the PIPAL dataset [68] during the 2021 NTIRE challenge [69], because the winners [70], [71], [72] released no official implementations or model weights at the time of our experiments.

For the comparison, we collect 14,700 ratings from 7 trained radiologists to evaluate the quality of reconstructed images based on three main criteria of quality: perceived level of noise (SNR), perceived soft tissue contrast (CNR), and the presence of artefacts, making this work the most comprehensive study of MRI image quality assessment to date.

IV. MEDICAL EVALUATION

The key goal of this study is to evaluate popular selected IQMs on MRI data. Previous works [11], [34] evaluated the ability of certain IQMs to assess overall quality of data after to various types of artificial distortions.[1] However, in practice, the overall image quality (IQ) rating may be insufficient due to its ambiguity: *e.g.*, one could not truly interpret the reasons for poor or good scoring. At the same time, asking the medical experts these general questions may be challenging because of many factors, ranging from the specifics of certain clinical workflows to personal preferences.

In this work, we aspire to solve these problems by proposing the following study. First, we evaluate IQMs with regard to their ability to reflect radiologists' perception of the quality of distorted images, comparing them to the fully-sampled artifact-free ones. We range the metrics based on three IQ criteria that are crucial for making clinical decisions: perceived level of noise (SNR), perceived level of contrast (CNR), and the presence of artefacts. Second, instead of corrupting images with artificial perturbations, for the first time in the community, we validate these metrics using the actual outputs of deep learning networks trained to solve common MRI-related tasks. As such, the artefacts originate from the imperfect solutions to the common real-world problems of motion correction, scan acceleration, and denoising.

A group of trained radiologists rated the quality of distorted images compared to the clean reference images on a scale from 1 to 4 for the three IQ criteria. Unlike the five-point Likert scale, the simplified scale balances the descriptiveness of the score with the noise in the votes of the radiologists. Our mock experiments showed that the respondents considered the selection between too many options difficult, with the five-point scale having a diluted difference between the options; whereas, the three-point scale was deemed insufficient.

After the evaluation, the aggregated results were compared with the values of selected IQA algorithms to identify the top performers – the metrics that correlate the highest with the radiologists' votes.

A. IMAGE LIBRARY GENERATION

As a data source, we use the largest publicly available repository of raw multi-coil MRI *k*-space data – the FastMRI dataset, containing the knee and the brain scans [73], [74]. The knee subset of FastMRI contains 1,500 fully sampled MRIs acquired with a 2D protocol in the coronal direction with 15 channel knee coil array on 3 and 1.5 Tesla Siemens MRI machines. The data consists of approximately equal number of scans acquired using the proton density weighting with (PDFS) and without (PD) fat suppression pulse sequences with the pixel size of 0.5 mm × 0.5 mm and the slice sickness of 3 mm.

[1] These artificial distortions, *e.g.*, blurring or JPEG artefacts, are rarely encountered or even impossible in MRI practice.

FIGURE 1. Distortions introduced to initial artefact-free scans during training and inference. Using the raw *k*-space data of the reference images, we undersample them with the acceleration factor of 4, impose rigid motion of a moderate amplitude, and introduce mild Gaussian noise. Note how the distortions differ from those in the Natural Images, on which the common IQMs were developed. We adjusted brightness for viewer's convenience.

The knee subset is divided into 4 categories: train (973 volumes), validation (199 volumes), test (118 volumes), and challenge (104 volumes). Only the multi-coil scans were selected for this study, omitting the single-coil data.

The brain subset includes 6,970 1.5 and 3 Tesla scans collected on Siemens machines using T1, T1 post-contrast, T2, and FLAIR acquisitions. Unlike the knee subset, this data are of a wide variety of reconstruction matrix sizes. For the purpose of de-identification, authors of the dataset limited the data to only 2D axial images, and replaced *k*-space slices $\gtrapprox$ 5 mm below the orbital rim with zero matrices. The brain subset is divided into 6 categories: train (4,469 volumes), validation (1,378 volumes), test 4× (281 volumes), test 8× (277 volumes), challenge 4× (303 volumes), and challenge 8× (262 volumes).

Starting with the clean knee and brain data, we first generate images corrupted with three types of distortions: scan acceleration, motion, and noise. The examples of the distorted images are presented in Fig. 1. After that, we train two reconstruction models for each type of distortions using PyTorch [75]. The first model is trained until the validation loss is stabilized. The second model is trained for half as long to purposely produce imperfect reconstructions, oftentimes encountered in practice. Examples of corrupted images and the corresponding reconstructions can be found in Appendix B. The reduced training time was a conscious choice, enabling the model to produce some visible reconstruction errors. More specifically, we interrupted the training when the 90% of the loss plateau is reached, which allows for good performing models with imperfections we wanted to test for.[2] Finally, we use the trained models to reconstruct the corrupted images in the *validation subset* of the FastMRI, from which we generate the labeling dataset.

1) SCAN ACCELERATION

Scan-acceleration data are generated from the ground truth images by undersampling the *k*-space data. To train the model, we selected only T1 weighted scans (T1, T1-PRE and T1-POST) from the train category of the FastMRI brain data. The same subset of data was used for training of motion correction and denoising models. The *k*-space data were subsampled using a Cartesian mask, where *k*-space lines are set to zero in the phase encoding direction. The sampled lines are selected randomly, with the total sampling density depending on the chosen acceleration rate. Following the data generation process from the FastMRI challenge [73], all masks are fully sampled in the central area of *k*-space (the low frequencies). For the 4× accelerated scans, this corresponds to 8%, and for the 8× acceleration, it equals to 4%. Besides making the reconstruction problem easier to solve, such lines allow computing the low-pass filtered versions of the images for assessing the coil sensitivity maps.

To compensate for the undersampling, we used the 2019 FastMRI challenge winner Adaptive-CS-Net model [77]. Based on the Iterative Shrinkage-Thresholding Algorithm (ISTA) [78], this model consists of several trainable convolutional multi-scale transform blocks between which several prior knowledge-based computations are implemented. For scalability reasons and without substantially impacting the reconstruction results, in this study, we trained a simplified light-weight version of the Adaptive-CS-Net model. The resulting model consists of only 10 trainable blocks and 267k parameters. Unlike the full Adaptive-CS-Net model with three MRI-specific physics-inspired priors, the simplified version has only one prior module between the reconstruction blocks – the soft data consistency step. Specifically, the update for the block B_{i+1} in the simplified Adaptive-CS-Net model is defined as follows:

$$B_{i+1}(x_i) = x_i + \hat{\mathcal{U}}_i(soft(\mathcal{U}_i(x_i, e_i), \lambda_{s,f_s})), \tag{1}$$

where x_i denotes the i-th estimate of reconstruction, $\mathcal{U}$ and $\hat{\mathcal{U}}$ are the multi-scale transform and its inverse that consist of 2D convolutions and a nonlinearity in the form of Leaky-ReLU. The feature maps produced at the different scales are thresholded using the soft-max function $soft(\cdot)$,[3] parameterized by a learned parameter λ_{s,f_s} for each feature channel f_s and scale s. In Eq. 1, the soft data consistency step e_i is defined as follows:

$$e_i = \mathcal{F}^{-1}(M\mathcal{F}x_i - My), \tag{2}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote Fourier transform and its inverse, My is the data measured with the sampling mask M.

We trained the simplified Adaptive-CS-Net model using RMSprop optimizer [79] to minimize L1 loss function between the reconstruction estimate and the ground truth image obtained from the fully sampled data. We used a

[2] It is a standard way to broaden the image distribution from which the samples are drawn for evaluation and voting (*e.g.*, see [76]).

[3] Defined as $soft(u, \lambda) = \max(|u| - \lambda, 0) \cdot \frac{u}{|u|}$.

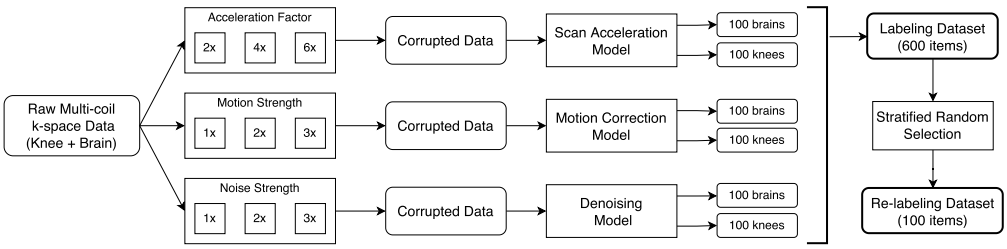

Image Group	N	Ratings per image
By radiologist		
Radiologists 1-7	600	1
Combined	600	7
By degradation type		
Scan Acceleration	200	7
Motion Correction	200	7
Denoising	200	7
By modality		
T1w	300	7
TSE	300	7

FIGURE 2. Formation of Labeling and Re-labeling datasets for annotation (left) and the content of each group of images for the medical evaluation and labeling by radiologists in the Labeling dataset (right). Starting from the clean validation data, we first generate corrupted data with the acceleration artefacts, the motion artefacts, and the Gaussian noise. Then, we reconstruct the corrupted data using trained neural network models and randomly select scans to form labeling and re-labeling pairs for the experts to grade.

step-wise learning rate decay of 10^{-4} and the batch size of 8 to reconstruct the data for various acceleration factors (from $2\times$ to $8\times$).

2) MOTION CORRECTION

The in-plane motion artefacts, including rigid translation and rotation, were introduced into the Fourier-transformed data following the procedure described in [80]. For each input image, the assumed echo-train length of the turbo spin-echo readout was chosen randomly in the 8—32 range. Similarly, the assumed extent of zero-padding in *k*-space was chosen randomly in the range of 0–100. The motion trajectories (translation/rotation vectors as a function of scan time) were generated randomly to simulate the realistic artefacts. In this study we utilized the protocol for ''sudden motion'' simulation. Here, the subject is assumed to lie still for a large part of the examination, until a swift translation or rotation of the head occurs. The time point of the sudden motion was taken randomly as a fraction of the total scan time in the range of one-third to seven-eighths. The maximum magnitude of the motion was chosen randomly from the range of [1], [4] pixels for the translation and [0.5,4.0] degrees for the rotation artefacts. The center of rotation was also varied randomly in the range of [0,100] pixels in each direction. These parameter ranges were selected empirically to generate a large variety of realistic artefacts and were used consistently in the training and in the validation runs.

To compensate for the motion artefacts of various extent, we trained U-Net models [81] with 209k parameters. While more advanced architectures exist, we found the basic U-Net to be more than sufficient for the scope of the proposed IQA study, as it is enough to capture imperfections which are often generated by deep learning models.

The model received the motion corrupted data as the input and learned to predict the motion artefacts in a residual manner, *i.e.*, the output of the model was a predicted image of motion in the input data. The model was trained to minimise L1 loss between the ground-truth and the predicted residual with Adam [82] optimizer using the step-wise learning rate decay of 10^{-4} and the batch size of 8. Preserving the same nature of artefacts, we trained our models for a range of amplification factors (from 1 to 3). For that, throughout the training, the motion amplitude was scaled by the amplification factor, yielding a consistently diverse appearance of the motion artefacts that could be met in practice.

3) DENOISING

In our study, noisy *magnitude* images are generated from the complex *k*-space data with the Gaussian distribution taken as the representative noise model. Below, the standard deviation of the Gaussian noise is reported for a region of interest in the background of the magnitude image, as proposed in [83]. The parameters of the noise distribution for each volume are drawn from the last slice of this volume. Then, the Gaussian noise with the estimated distribution parameters is generated, scaled by an amplification factor, and added to all images of the volume. We used the amplification factor of 2 for the training and the amplification factors of 1, 2, and 3 for the test data generation to enrich the variety of the tested image qualities in the resulting dataset.

To compute the denoised images, we trained DnCNN models [84] with 556k parameters on the brain multi-coil train data using the RMSprop optimizer [79] and a step-wise learning rate decay of 10^{-4} with the batch size of 8. Similarly to the other tasks considered herein, we are not looking for the most powerful denoising algorithms but consider a very commonplace model DnCNN instead, merely to rank the modern IQA metrics for the specific task of denoising.

4) FINAL DATASET FOR LABELING

We started the formation of the labeling dataset from the clean volumes from the validation subsets of brain and knee FastMRI datasets; hence, these scans were not used to train the artefacts correction models. In total, both validation subsets contain 1,577 volumes, resulting in 28,977 images: 199 knee volumes with 7,135 slices and 1,378 brain volumes with 21,842 slices. In each brain volume, the lower 2 and top 3 slices were discarded to restrict the analysis to clinically relevant parts of the scan. In each knee volume, the first 3 slices were discarded for the same reason. To limit the number of data points and decrease the overall variability of data types, we selected only T1-weighted (T1, T1-PRE and T1-POST) brain volumes and proton-density weighted without fat suppression (PD) knee volumes.

The data generation pipeline is summarized in Fig. 2.

Using the selected subset of clean validation data, we simulated images for the reconstruction:

- For the scan acceleration task, we simulated acceleration artefacts for undersampling rates of 2×, 4×, and 6×, following the data generation process from the FastMRI challenge;
- For the motion correction task, we simulated motion artefacts of three different strengths using the rigid motion simulation framework described above;
- For the denoising task, we simulated Gaussian noise with amplification factors of 1, 2 and 3 using the noise generation procedure described above.

After that, all generated corrupted data were reconstructed using the reconstruction models trained for the corresponding tasks. Note that we deliberately generated a fraction of data with parameters different from the ones used to train the reconstruction models. We found this approach yields various levels of artefacts typically appearing after the reconstruction process.

From the large pool of reconstructed images, we select 100 pairs of images (clean - reconstructed) for each task (scan acceleration, motion correction, denoising) and each anatomy (knee, brain), evenly distributing the data to represent each reconstruction parameter (*e.g.*, the acceleration rate for the scan acceleration task). This strategy results in the labeling dataset of 600 pairs of images in total (3 tasks × 2 anatomies).

To reach the goal labeling dataset size, we utilized the following data selection procedure:

1) Compute values of IQMs for all reconstructed images (for NR IQMs) or image pairs (for FR IQMs);
2) Normalize each IQM value to [0, 1];
3) Compute variance between IQM values for all items;
4) Sort all items by the value of variance;
5) Select 25% of data for each task-anatomy combination from the data items with the highest variance, assuming that items with the biggest disagreement between IQMs are the most informative;
6) Select the rest 75% of data pseudo-randomly (preserve distribution of reconstruction parameters) to avoid introducing any bias from the variance computation.

Lastly, we deliberately duplicated 100 of the 600 prepared items for the purpose of verification of radiologists' self-consistency, resulting in 700 image pairs to be labelled by each radiologist.

B. EXPERIMENT SETUP

Within the paradigm of the model observer framework [85], the quality of a medical image can be defined as how well a clinical task (*e.g.*, diagnostics) can be performed on it [86]. This means that the perfect MRI IQM would be some task-based score, such as the diagnostic accuracy. However, such a metric is difficult to implement due to a great diversity of diagnostic outcomes that radiologists deal with in practice. Because of that, the convention is to use a *subjective estimation* of the overall diagnostic value instead [11].

However, we argue that a single score is not sufficient to reflect the abundance of anatomies, pathologies, and artefactual cases that the radiologists work with. Instead, we propose to subdivide the score of the overall diagnostic quality into three main criteria that can be important for a clinical practitioner to make their decision: i) perceived level of noise, ii) perceived level of soft-tissue contrast, and iii) presence of artefacts.

1) SUBJECTIVE EVALUATION

Seven trained radiologists with 7 to 20 years of experience took part in this study. The participants were asked to score pairs of reconstructed-reference images using three main IQ criteria. For each image pair and each criterion, radiologists scored the perceived diagnostic quality of the reconstructed image compared to the ground-truth using a four-point scale: not acceptable (1), weakly acceptable (2), rather acceptable (3), and fully acceptable (4). The four-point scale was selected over the five-point Likert scale, previously used in [11].

Each participant performed the labeling individually using a dedicated instance of the Label Studio [89] software accessible via a web interface. The experts were asked to make all judgments about the image quality with regard to a particular diagnostic task that they would normally perform in their practice (*e.g.*, the ability to discriminate relevant tissues, the confidence in using the image to detect a pathology, *etc.*). The interface provided additional functionality of scaling (zooming) the images to closer mimic the real-life workflow. The pairs of images were displayed in a random order until all pairs were labelled. Participants had an opportunity to re-label the pairs they have already scored at any point until the experiment is finished.

During the main part of the experiment, each participant labelled 600 pairs of images based on the 3 quality criteria, resulting in 4,200 annotated pairs and 12,600 labels in total. The results of the main labeling session were used for further evaluation of the IQMs. After finishing the main part of the experiment, the participants were asked to additionally label 100 randomly selected pairs from the same dataset, yielding additional 2,100 labels. The results of this additional re-labeling were used to evaluate the self-consistency of each annotator.

2) METRICS COMPUTATION

Unlike FR and NR IQMs, designed to compute an image-wise distance, the DB metrics compare distributions of *sets* of images. This makes them less practical for traditional IQA, the goal of which is to compute a score for a given image pair. Moreover, the need to have sets of images hinders the vote-based evaluation via the mean subjective opinion scores.

To address these problems, we adopt a different way of computing the DB IQMs. Instead of extracting features from the whole images, we crop them into overlapping tiles of size 96×96 with $stride = 32$. This pre-processing allows us to

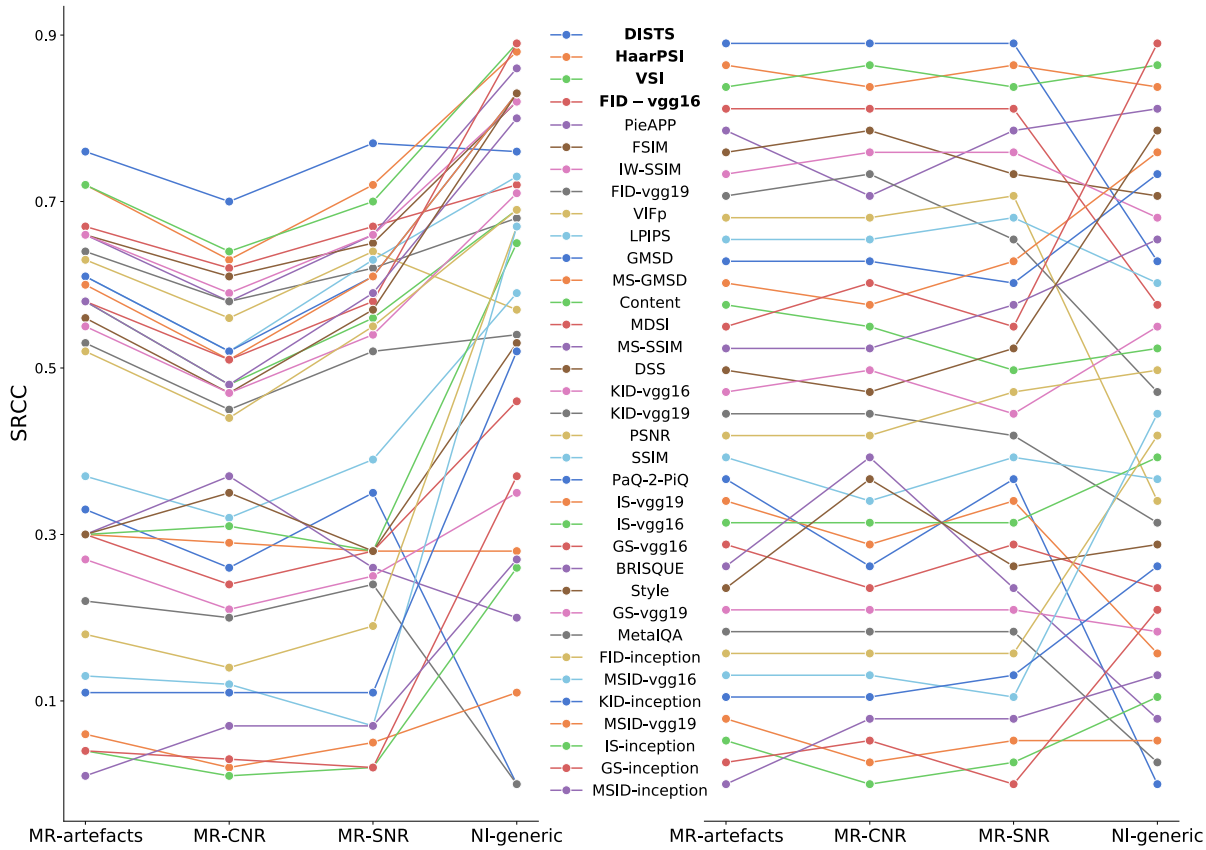

FIGURE 3. Performance of IQMs on different MRI tasks and on Natural Images (NI), compared by their correlation with the expert votes (SRCC values, left) and sorted top-to-bottom by their rank (right). The ordering reflects the performance on the MRI data only. The same color-coding is used in both plots. *NI-generic* scores are the average between TID2013 [87] and KADID-10k [88] datasets. Note higher correlation of IQMs on NI and poor translation of ranking to MRI domain. Refer to data in Table 1 for numerical values.

treat each pair of images as a pair of distributions of tiles, enabling further comparison. The other stages of computing the DB IQMs are kept intact.

C. DATA ANALYSIS

Here, we adapt the analysis of the scoring data proposed in [24] to the multiple IQ criteria. The voting scores for each scoring criteria are not analyzed in their raw format. Instead, they are converted to z-scores (averaged and re-scaled from 0 to 100 for each radiologist to account for their different scoring):

$$z_{nmk} = (D_{nmk} - \mu_{mk})/\sigma_{mk}\,, \tag{3}$$

where μ_{mk} and σ_{mk} are the mean and the standard deviation of the difference scores of the m^{th} radiologist on the k^{th} scoring criteria, and D_{nmk} are the difference scores for n^{th} degraded image defined as follows:

$$D_{nmk} = s_{mk,\text{ref}} - s_{nmk}\,. \tag{4}$$

In Eq. (4), $s_{mk,\text{ref}}$ is the raw score of the m^{th} radiologist on the k^{th} scoring criteria for the reference image corresponding to the n^{th} degraded image, and s_{nmk} is the raw score of the m^{th} radiologist on the n^{th} degraded image on the k^{th} scoring criteria. Note that in this study, the radiologists were asked to perform pair-wise comparison between degraded and reference images. Hence, it is possible to treat the raw labeling scores as the difference scores D_{nmk}.

After standardizing the expert votes by Eq. (3), their correlation statistics with each IQM were computed in the form of SRCC and KRCC coefficients, defined as follows:

$$\text{SRCC} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{5}$$

where d_i is the difference between the i-th image's ranks in the objective and the subjective ratings and n is the number of observations.

$$\text{KRCC} = \frac{2}{n(n-1)} \sum_{i<j} \text{sign}(x_i - x_j)\ \text{sign}(y_i - y_j)\,, \tag{6}$$

where $(x_1, y_1), \ldots, (x_n, y_n)$ are the observations: the objective and the subjective score pairs.

We use SRCC as the main measure of an IQM performance, due to the non-linear relationship between the subjective and the objective scores.[4]

[4]The non-linear relationship is evident in Fig. 4 below.

The sizes of each batch of data are described in Fig. 2 (right).

A non-linear regression was performed on the IQM scores according to the quality Q to fit the subjective votes:

$$Q(x) = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + \exp(\beta_2(x - \beta_3))} \right) + \beta_4 x + \beta_5, \quad (7)$$

where x are the original IQM scores and $\beta_1, \ldots, \beta_5$ are the fitting coefficients.

V. RESULTS

Figs. 3 and 4 and Table 1 summarize the correlation study between the radiologists' scores and the IQM values for the three proposed evaluation criteria. The figures also show the results for the natural image domain. Top 4 performers in each category are marked in bold. The best and the worst examples of the reconstructions, as judged by different metrics, are presented in Fig. 5, and the aggregate scores for the top-performing metrics in each application in Fig. 6.

VI. DISCUSSION

The visual inspection of the outputs of the models in Fig. 5 makes it evident how the top metrics are superior in reflecting the actual reconstruction quality over the conventional PSNR and SSIM. The latter are known to misjudge shifts of brightness or a blur, indicating high quality for the bad images, whereas the more advanced FR and DB IQMs correlate with the visual perception and the subjective scores ably. Henceforth, out of the 35 metrics considered, we only discuss the best ones, according to their rank in the correlation study (VSI, HaarPSI, DISTS, $\text{FID}_{\text{VGG16}}$) and the widely used PSNR and SSIM.

As the key observation in the first systematic study of the DB metrics, we affirm that the choice of the feature extractor plays a crucial role. In particular, the correlation scores show that the Inception-based features are almost always worse than those from VGG16 (except for the MSID metric). Moreover, we see that, despite having been designed for the evaluation of *realism* of generative models data, FID shows competitive SRCC scores, thus, becoming a new recommended metric for the MRI image assessment tasks.

The non-linear relationship between the subjective and the objective scores, seen in Fig. 4, portrays intricate behavior with evident dependence on the anatomy and the target task, as well as a clear clustering of the points, instrumental for selecting a proper metric in a particular application. Notable are the generally lower IQM correlation scores when the difficulty of the reconstruction routine increases (compare trends in the scan acceleration data to those in the more complex denoising and the motion correction models). Also, the evaluation values for the knee reconstruction are generically lower, which could be caused by the greater variety of anatomical structures present in the knee data, as well as the more strict pertinent medical evaluation criteria [33].

Fig. 6 aggregates the outcomes per each task, anatomy, and evaluation criteria studied in our work, with the relation between the subjective and the objective scores highlighting the differences in the average performance of the top metrics. Notably, these selected IQMs have the highest correlation with expert judgment in the scan acceleration task. However, all metrics equally struggle reflecting the opinion of the radiologists in denoising and, sometimes, in motion correction tasks, especially on brain data. We also observe that some metrics perform consistently in terms of all three evaluation criteria and all tasks for given anatomy. For instance, GMSD and DISTS, despite not being of the highest SRCC rank overall, still show consistently high correlation scores on knee data, which proffers both of them as universal choices for the IQA in orthopedic applications. On the other hand, HaarPSI consistently rates the highest for both anatomies in the scan acceleration task, an instrumental fact to know when a single machine is used to scan various body parts or when the pertinent cross-anatomy inference [91] is performed.

TABLE 1. SRCC values of all 35 metrics on Natural and MRI Data. Top 4 performers in all categories are marked in bold. * denotes values taken directly from [90].

	Natural Images		MRI Data		
	TID2013	KADID-10k	Artefacts	CNR	SNR
PSNR	0.69	0.68	0.52	0.44	0.55
SSIM [2]	0.55	0.63	0.37	0.32	0.39
MS-SSIM [50]	0.80	0.80	0.58	0.48	0.59
IW-SSIM [51]	0.78	0.85	0.66	0.59	0.66
VIFp [35]	0.46	0.67	0.63	0.56	0.64
GMSD [52]	0.80	0.85	0.61	0.52	0.61
MS-GMSD [53]	0.81	0.85	0.60	0.51	0.61
FSIM [36]	0.80	0.84	0.66	0.61	0.65
VSI [54]	**0.89**	**0.88**	**0.72**	**0.64**	**0.70**
MDSI [55]	**0.89**	**0.89**	0.58	0.51	0.58
HaarPSI [56]	**0.87**	**0.88**	**0.72**	**0.63**	**0.72**
$\text{Content}_{\text{VGG16}}$ [57]	0.67	0.71	0.58	0.48	0.56
$\text{Style}_{\text{VGG16}}$ [57]	0.50	0.56	0.30	0.35	0.28
$\text{LPIPS}_{\text{VGG16}}$ [58]	0.67	0.78	0.61	0.52	0.63
DISTS [59]	0.71	0.81	**0.76**	**0.70**	**0.77**
PieAPP [60]	0.84	**0.87**	0.66	0.58	0.66
DSS [61]	0.79	0.86	0.56	0.47	0.57
No-reference metrics					
BRISQUE [4]	0.20	0.20	0.30	0.37	0.26
PaQ-2-PiQ [62]	**0.86***	0.84*	0.33	0.26	0.35
MetaIQA [63]	**0.86***	0.76*	0.22	0.20	0.24
Distribution-based metrics					
$\text{KID}_{\text{InceptionV3}}$ [7]	0.42	0.63	0.11	0.11	0.11
$\text{FID}_{\text{InceptionV3}}$ [6]	0.67	0.66	0.18	0.14	0.19
$\text{GS}_{\text{InceptionV3}}$ [64]	0.37	0.37	0.04	0.03	0.02
$\text{IS}_{\text{InceptionV3}}$ [5]	0.26	0.25	0.04	0.01	0.02
$\text{MSID}_{\text{InceptionV3}}$ [8]	0.21	0.32	0.01	0.07	0.07
$\text{KID}_{\text{VGG16}}$ [7]	0.70	0.71	0.55	0.47	0.54
$\text{FID}_{\text{VGG16}}$ [6]	0.67	0.66	**0.67**	**0.62**	**0.67**
GS_{VGG16} [64]	0.47	0.45	0.30	0.24	0.28
IS_{VGG16} [5]	0.64	0.65	0.30	0.31	0.28
$\text{MSID}_{\text{VGG16}}$ [8]	0.69	0.64	0.13	0.12	0.07
$\text{KID}_{\text{VGG19}}$ [7]	0.54	0.59	0.53	0.45	0.52
$\text{FID}_{\text{VGG19}}$ [6]	0.68	0.75	0.64	0.58	0.62
GS_{VGG19} [64]	0.35	0.41	0.27	0.21	0.25
IS_{VGG19} [5]	0.28	0.33	0.30	0.29	0.28
$\text{MSID}_{\text{VGG19}}$ [8]	0.11	0.13	0.06	0.02	0.05

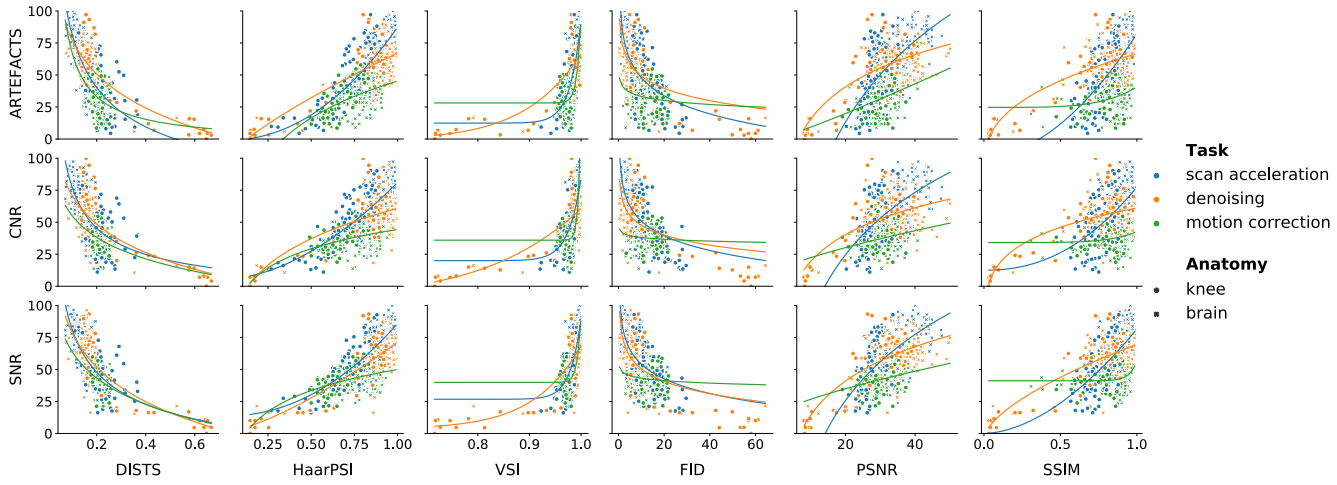

FIGURE 4. Relationship between processed subjective scores and IQM values for 3 evaluation criteria, 3 target tasks, and 2 anatomies (600 annotated image pairs in total). The solid lines are fits, plotted using the non-linear regression (7) on the subsets of images split by the tasks. The top 4 metrics (along with PSNR and SSIM, as the most commonplace) are shown in the decreasing order left to right, using SRCC to gauge the performance.

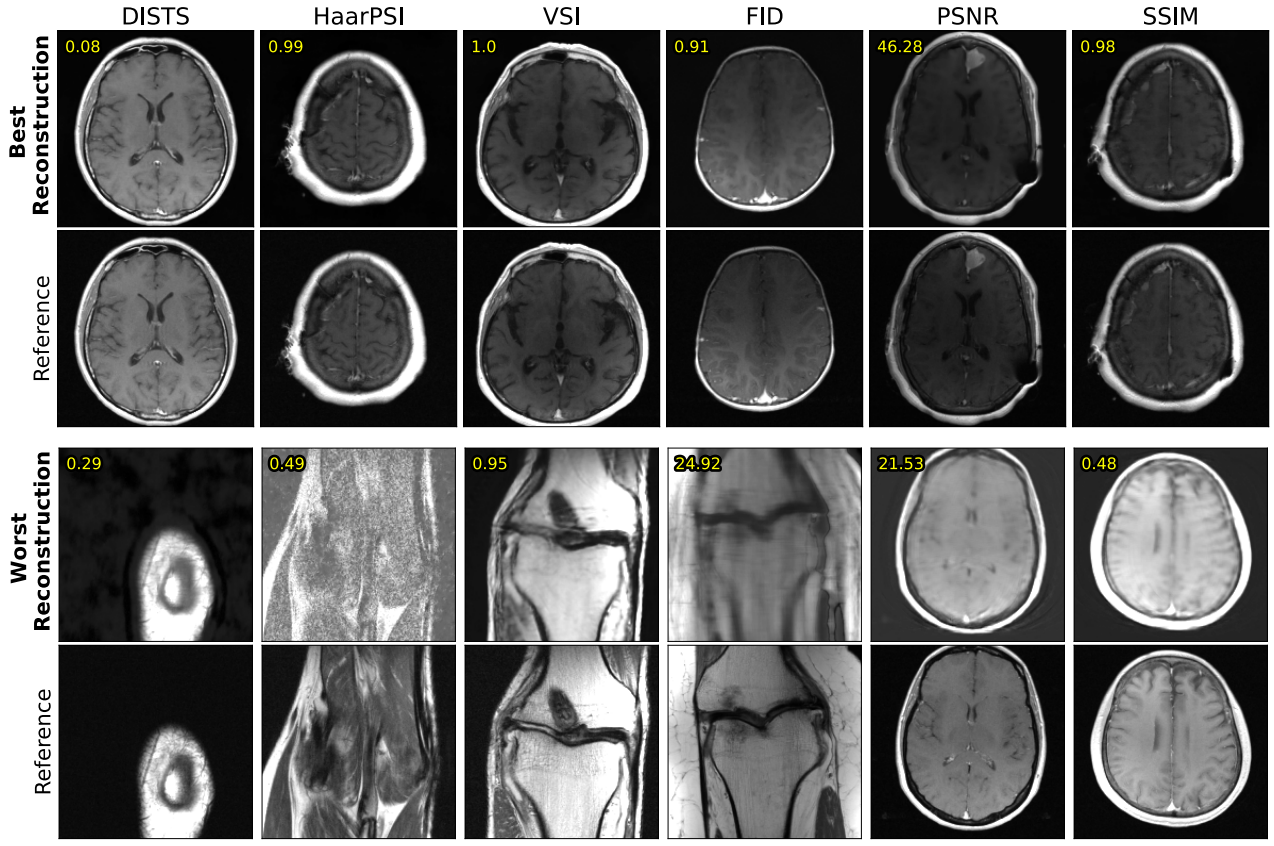

FIGURE 5. The best and the worst reconstruction-reference pairs according to different metrics (their values are shown in yellow). Note how the top 4 metrics (first four columns) reflect the actual reconstruction quality better than PSNR and SSIM (which are prone to misjudging a simple shift of brightness or a blur). The brightness is adjusted for viewer's convenience.

A. NATURAL vs. MRI IMAGES

A frequent IQA-related question is how generalizable are the performance benchmarks across different datasets and image domains. To study that, we analyzed the applicability of all 35 IQMs considered herein both in the MRI and the natural image (NI) domains (Table 1). For the latter, the popular TID2013 [87] and KADID-10k [88] datasets of NIs were used. Fig. 3 illustrates the effect of the shift between

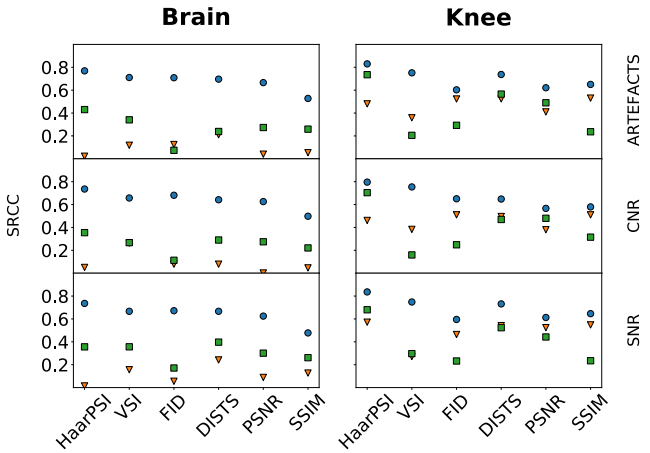

FIGURE 6. Aggregate relationship between the objective and the subjective scores for 3 evaluation criteria (rows), 2 anatomies (columns), and 3 tasks: scan acceleration (●), denoising (▼), and motion correction (■). The IQMs are ordered by decreasing average SRCC for the artefacts criterion on the brain data. This order is kept throughout all results for consistency. Note the tendency of the metrics to perform poorly in some task-anatomy combinations, *e.g.*, in denoising the brain data.

FIGURE 7. Correlation between the subjective scores in labeling and re-labeling sessions on the same data, with each column/color corresponding to an individual radiologist. This plot shows scoring self-consistency of the experts and the average time spent labeling one pair of images. Apparently, the time spent on labeling is not the major factor affecting the self-consistency of experienced radiologists.

the NI and the MRI domains, featuring an expected drop of the correlation values for most metrics.[5] However, the domain shift affects the ranks of the IQMs differently. Some top NI metrics, such as MDSI and MS-GMSD, naturally take lower standings in the MRI domain; however, others, such as HaarPSI and VSI, remain well-correlated with the radiologists' perception of quality. Further examples of IQMs robust to the domain shift are DISTS and $\text{FID}_{\text{VGG16}}$.

B. LABELING DISCREPANCIES AND SELF-CONSISTENCY STUDY

Another IQA-related question encountered in survey-based studies is the trustworthiness of the votes themselves. Given

[5]Not surprising, given these IQMs were designed for NI in the first place.

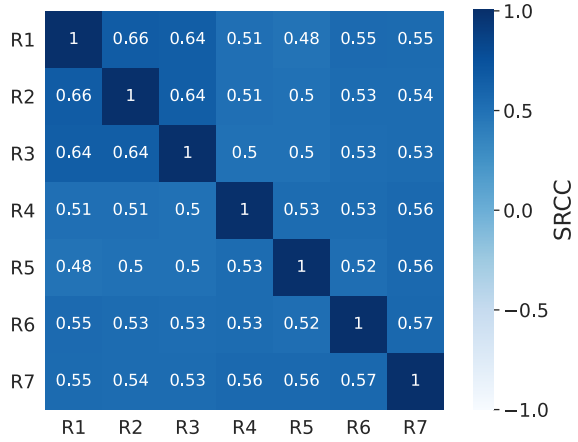

FIGURE 8. Pair-wise Spearman's rank correlation coefficient between *z*-scores from seven radiologists participating in the survey. According to [92], this pattern corresponds to a strong agreement between the experts.

that only reputable radiologists were engaged in our labeling routine, we have no grounds for doubting their annotations as far as the domain knowledge is concerned. Therefore, feasible discrepancies among their votes can be assumed to originate either from such factors as the study design, its duration, and fatigue, or from a previous experience which sometimes forms *a posteriori* intuition and, allegedly, influences the experts to make decisions different from the others.

While the latter is too subjective and difficult to regulate, the former could be controlled. We put effort to simplify the user experience and allowed the radiologists to approach the labeling assignment in batches at their own pace (see Appendix C). The average lead time spent labeling a pair of images,[6] an arguable indicator of the scrupulousness of an annotator, is plotted in Fig. 7, where we also summarize the results of the self-consistency study. The study reports Weighted Cohen's Kappa scores, computed between the votes provided in the main and in the additional re-labeling experiments on the same data. Interestingly, there is no significant correlation between self-consistency and the labeling time, placing other factors mentioned above, such as individual experience, at the forefront.

We also opted for evaluating the agreement between the radiologists' opinions by assessing the monotonic correlation between the *z*-scores computed earlier, which should account for the individual scoring preferences. The SRCC correlation values, shown in Fig. 8, never drop below 0.50, with a mean of 0.55 and a median of 0.53 (corresponds to *strong* relationship between variables).

In Fig. 7, the Weighted Cohen's Kappa values correspond to *moderate to substantial* consistency of scoring (according to [93]). And, according to [92], the SRCC range in Fig. 8 corresponds to a *strong* agreement. Given the sufficiently

[6]We discarded 5% of the shortest and the longest lead times to account for erroneous clicks and breaks between the labeling sessions.

TABLE 2. List of Abbreviations.

Abbreviation	Meaning
BRISQUE [4]	Blind/Referenceless Image Spatial Quality Evaluator
CNR	Contrast-to-Noise Ratio
DB	Distribution-based
DISTS [59]	Deep Image Structure and Texture Similarity
DSS [61]	Discrete Cosine Transform Subband Similarity
FID [6]	Frechet Inception Distance
FLAIR	Fluid-attenuated Inversion Recovery
FR	Full-reference
FSIM [36]	Feature Similarity Index Measure
GAN	Generative Adversarial Network
GMSD [52]	Gradient Magnitude Similarity Deviation
GS [64]	Geometry Score
HaarPSI [56]	Haar Perceptual Similarity Index
IQ	Image Quality
IQA	Image Quality Assessment
IQM	Image Quality Metric
IS [5]	Inception Score
IW-SSIM [51]	Information Content Weighted Structural Similarity
KID [7]	Kernel Inception Distance
KRCC	Kendall Rank Correlation Coefficient
LPIPS [58]	Learned Perceptual Image Patch Similarity
MDSI [55]	Mean Deviation Similarity Index
MRI	Magnetic Resonance Imaging
MS-GMSD [53]	Multi-Scale Gradient Magnitude Similarity Deviation
MS-SSIM [50]	Multi-Scale Structural Similarity
MSID [8]	Multi-Scale Intrinsic Distance
MetaIQA [63]	Meta Image Quality Assessment
NI	Natural Image
NR	No-reference
PD	Proton Density Weighting without Fat Suppression
PDFS	Proton Density Weighting with Fat Suppression
PIQ	PyTorch Image Quality
PSNR	Peak Signal-to-Noise Ratio
PaQ-2-PiQ [62]	Patches to Pictures
PieAPP [60]	Perceptual Image-Error Assessment
SNR	Signal-to-Noise Ratio
SRCC	Spearman's Rank Correlation Coefficient
SSIM [2]	Structural Similarity
T1-POST	T1 acquisition, post-contrast
T1-PRE	T1 acquisition, before contrast
VIFp [35]	Visual Information Fidelity
VSI [54]	Visual Saliency-induced Index

trustworthy labeling, the spread of the correlation scores for the modern IQA metrics in Fig. 6, and the non-trivial correlation patterns in Fig. 4, one can conclude that *the optimal* MRI metric is yet to be devised.

Besides a blunt umbrella metric aggregating the top-performing predictions (*e.g.*, those of VSI, HaarPSI, DISTS, and $\text{FID}_{\text{VGG16}}$), the future effort should be dedicated to additional forays into modeling *MRI-specific perception* of the radiologists and to *interpreting* their assessment using formalized rules taken from the medical textbooks. Such interpretatable metrics will be especially in demand, given the recent appearance of the MRI sampling approaches aimed towards optimizing downstream tasks [94], including the recently annotated FastMRI dataset [95]. Another line of future work could be 'borrowed' from the NI domain, where the abundance of data has led to the emergence of several NR IQMs. Although, in our study, all such metrics (classic BRISQUE [4] and the more recent PaQ-2-PiQ [62] and MetaIQA [63]) showed equally mediocre performance compared to the other IQMs, we believe their value in the MRI domain is bound to improve with the growth of available data.

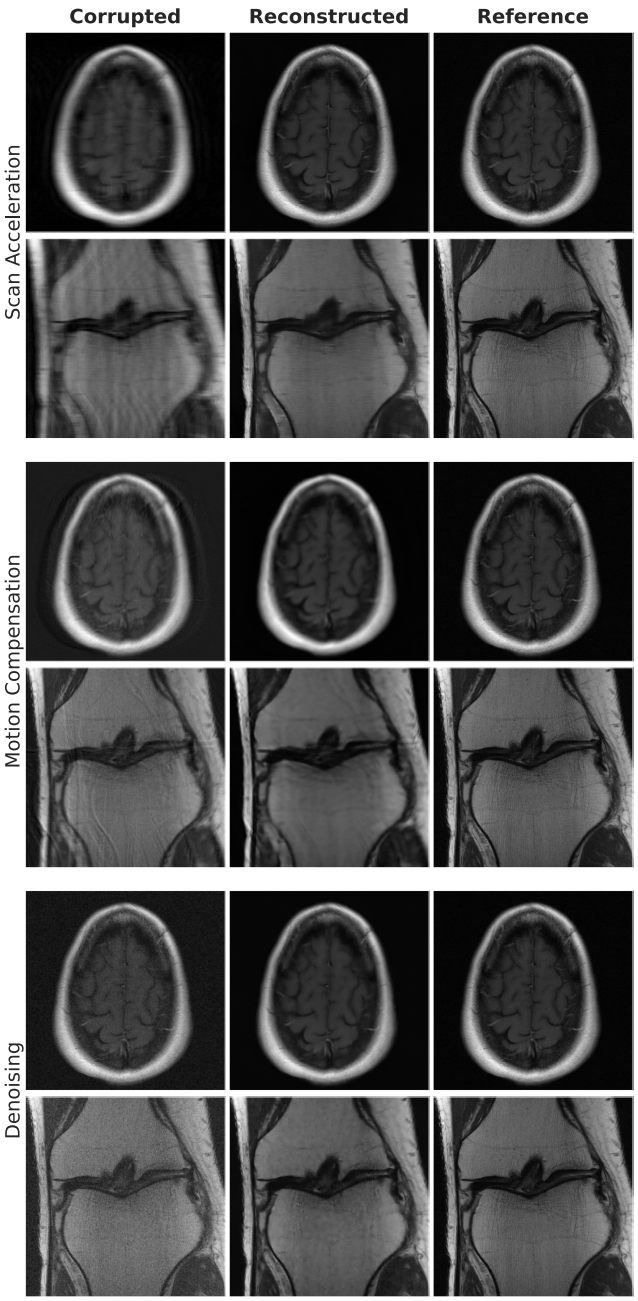

FIGURE 9. Examples of corrupted images used as inputs to the reconstruction models (left column), the reconstruction results (middle column), and the artefact-free reference images (right column). Examples with medium strength corruptions are displayed to showcase possible imperfect reconstruction results.

VII. CONCLUSION

This manuscript reports the most extensive study of the image quality metrics for Magnetic Resonance Imaging to date, evaluating 35 modern metrics and using 14,700 subjective votes from experienced radiologists.

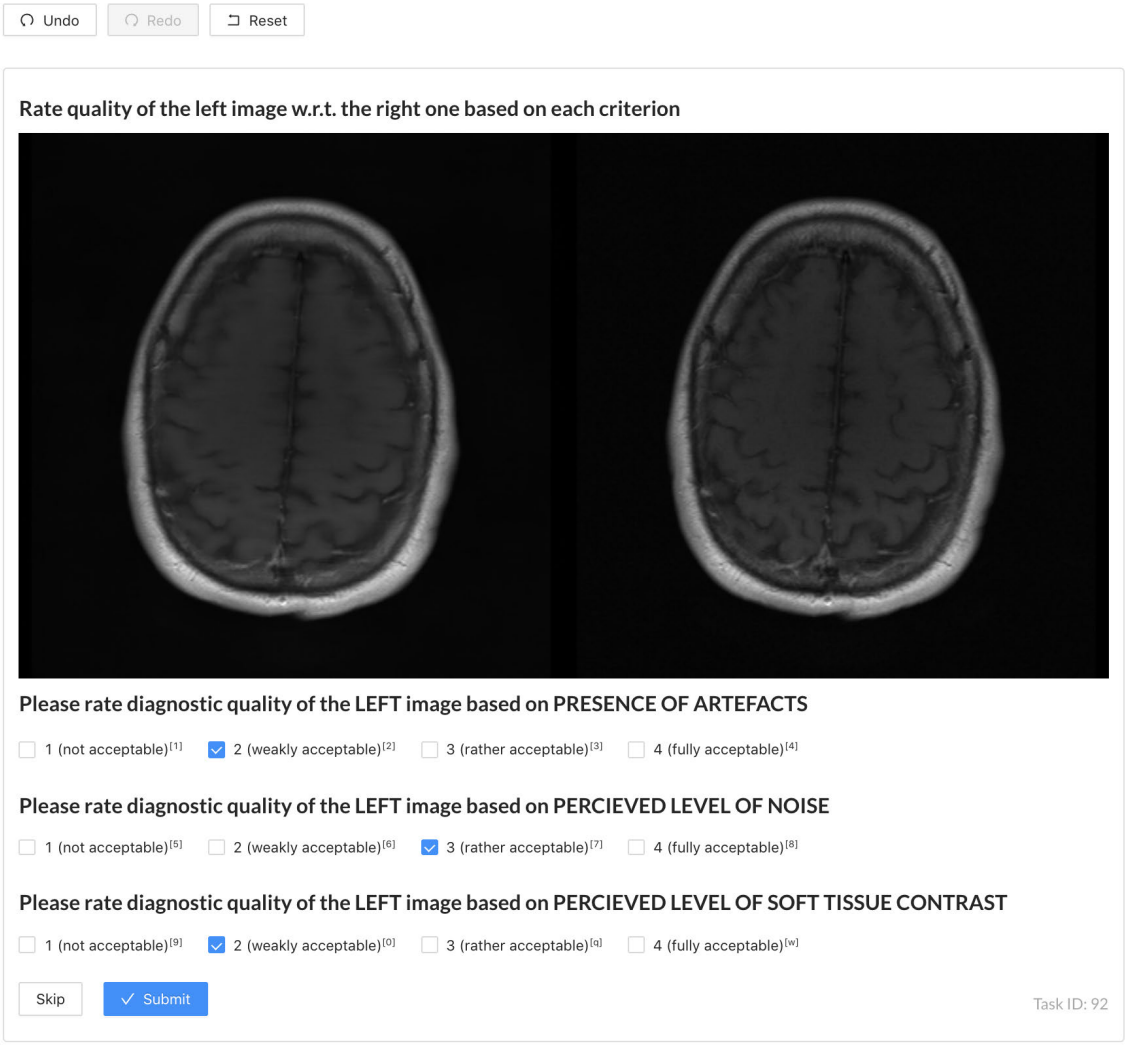

FIGURE 10. Web interface of the Label Studio software released to the expert radiologists to perform the labeling. The participants selected their answers using the proposed scale from 1 to 4, rating the images based on each proposed IQA criteria.

The applicability of full-reference, no-reference, and distribution-based metrics is discussed from the standpoint of MRI-specific image reconstruction tasks (scan acceleration, denoising, and motion correction). Unlike previous IQA studies analyzing IQMs with manual distortions, we use the outputs of neural network models trained to perform these particular tasks, enabling a realistic evaluation. Different from the natural images, the MRI scans are proposed to be assessed according to the most diagnostically influential criteria for the quality of MRI scans: signal-to-noise ratio, contrast-to-noise ratio, and the presence of artefacts.

The top performers – DISTS, HaarPSI, VSI, and $\text{FID}_{\text{VGG16}}$ – are found to be efficient across three proposed quality criteria, for all considered anatomies and the target tasks.

APPENDIX A
ABBREVIATIONS

A list of abbreviations is provided in Table 2.

APPENDIX B
RECONSTRUCTION EXAMPLES

During the labeling experiment, experts were asked to label pairs of images. Each pair contained a low-quality image placed side-by-side with a corresponding high-quality reference. Each low-quality image was obtained by, first, corrupting the corresponding reference and, then, by reconstructing it with the models trained to solve one of the tasks described in the main text (scan acceleration, motion compensation, or denoising). Fig. 9 showcases typical pairs of images used in the experiment. The following corruption parameters were used to generate the images: acceleration factor of 4, motion amplification factor of 0.6, noise amplification factor of 2. These parameters correspond to *medium* strength corruptions, showcasing possible imperfect reconstruction results.

APPENDIX C
LABELING USER INTERFACE

During the labeling experiment, the participants were asked to score pairs of reconstructed-reference images presented to

them side-by-side in a web interface of the Label Studio [89]. The web interface is shown in Fig. 10.

The labeling was done using three main IQ criteria: the presence of artefacts, the perceived level of noise, and the perceived level of soft-tissue contrast. The participants were able to select their answers using the mouse pointer or some keys on the keyboard. During the quality assessment process, the participants were able to zoom images, re-label previously labeled examples, pause and divide their evaluation session into as many labeling rounds as they wished. All labeling results were continuously saved on a remote server to eliminate the possibility of data loss. After the complete labeling process, the participants were offered the last chance to fix the scoring of the borderline examples.

ACKNOWLEDGMENT

The authors acknowledge the effort of radiologists from the Philips Clinical Application Team (PD CEER) for their help with data labeling and thank the supporters of their GitHub Project (https://github.com/photosynthesis-team/piq/), where each metric was independently implemented and tested.

They also declare no conflict of interest and no personal bias towards particular image quality metrics.

REFERENCES

[1] Z. Wang, "Applications of objective image quality assessment methods [applications corner]," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 137–142, Nov. 2011.

[2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[3] Z. Wang and A. C. Bovik, "Modern image quality assessment," *Synth. Lect. Image, Video, Multimedia Proc.*, vol. 2, no. 1, pp. 1–156, 2006.

[4] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[5] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Proc. Adv. Neural Inform. Process. Syst.*, 2016, pp. 2234–2242.

[6] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Adv. Neural Inform. Process. Syst.*, pp. 6626–6637, 2017.

[7] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in *Proc. ICLR*, 2018, pp. 1–36.

[8] A. Tsitsulin, M. Munkhoeva, D. Mottin, P. Karras, A. Bronstein, I. Oseledets, and E. Müller, "The shape of data: Intrinsic distance for data distributions," in *Proc. ICLR*, 2020, pp. 1–20.

[9] S. Athar and Z. Wang, "A comprehensive performance evaluation of image quality assessment algorithms," *IEEE Access*, vol. 7, pp. 140030–140070, 2019.

[10] R. A. Manap and L. Shao, "Non-distortion-specific no-reference image quality assessment: A survey," *Inf. Sci.*, vol. 301, p. 141, 2015.

[11] A. Mason, J. Rioux, S. E. Clarke, A. Costa, M. Schmidt, V. Keough, T. Huynh, and S. Beyea, "Comparison of objective image quality metrics to expert radiologists' scoring of diagnostic quality of MR images," *IEEE Trans. Med. Imag.*, vol. 39, no. 4, pp. 1064–1072, Apr. 2020.

[12] M. B. Williams, E. A. Krupinski, K. J. Strauss, W. K. Breeden III, M. S. Rzeszotarski, K. Applegate, M. Wyatt, S. Bjork, and J. A. Seibert, "Digital radiography image quality: Image acquisition," *J. Amer. College Radiol.*, vol. 4, no. 6, pp. 371–388, 2007.

[13] A. J. Ahumada, "Computational image quality metrics: A review," *SID Dig.*, vol. 24, no. Jan. 1993, pp. 305–308, 1993.

[14] J. Kumar, P. Ye, and D. Doermann, "A dataset for quality assessment of camera captured document images," in *Camera-Based Document Analysis and Recognition* (Lecture Notes in Computer Science), vol. 8357. Cham, Switzerland: Springer, 2013, pp. 113–125.

[15] H. Yang, Y. Fang, and W. Lin, "Perceptual quality assessment of screen content images," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4408–4421, Nov. 2015.

[16] E. Wang, B. Yang, and L. Pang, "Superpixel-based structural similarity metric for image fusion quality evaluation," *Sens. Imag.*, vol. 22, no. 1, pp. 1–25, Dec. 2021.

[17] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3674–3683.

[18] O. Ieremeiev, V. Lukin, K. Okarma, and K. Egiazarian, "Full-reference quality metric based on neural network to assess the visual quality of remote sensing images," *Remote Sens.*, vol. 12, no. 15, p. 2349, Jul. 2020.

[19] A. H. Baker, D. M. Hammerling, and T. L. Turton, "Evaluating image quality measures to assess the impact of lossy data compression applied to climate simulation data," *Comput. Graph. Forum*, vol. 38, no. 3, pp. 517–528, Jun. 2019.

[20] N. Avadhanam and V. R. Algazi, "Evaluation of a human-vision-system-based image fidelity metric for image compression," in *Proc. SPIE*, vol. 3808, pp. 569–579, Oct. 1999.

[21] E. Allen, S. Triantaphillidou, and R. Jacobson, "Image quality of JPEG vs JPEG 2000 image compression schemes, Part 1: Psychophysical measurements," *IS&T J. Imag. Sci. Technol*, vol. 51, no. 3, p. 248, 2007.

[22] S. Triantaphillidou, E. Allen, and R. Jacobson, "Image quality of JPEG vs JPEG 2000 image compression schemes, Part 2: Scene analysis," *IS&T J. Imag. Sci. Technol*, vol. 51, no. 3, pp. 259–270, 2007.

[23] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Comparison of full-reference image quality models for optimization of image processing systems," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1258–1281, 2021.

[24] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[25] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, pp. 011006-1–011006-21, Jan. 2010.

[26] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "A comprehensive evaluation of full reference image quality assessment algorithms," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 1477–1480.

[27] P. Mohammadi, A. Ebrahimi-Moghadam, and S. Shirani, "Subjective and objective quality assessment of image: A survey," 2014, *arXiv:1406.7799*.

[28] M. Pedersen, "Evaluation of 60 full-reference image quality metrics on the CID: IQ," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 1588–1592.

[29] C. Cavaro-Menard, L. Zhang, and P. Le Callet, "Diagnostic quality assessment of medical images: Challenges and trends," in *Proc. 2nd Eur. Workshop Vis. Inf. Process. (EUVIP)*, Jul. 2010, pp. 277–284.

[30] R. Li, G. Dai, Z. Wang, S. Yu, and Y. Xie, "Using signal-to-noise ratio to connect the quality assessment of natural and medical images," in *Proc. SPIE*, vol. 10806, Aug. 2018, Art. no. 108064Q.

[31] H. Rajagopal, "Subjective versus objective assessment for magnetic resonance images," in *Proc. 17th Int. Conf Communic Inf. Technol. Eng.*, vol. 9, 2015, pp. 1–6.

[32] L. S. Chow, H. Rajagopal, and R. Paramesran, "Correlation between subjective and objective assessment of magnetic resonance (MR) images," *Magn. Reson. Imag.*, vol. 34, no. 6, pp. 820–831, 2016.

[33] A. Keshavan, J. D. Yeatman, and A. Rokem, "Combining citizen science and deep learning to amplify expertise in neuroimaging," *Frontiers Neuroinform.*, vol. 13, p. 29, May 2019.

[34] G. P. Renieblas, A. T. Nogués, A. M. González, N. Gómez-Leon, and E. G. del Castillo, "Structural similarity index family for image quality assessment in radiological images," *J. Med. Imag.*, vol. 4, no. 3, Jul. 2017, Art. no. 035501.

[35] H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessment," in *Proc. 1st Int. Workshop VPQM Consum. Electron.*, vol. 7, 2005, p. 2.

[36] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[37] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 636–650, Apr. 2000.

[38] L. S. Chow and R. Paramesran, "Review of medical image quality assessment," *Biomed. Signal Process. Control*, vol. 27, no. 1, pp. 145–154, 2016.

[39] J. Jang, K. Bang, H. Jang, and D. Hwang, "Quality evaluation of no-reference MR images using multidirectional filters and image statistics," *Magn. Reson. Med.*, vol. 80, no. 3, pp. 914–924, Sep. 2018.

[40] M. Oszust, A. Piórkowski, and R. Obuchowicz, "No-reference image quality assessment of magnetic resonance images with high-boost filtering and local features," *Magn. Reson. Med.*, vol. 84, no. 3, pp. 1648–1660, Sep. 2020.

[41] R. Obuchowicz, M. Oszust, M. Bielecka, A. Bielecki, and A. Piórkowski, "Magnetic resonance image quality assessment by using non-maximum suppression and entropy analysis," *Entropy*, vol. 22, no. 2, p. 220, Feb. 2020.

[42] I. Stpień, R. Obuchowicz, A. Piórkowski, and M. Oszust, "Fusion of deep convolutional neural networks for no-reference magnetic resonance image quality assessment," *Sensors*, vol. 21, no. 4, p. 1043, Feb. 2021.

[43] V. R. Simi, D. R. Edla, and J. Joseph, "A no-reference metric to assess quality of denoising for magnetic resonance images," *Biomed. Signal Process. Control*, vol. 70, Sep. 2021, Art. no. 102962.

[44] T. Küstner, S. Gatidis, A. Liebgott, M. Schwartz, L. Mauch, P. Martirosian, H. Schmidt, N. F. Schwenzer, K. Nikolaou, F. Bamberg, and B. Yang, "A machine-learning framework for automatic reference-free quality assessment in MRI," *Magn. Reson. Imag.*, vol. 53, pp. 134–147, Nov. 2018.

[45] B. Mortamet, M. A. Bernstein, C. R. Jack Jr., J. L. Gunter, C. Ward, P. J. Britson, R. Meuli, J.-P. Thiran, and G. Krueger, "Automatic quality assessment in structural brain magnetic resonance imaging," *Magn. Reson. Medicine: An Off. J. Int. Soc. Magn. Reson. Med.*, vol. 62, no. 2, pp. 365–372, 2009.

[46] R. A. Pizarro, X. Cheng, A. Barnett, H. Lemaitre, B. A. Verchinski, A. L. Goldman, E. Xiao, Q. Luo, K. F. Berman, J. H. Callicott, D. R. Weinberger, and V. S. Mattay, "Automated quality assessment of structural magnetic resonance brain images based on a supervised machine learning algorithm," *Frontiers Neuroinform.*, vol. 10, p. 52, Dec. 2016.

[47] M. S. Treder, R. Codrai, and K. A. Tsvetanov, "Quality assessment of anatomical MRI images from generative adversarial networks: Human assessment and image quality metrics," *J. Neurosci. Methods*, vol. 374, May 2022, Art. no. 109579.

[48] S. Masoudi, S. Harmon, S. Mehralivand, N. Lay, U. Bagci, B. J. Wood, P. A. Pinto, P. Choyke, and B. Turkbey, "No-reference image quality assessment of T2-weighted magnetic resonance images in prostate cancer patients," in *Proc. IEEE 18th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2021, pp. 1201–1205.

[49] I. Stpień and M. Oszust, "A brief survey on no-reference image quality assessment methods for magnetic resonance images," *J. Imag.*, vol. 8, no. 6, p. 160, Jun. 2022.

[50] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, Jul. 2003, pp. 1398–1402.

[51] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2010.

[52] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2013.

[53] B. Zhang, P. V. Sander, and A. Bermak, "Gradient magnitude similarity deviation on multiple scales for color image quality assessment," in *Proc. IEEE ICASSP*, Mar. 2017, pp. 1253–1257.

[54] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270–4281, Aug. 2014.

[55] H. Z. Nafchi, A. Shahkolaei, R. Hedjam, and M. Cheriet, "Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator," *IEEE Access*, vol. 4, pp. 5579–5590, 2016.

[56] R. Reisenhofer, S. Bosse, G. Kutyniok, and T. Wiegand, "A Haar wavelet-based perceptual similarity index for image quality assessment," *Signal Process., Image Commun.*, vol. 61, pp. 33–43, Feb. 2018.

[57] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*, 2016, pp. 694–711.

[58] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[59] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," 2020, *arXiv:2004.07728*.

[60] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "PieAPP: Perceptual image-error assessment through pairwise preference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1808–1817.

[61] A. Balanov, A. Schwartz, Y. Moshe, and N. Peleg, "Image quality assessment based on DCT subband similarity," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 2105–2109.

[62] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3572–3582.

[63] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "MetaIQA: Deep meta-learning for no-reference image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14131–14140.

[64] V. Khrulkov and I. Oseledets, "Geometry score: A method for comparing generative adversarial networks," in *Proc. ICML*, 2018, pp. 2626–2634.

[65] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[66] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.

[67] S. Kastryulin, D. Zakirov, and D. Prokopenko. (2019). *PyTorch Image Quality: Metrics and Measure for Image Quality Assessment*. [Online]. Available: https://github.com/photosynthesis-team/piq

[68] G. Jinjin, C. Haoming, C. Haoyu, Y. Xiaoxing, J. S. Ren, and D. Chao, "PIPAL: A large-scale image quality assessment dataset for perceptual image restoration," in *Computer Vision* (Lecture Notes in Computer Science), vol. 12356. Cham, Switzerland: Springer, 2020, pp. 633–651.

[69] J. Gu, H. Cai, C. Dong, J. S. Ren, R. Timofte, Y. Gong, S. Lao, S. Shi, J. Wang, S. Yang, and T. Wu, "NTIRE 2022 challenge on perceptual image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 951–967.

[70] M. Cheon, S.-J. Yoon, B. Kang, and J. Lee, "Perceptual image quality assessment with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 433–442.

[71] H. Guo, Y. Bin, Y. Hou, Q. Zhang, and H. Luo, "IQMA network: Image quality multi-scale assessment network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 443–452.

[72] S. Shi, Q. Bai, M. Cao, W. Xia, J. Wang, Y. Chen, and Y. Yang, "Region-adaptive deformable network for image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 324–333.

[73] F. Knoll, J. Zbontar, A. Sriram, M. J. Muckley, M. Bruno, A. Defazio, M. Parente, K. J. Geras, and J. Katsnelson, and H. Chandarana, "FastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning," *Radiol., Artif. Intell.*, vol. 2, no. 1, Jan. 2020, Art. no. e190007.

[74] J. Zbontar, F. Knoll, A. Sriram, T. Murrell, Z. Huang, M. J. Muckley, A. Defazio, R. Stern, P. Johnson, and M. Bruno, "FastMRI: An open dataset and benchmarks for accelerated MRI," 2019, *arXiv:1811.08839*.

[75] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, and L. Antiga, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[76] A. Effland, E. Kobler, K. Kunisch, and T. Pock, "An optimal control approach to early stopping variational methods for image restoration," 2019, *arXiv:1907.08488*.

[77] N. Pezzotti, S. Yousefi, M. S. Elmahdy, J. H. F. Van Gemert, C. Schuelke, M. Doneva, T. Nielsen, S. Kastryulin, B. P. F. Lelieveldt, and M. J. P. Van Osch, "An adaptive intelligence algorithm for undersampled knee MRI reconstruction," *IEEE Access*, vol. 8, p. 204825, 2020.

[78] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, Jan. 2009.

[79] T. Tieleman and G. Hinton, "Lecture 6.5-RMSProp, COURSERA: Neural networks for ML," Univ. Toronto, Tech. Rep., 2012.

[80] K. Sommer, A. Saalbach, T. Brosch, C. Hall, N. M. Cross, and J. B. Andre, "Correction of motion artifacts using a multiscale fully convolutional neural network," *Amer. J. Neuroradiol.*, vol. 41, no. 3, pp. 416–423, Mar. 2020.

[81] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention* (Lecture Notes in Computer Science), vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[82] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR Conf.*, 2015, pp. 1–15.
[83] J. Sijbers, D. Poot, A. J. den Dekker, and W. Pintjens, "Automatic estimation of the noise variance from the histogram of a magnetic resonance image," *Phys. Med. Biol.*, vol. 52, no. 5, p. 1335, 2007.
[84] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2015.
[85] H. H. Barrett, J. Yao, J. P. Rolland, and K. J. Myers, "Model observers for assessment of image quality," *Proc. Nat. Acad. Sci. USA*, vol. 90, pp. 9758–9765, Nov. 1993.
[86] X. He and S. Park, "Model observers in medical imaging research," *Theranostics*, vol. 3, no. 10, p. 774, 2013.
[87] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, and F. Battisti, "Image database TID2013: Peculiarities, results and perspectives," *Signal Process., Image Commun.*, vol. 30, pp. 57–77, Jan. 2015.
[88] H. Lin, V. Hosu, and D. Saupe, "KADID-10k: A large-scale artificially distorted IQA database," in *Proc. 11th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–3.
[89] M. Tkachenko, M. Malyuk, N. Shevchenko, A. Holmanyuk, and N. Liubimov. (2020). *Label Studio: Data Labeling Software*. Accessed: Mar. 3, 2022. [Online]. Available: https://github.com/heartexlabs/label-studio
[90] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," 2021, *arXiv:2108.06858*.
[91] A. Belov, J. Stadelmann, S. Kastryulin, and D. V. Dylov, "Towards ultrafast MRI via extreme k-space undersampling and superresolution," in *Medical Image Computing and Computer-Assisted Intervention* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2021, pp. 254–264.
[92] C. P. Dancey and J. Reidy, *Statistics Without Maths for Psychology*. London, U.K.: Pearson, 2007.
[93] A. Viera and J. Garrett, "Understanding interobserver agreement: The Kappa statistic," *Family Med.*, vol. 37, pp. 3–360, Jun. 2005.
[94] A. Razumov, O. Y. Rogov, and D. V. Dylov, "Optimal MRI undersampling patterns for ultimate benefit of medical vision tasks," 2021, *arXiv:2108.04914*.
[95] R. Zhao, B. Yaman, Y. Zhang, R. Stewart, A. Dixon, F. Knoll, Z. Huang Y. W. Lui, M. S. Hansen, and M. P. Lungren, "FastMRI+: Clinical pathology annotations for knee and brain fully sampled multi-coil MRI data," 2021, *arXiv:2109.03812*.

SERGEY KASTRYULIN received the M.S. degree in computer science from Southern Federal University, Rostov-on-Don, Russia, in 2019. He is currently pursuing the Ph.D. degree in artificial intelligence with the Skolkovo Institute of Science and Technology, Moscow, Russia. From 2019 to 2022, he was a Research Scientist with Philips Research, Russia. His research interests include deep learning and metrics for image quality assessment.

JAMIL ZAKIROV received the B.Sc. degree in applied mathematics and physics from the Moscow Institute of Physics and Technology (MIPT), in 2019, and the M.Sc. degree in data science from the Skolkovo Institute of Science and Technology (Skoltech), in 2021. In 2020, he was a Junior Research Scientist at Philips Research. He is currently an Independent Researcher. His research interests include image synthesis, image quality evaluation, and model compression using reduced bit precision.

NICOLA PEZZOTTI received the B.Sc. and M.Sc. degrees in computer science and engineering from the University of Brescia, Italy, in 2009 and 2011, respectively, and the Ph.D. degree (cum laude) from the Delft University of Technology, The Netherlands, in 2018. He was a Research and Development Engineer at Open Technology S.r.l., from 2011 to 2014; and a Visiting Scientist with INRIA Saclay, Paris, in 2017, and Google AI, Zürich, in 2018. He is currently a Senior Scientist with Philips Research, Eindhoven, The Netherlands, and an Assistant Professor with the Eindhoven University of Technology. His research interests include machine learning, visual analytics, explainable AI, optimization techniques, and software engineering. He was a recipient of several awards, including the IEEE VGTC Best Dissertation Award, the TU Delft Excellence in Research, and the Dirk Bartz Prize for visual computing in medicine.

DMITRY V. DYLOV (Member, IEEE) received the M.Sc. degree in applied physics and mathematics from the Moscow Institute of Physics and Technology, Moscow, Russia, in 2006, and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, USA, in 2010. He is currently an Associate Professor and the Head of the Computational Imaging Group, Skolkovo Institute of Science and Technology (Skoltech), Moscow. His research interests include computational imaging, computer/medical vision, and fundamental aspects of image formation.