UNIVERSITÉ DE GENÈVE

Archive ouverte UNIGE
https://archive-ouverte.unige.ch

Article scientifique | Article | 2020

Published version | Open Access

This is the published version of the publication, made available in accordance with the publisher's policy.

Deep Learning to Automate Reference-Free Image Quality Assessment of Whole-Heart MR Images

Piccini, Davide; Demesmaeker, Robin; Heerfordt, John; Yerly, Jérôme; Di Sopra, Lorenzo; Masci, Pier Giorgio; Schwitter, Juerg; Van De Ville, Dimitri; Richiardi, Jonas; Kober, Tobias; Stuber, Matthias

How to cite

PICCINI, Davide et al. Deep Learning to Automate Reference-Free Image Quality Assessment of Whole-Heart MR Images. In: Radiology: Artificial Intelligence, 2020, vol. 2, n° 3, p. e190123. doi: 10.1148/ryai.2020190123

This publication URL: https://archive-ouverte.unige.ch/unige:156308

Publication DOI: 10.1148/ryai.2020190123

© This document is protected by copyright. Please refer to copyright holder(s) for terms of use.

Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Deep Learning to Automate Reference-Free Image Quality Assessment of Whole-Heart MR Images

*Davide Piccini, PhD** • *Robin Demesmaeker, MSc** • *John Heerfordt, MSc* • *Jérôme Yerly, PhD* • *Lorenzo Di Sopra, MSc* • *Pier Giorgio Masci, MD* • *Juerg Schwitter, MD* • *Dimitri Van De Ville, PhD* • *Jonas Richiardi, PhD* • *Tobias Kober, PhD* • *Matthias Stuber, PhD*

From Advanced Clinical Imaging Technology, Siemens Healthcare, Lausanne, Switzerland (D.P., R.D., J.H., J.R., T.K.); Department of Diagnostic and Interventional Radiology, Lausanne University Hospital and University of Lausanne, Rue de Bugnon 46, BH 8.80, 1011 Lausanne, Switzerland (D.P., J.H., J.Y., L.D.S., J.R., T.K., M.S.); LTS5, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland (D.P., J.R., T.K.); Institute of Electrical Engineering, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland (R.D.); Institute of Bioengineering/Center for Neuroprosthetics, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland (R.D., D.V.D.V.); Center for Biomedical Imaging (CIBM), Lausanne, Switzerland (J.Y., M.S.); Division of Cardiology and Cardiac MR Center, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland (P.G.M., J.S.); and Department of Radiology and Medical Informatics, University Hospital of Geneva (HUG), Geneva, Switzerland (D.V.D.V.). Received July 19, 2019; revision requested September 16; revision received March 3, 2020; accepted March 11. **Address correspondence to** D.P. (e-mail: *piccinidavide@gmail.com*).

*D.P. and R.D. contributed equally to this work.

Work supported by Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung (143923, 173129).

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2020; 2(3):e190123 • https://doi.org/10.1148/ryai.2020190123 • Content codes: CA IN MR

Purpose: To develop and characterize an algorithm that mimics human expert visual assessment to quantitatively determine the quality of three-dimensional (3D) whole-heart MR images.

Materials and Methods: In this study, 3D whole-heart cardiac MRI scans from 424 participants (average age, 57 years ± 18 [standard deviation]; 66.5% men) were used to generate an image quality assessment algorithm. A deep convolutional neural network for image quality assessment (IQ-DCNN) was designed, trained, optimized, and cross-validated on a clinical database of 324 (training set) scans. On a separate test set (100 scans), two hypotheses were tested: *(a)* that the algorithm can assess image quality in concordance with human expert assessment as assessed by human-machine correlation and intra- and interobserver agreement and *(b)* that the IQ-DCNN algorithm may be used to monitor a compressed sensing reconstruction process where image quality progressively improves. Weighted κ values, agreement and disagreement counts, and Krippendorff α reliability coefficients were reported.

Results: Regression performance of the IQ-DCNN was within the range of human intra- and interobserver agreement and in very good agreement with the human expert (R^2 = 0.78, κ = 0.67). The image quality assessment during compressed sensing reconstruction correlated with the cost function at each iteration and was successfully applied to rank the results in very good agreement with the human expert.

Conclusion: The proposed IQ-DCNN was trained to mimic expert visual image quality assessment of 3D whole-heart MR images. The results from the IQ-DCNN were in good agreement with human expert reading, and the network was capable of automatically comparing different reconstructed volumes.

Supplemental material is available for this article.

© RSNA, 2020

Image quality assessment is essential for many radiology applications (1). There are various methods for determining image quality, including visual inspection by human experts and extraction of quantitative endpoints (2–5). However, to promote an automated workflow, it is required that more flexible and reproducible quantitative metrics are extracted from images. An automated workflow would facilitate the immediate assessment of image quality during the patient examination and may potentially allow for a timely rescan when needed. One of the main challenges for generating an automated image quality extraction algorithm relates to its capability of providing a meaningful correlation with the general visual perception of human experts (6), instead of only quantifying one specific feature of the image. In many cases, mathematical formulas, such as the spatial variation of the image intensity, gradient entropy, temporal total variation, or other metrics, lead to computational results that tend to indicate agreement with human expert perception (6). However, these formulas are often independent of the specific context of the images they are used to assess.

In cardiac MRI, aside from classic measures such as signal-to-noise ratio or contrast-to-noise ratio, several anatomy-specific metrics can be used to assess image quality, including coronary sharpness, visible vessel length (7), or myocardial border sharpness (8). Alternative measures for assessing image quality include the precision and accuracy of the functional analysis for cine imaging (9) or, more generically, one of many diagnostic quality scales (eg, Likert scale) (10). Only recently (11), deep learning has been identified as a way of implicitly learning image features that may inform about quality. Current examples in the literature mainly focus on natural image processing, where extensive databases are freely available, and the

This copy is for personal use only. To order printed copies, contact reprints@rsna.org

Abbreviations

DCNN = deep convolutional neural network, IQ-DCNN = DCNN for image quality assessment, 3D = three dimensional, 2D = two dimensional

Summary

An artificial intelligence–based algorithm can mimic expert visual image quality assessment and allows for fast and automated image quality grading of three-dimensional whole-heart MR images.

Key Points

- The proposed deep learning framework shows that it is possible to train a neural network to reproduce human expert image quality assessment on a predefined scale with a performance that approaches that of human expert interobserver agreement.
- The proposed deep learning framework can be used to compare image volumes that are reconstructed with different algorithms from the same acquisition and to select the image with the best quality.

meaning of image quality is straightforward. For example, the classic structural similarity index (12) was developed to assess two-dimensional (2D) image degradation under compression; and face-specific quality metrics have been evaluated on face image databases with multiple quality settings (13). Deep neural networks have already been successfully used to provide a quality grade on a predefined scale (14,15). However, few applications for medical imaging have been proposed thus far. Examples of image quality assessment algorithms include those used for classification of fundus images of the retina in two categories to assess in real time whether the image should be reacquired (16,17), automated artifact detection (binary classification) on MRI scans (18), classifying T2-weighted liver MR images as "diagnostic" or "nondiagnostic" (19), or transperineal US image quality (20). In cardiac MRI, Zhang et al used a convolutional neural network to detect missing apical and basal slices (21) and a generative adversarial network to identify complete left ventricular coverage (22).

In this work, we applied an alternative approach to set out to understand the intrinsic concept of "image quality" as assessed by human experts by using a predefined scale. The hypothesis for this study was that a deep convolutional neural network (DCNN) could be trained to reproduce the grading performance of an expert observer. The results obtained by the network would therefore be within the range of the human inter- and intraobserver variability. We tested our hypothesis in the specific scenario of whole-heart MRI acquisitions. By employing a DCNN for image quality assessment (IQ-DCNN), the proposed algorithm was first trained and then cross-validated using a patient database of three-dimensional (3D) MR images of the heart that were graded for image quality by human experts. The performance of the trained and validated IQ-DCNN was subsequently tested using patient data that were not included during the training and validation phases. The link between human and IQ-DCNN grading was then studied and expressed relative to expert intra- and interobserver agreement. Finally, we investigated the combination of this algorithm with an iterative compressed sensing image reconstruction process as an exemplary scenario where the IQ-DCNN would be used to assess quality differences in the same dataset under different conditions (eg, increasing number of iterations or different respiratory phases). An automated workflow that can reproduce expert image quality assessment could potentially facilitate rescan decisions in a timely manner and may be useful to optimize iterative reconstruction algorithms in terms of number of iterations and final image quality.

Table 1: Image Quality Scale

Grade	Description
0	Nondiagnostic
1	Marked blurring, limited diagnostic value
2	Moderate blurring, but diagnostic value
3	Mild blurring, good diagnostic value
4	Excellent diagnostic value

Note.—Image quality scale used for the human expert grading. Half grades were allowed when the expert reader was in doubt between two grades.

Materials and Methods

Patient Imaging Database

A clinical database consisting of 3D whole-heart MR images was used in this study. Written informed consent was obtained from all participants. All images were acquired between 2013 and 2016. This retrospective study was approved by the local ethics committee at the University Hospital of Lausanne. Scans from 424 randomly selected patients (one scan per patient) referred to general clinical cardiac MRI (average age, 57 years ± 18 [standard deviation], 66.5% men) were collected and fully anonymized with no specific inclusion criteria.

MRI Parameters

All scans were acquired with a 1.5-T clinical MR scanner (MAGNETOM Aera, Siemens Healthcare, Erlangen, Germany) using a prototype free-breathing and respiratory self-navigated electrocardiographically triggered 3D radial golden-angle balanced steady-state free precession sequence (23–26). The field of view was 210 × 210 × 210 mm^3, with 1-mm^3 isotropic spatial resolution. Typical sequence parameters were as follows: repetition time msec/echo time msec = 3.1/1.56, radiofrequency excitation angle = 90°–115°, and receiver bandwidth = 900 Hz/pixel. A total of approximately 15 000 radial readouts were acquired, equally divided over about 400–500 heartbeats, depending on the individual heart rate of each patient and with an overall sampling ratio of 20% of the Nyquist limit. The trigger delay to the most quiescent middiastolic or late systolic cardiac phase was set by the cardiologist operating the MRI unit, using a midventricular short-axis cine image series acquired before the whole-heart scan.

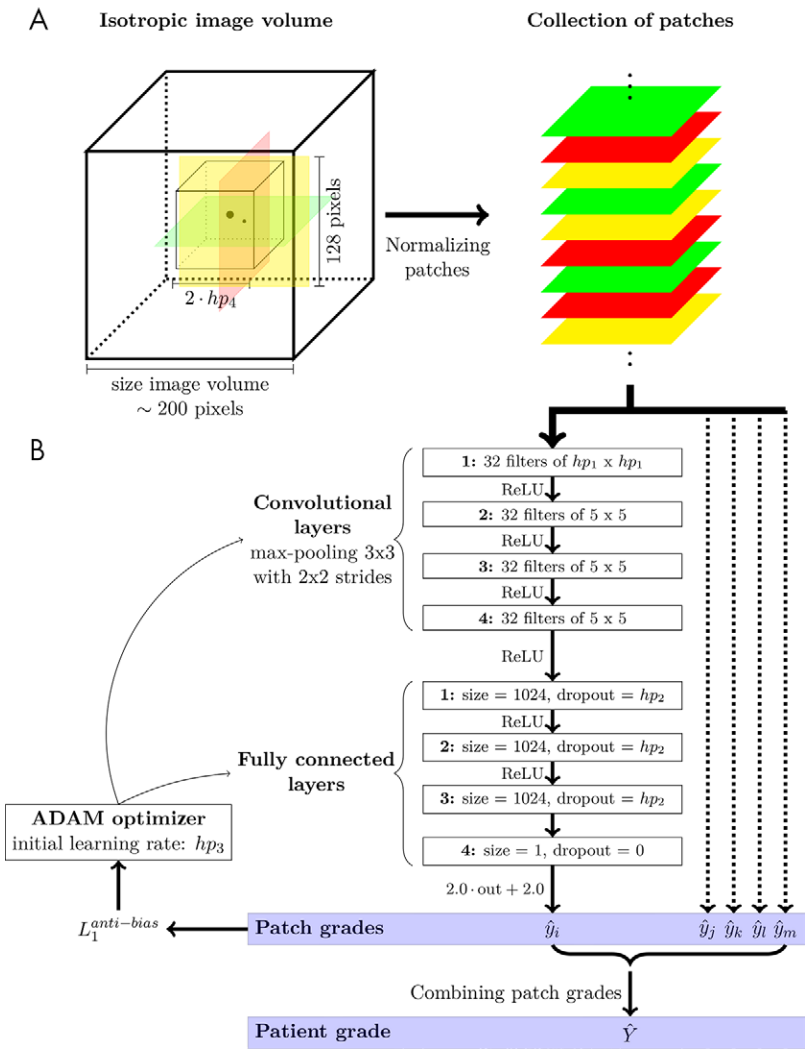

Figure 1: Visual representation of the deep neural network used in this study. *A*, The network input is a two-dimensional (2D) patch extracted from the three-dimensional volume. Batches of 2D patches in the three orthogonal orientations are used for training. *B*, Four convolutional layers are followed by three fully connected layers. The final regression layer provides a grade for each patch. The mean value of patch grades for one volume corresponds to the patient grade. The hyperparameters hp_1 to hp_4 were optimized during cross-validation. ReLU = rectified linear unit.

Reader Assessment for Image Quality

Readers assessed images at two different stages. Reference standard image quality grading was established using a diagnostic quality scale (10) from 0 to 4, in steps of 0.5 for finer scaling, according to artifact level, blurring, vessel sharpness (7), and noise content (Table 1). One expert reader (D.P., with 11 years of experience in coronary MR angiography), whose grading was considered as the ground truth in this work, graded all 424 anonymized datasets, and the resulting grade distribution was studied. The expert was blinded to the patient's identity, diagnosis, treatment, and any other patient-specific information. Time spent by the expert to grade each dataset was recorded.

After IQ-DCNN assessed image quality (described in the next section), the test dataset of 100 scans was assessed by another expert reader (J.Y., with 6 years of experience in coronary MR angiography), as well as a second time by the first reader more than 1 month after the first assessment. Assessment from the two readers was compared with the IQ-DCNN assessment, and inter- and intraobserver variability were determined.

Automated Image Quality Assessment Algorithm

A DCNN was designed to perform fully automated image quality assessment. The complete algorithm is presented graphically in Figure 1 and described in more detail in Appendix E1 (supplement).

In brief, 2D patches of 128 × 128 pixels were extracted in axial, sagittal, and coronal orientations from all 3D image volumes and used as input. The IQ-DCNN was designed with four convolutional layers followed by three fully connected layers. A final regression layer was used to combine the implicitly extracted image features into a quantitative image quality value for each patch. An optimizer minimized a cost-sensitive (27) antibiasing L_1 loss function ($\bar{L}_1^{\text{anti-bias}}$), which represents a measure of the similarity between the grades predicted by the network and the ground truth, while including intrinsic compensation for the nonuniform grade distribution (Eq 1).

$$\bar{L}_1^{\text{anti-bias}} = \frac{1}{\#G}\sum_{j\in G}\frac{1}{\#S_j}\sum_{i\in S_j}\left|\hat{y}_i - y_i^*\right| \quad (1)$$

Here, G is the set of all possible grades, S_j is the set of all patches having a true grade of j, $\hat{y}_i$ is the estimated grade, and y_i^* is the true grade. To prevent overfitting, dropout was used for regularization (28). Dropout is defined as a regularization technique by which neural network units and connections are randomly dropped during training.

The network architecture is broadly inspired by the VGG-16 model (29), while reducing the number of convolutional layers and having smaller fully connected layers to account for the small size of the training set. The specific architecture was selected after some initial exploratory experiments on the training and validation set.

The database was split into training and validation set (324 scans) and test set (100 scans) with similar grade and sex distributions. The training and validation set was further split into three equal parts (n = 108) for the optimization using threefold cross-validation of four preselected hyperparameters (generally defined as neural network settings that are not directly modified by optimization of the network to reduce the loss): hp_1 is size of receptive field of the first convolutional layer, hp_2 is probability of keeping connections while using dropout, hp_3 is initial learning rate, and hp_4 is half edge length of the cube of patch centers (Appendix E1

[supplement]). Finally, the patch quality grades were combined into a single grade using the average of the patch grades for each patient.

Testing the IQ-DCNN and Statistical Analysis

Comparison with expert reference.—After hyperparameter optimization, the IQ-DCNN with the lowest $\bar{L}_1^{\text{anti-bias}}$ was retrained on the training and validation set (n = 324) for evaluation on the previously unseen test data. The network's performance is reported as a boxplot of the network's quality estimation against the reference standard. Correlation coefficients, linear regression, and weighted κ statistics (Eq 2) were calculated. To validate the robustness of the training procedure, the network was retrained 10 times, and mean and standard deviation of the figures of merit were reported. Training and testing times were reported for a Linux server with a Nvidia Tesla K40c GPU, 512GB RAM, and Intel Xeon E5–2680 version 3 CPU.

Statistical analysis of intra- and interobserver variability.—The variability of the IQ-DCNN was compared with the inter- and intraobserver variability described earlier for the test set (n = 100). The weighted κ score (Eq 2) was used to quantify the inter- and intraobserver variability (30):

$$\kappa = 1 - \frac{\sum_{i,j \in G} |i-j| o_{ij}}{\sum_{i,j \in G} |i-j| e_{ij}}$$

with

$$o_{ij} = \frac{\#\left(O_1 = i \bigcap O_2 = j\right)}{N} \quad (2)$$

and

$$e_{ij} = \frac{\#\left(O_1 = i\right)\#\left(O_2 = j\right)}{N^2}$$

Here, G is the set of possible grades, O_1 and O_2 are two sets of observations, and N is the number of datasets. The IQ-DCNN grade was rounded to the nearest half integer.

κ was interpreted as follows: 0–0.2 = slight agreement, 0.2–0.4 = fair agreement, 0.4–0.6 = moderate agreement, 0.6–0.8 = substantial agreement, 0.8–1.0 = almost perfect agreement, and κ = 1 as perfect agreement. Corresponding heatmaps with agreement and disagreement counts were plotted. Krippendorff α reliability coefficients were computed for agreement between the IQ-DCNN and each of the human readers, as well as between the readers. An α value equal or greater than 0.8 was required to show agreement. Spearman rank correlation coefficients were additionally computed to account for the possible nonlinearity of the fit. Bland-Altman analysis was performed to assess the agreement between all human expert readers and the algorithm, as well as an analysis of the average variation per grade for intraobserver, interobserver, and network-observer comparisons. Variations in the response of the network with respect to the average grade between the two expert readers (considered as the expert consensus grading) were also calculated. Last, the mean absolute error and the root mean squared error between human raters and convolutional neural network output were calculated to report on the magnitude of the errors.

Heatmap generation.—To highlight the features that the DCNN uses to assign an overall quality value to a 3D volume, a heatmap was generated for one exemplary dataset. Such a heatmap depicts the relative importance of different regions as contributors for the overall image quality grade assigned by the network to the specific dataset. The heatmap can be obtained by masking different parts of an input image with a sliding window mask and evaluating the response in the output image quality score. The generation process consists of three main steps. First, for every valid mask position in slice: *(a)* mask out region and set values to zero, *(b)* grade masked image using the trained IQ-DCNN, and *(c)* add this grade to the nonmasked region in the heatmap. Second, compute average grade for every pixel in the heatmap considering the number of patches in which it was included. Finally, normalize all heatmap slices.

Validation in iterative compressed sensing image reconstruction.—Finally, the IQ-DCNN algorithm was used to evaluate the intermediate and final results of an iterative compressed sensing reconstruction. A total of 69 raw datasets from the test set were reconstructed with the pipeline described in Piccini et al (31), with four respiratory phases using extra dimensional golden-angle radial sparse parallel MRI (32). The IQ-DCNN was used to evaluate the quality of intermediate images at subsequent iterations and to compare image quality among the respiratory phases. The resulting quality evolution curves were plotted together with the evolution of the mathematical objective cost function used by the reconstruction. Paired comparisons were performed to evaluate the evolution of the image quality grades during iterative reconstruction, the average final image quality improvement, the image quality differences among the respiratory phases, and the correlation between final image quality grades and those obtained with conventional gridding. A P value ≤ .05 was considered significant. Finally, to further test the performance of the network in assessing the image quality of different reconstructions of the same dataset, an expert reader (D.P.) visually compared pairs of anonymized image volumes corresponding to the four reconstructed respiratory phases in 16 datasets. The reader was instructed to either select the phase with highest image quality of the two or assign equal quality. The expert evaluations were plotted against the same paired comparisons performed using the IQ-DCNN assessment. While the human expert can only say that one of the datasets has higher quality than the other or that they appear to be of equal quality, the IQ-DCNN always outputs a quality grade on a continuous scale and, therefore, always indicates one dataset to be of higher quality, even by only some decimals of a grade. As a consequence, and considering that the original scale was in steps of half a grade, an assumption was made that when the network indicates that there is a quality difference below one-quarter (0.25) of a grade between two datasets, a human expert would not be able to perceive it.

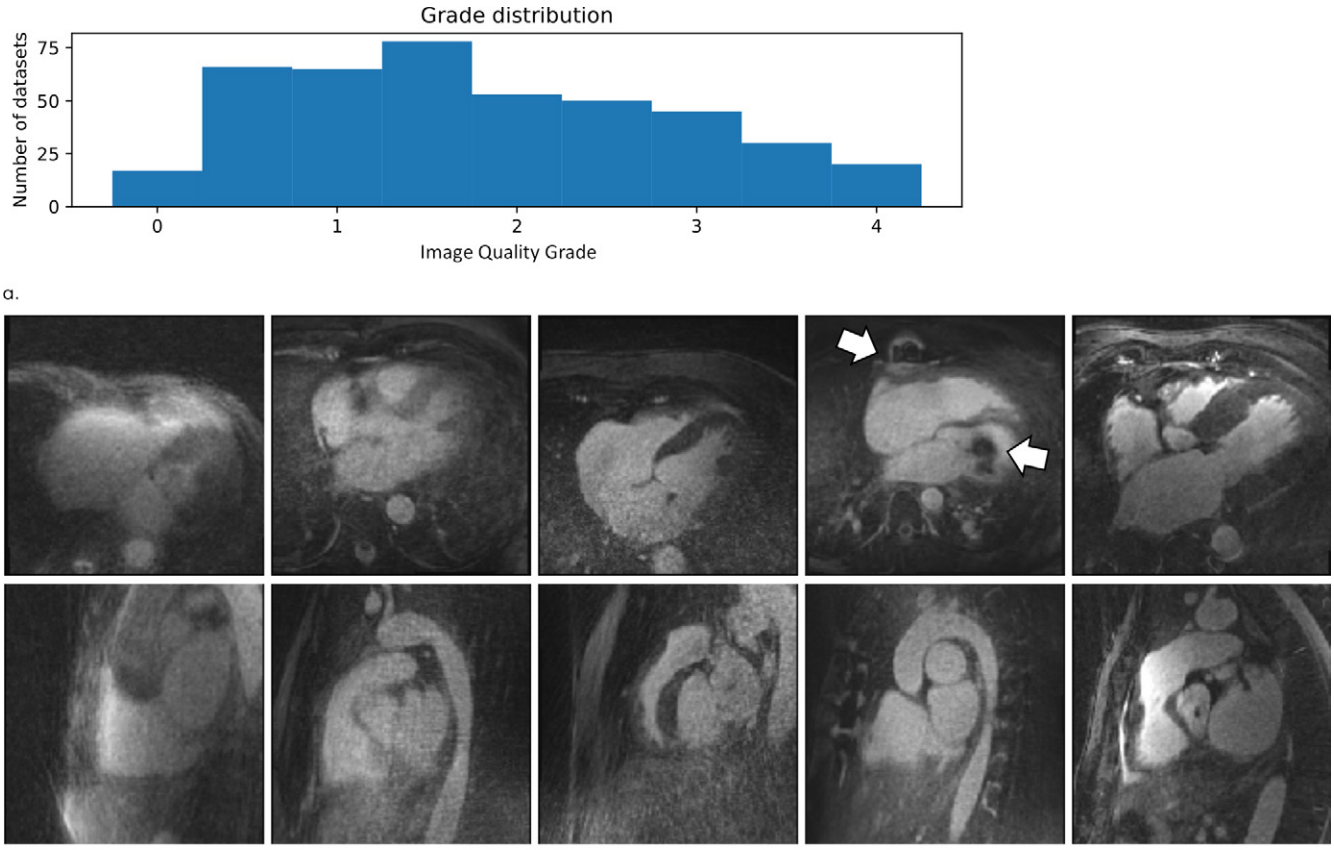

Figure 2: **(a)** Distribution of the grades in the database used in this work by observer 1. A standard image quality scale was extended to account for variations of half a grade when the expert readers were uncertain between two values. **(b)** Axial (top) and sagittal (bottom) reformats from datasets with corresponding grades from 0 to 4 (left to right) are shown for reference. Note how a dataset with a metal artifact in the chest and a clear flow artifact in the left ventricle (arrows) could still receive a high grade (grade 3) as the evaluation was performed merely on the quality of the heart structure.

Statistical analyses were performed using R version 3.5.1 (R Foundation for Statistical Computing, Vienna, Austria), with packages irr 0.84.1 and Metrics 0.1.4.

Results

Expert Image Quality Assessment for Reference Standards

Prior to IQ-DCNN training, images within the dataset (n = 424) were first assessed by an expert reader. The histogram of all grades given by the first reader (observer 1) is shown in Figure 2, together with examples of representative midventricular axial and sagittal slice reformats and their respective image quality grades. The time for an expert to grade one dataset was 100 seconds on average.

Optimization of IQ-DCNN Network Parameters

Analysis of the threefold cross-validation results for different combinations of the four hyperparameters under consideration yielded the following optimal values: size of receptive field first convolutional layer (hp_1) = 5 pixels; probability of keeping connections while using dropout (hp_2) = 0.3; initial learning rate (hp_3) = $5e^{-5}$; and half edge length of the cube of patch centers (hp_4) = 40 pixels. The mean $\bar{L}_1^{\text{anti-bias}}$ for the optimal network topology for all folds combined was 0.48 for the patches separately and 0.41 for the mean of all patches per patient.

Test Performance

After retraining the IQ-DCNN with the selected hyperparameters on the complete training set, the final performance was evaluated using the test set. Training time amounted to approximately 14 minutes per 1000 iterations, and evaluation took approximately 1.6 seconds per dataset. A boxplot of the network's patch grade output against the reference standard is shown in Figure 3a. The $\bar{L}_1^{\text{anti-bias}}$ was 0.44, the Pearson correlation (R^2) was 0.72, and the weighted κ score was 0.63. When considering the patient grades (the mean of all patch grades belonging to the same patient), $\bar{L}_1^{\text{anti-bias}}$ became 0.39, R^2 was 0.78, and κ was 0.67. The corresponding boxplot is shown in Figure 3b. Mean and standard deviation of the network's performance after 10 cycles of retraining were $\bar{L}_1^{\text{anti-bias}}$ = 0.45 ± 0.01 for the patch level and $\bar{L}_1^{\text{anti-bias}}$ = 0.40 ± 0.01, R^2 = 0.78 ± 0.01 and κ = 0.66 ± 0.01 on the patient level.

As the IQ-DCNN assigns a grade to every 2D patch, the estimated patch grade distribution within a single image volume is presented in Figure E2 (supplement) for six different datasets representative for all phenomena within the test set.

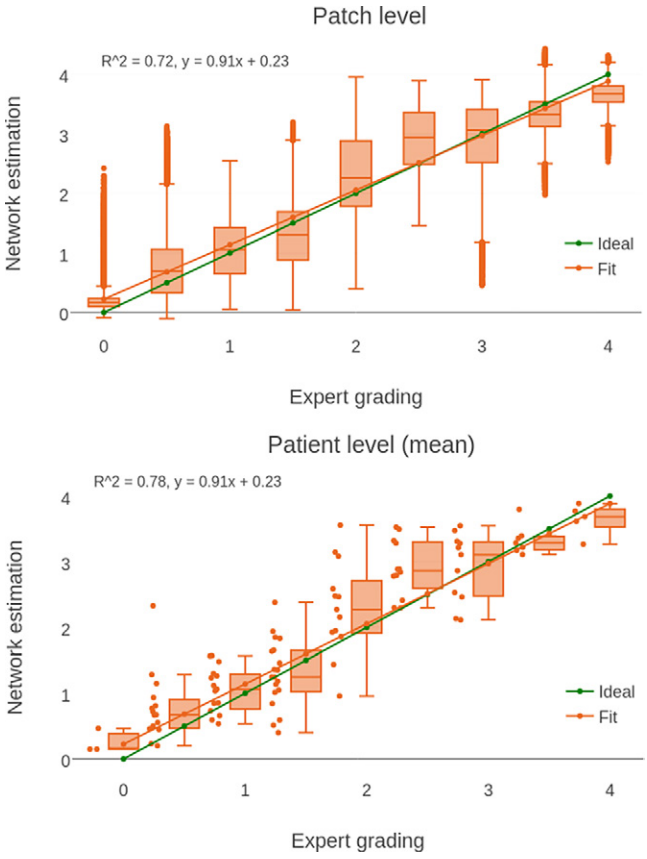

Figure 3: Boxplots and regression fits of the neural network when compared with the expert grading on the test set. While the correlation is quite high, the variability of the prediction depends on the grade. The regression fit is similar for both patch (top) and patient level (bottom), but R^2 is higher when considering all the patches for each patient. The boxplots represent the median and interquartile ranges. The dots in the patch level plot are the outliers. The dots in the patient level plot represent the actual image quality assigned to each volume of the test set.

Table 2: Bland-Altman Analyses

Rater	Observer 1b	Observer 2	IQ-DCNN
Observer 1a	0.01 ± 0.50	0.03 ± 0.57	0.07 ± 0.52
Observer 1b	…	0.04 ± 0.61	0.08 ± 0.49
Observer 2	…	…	0.04 ± 0.47

Note.—All values are means ± standard deviation. Observer 1 viewed images twice: a denotes the first time, and b denotes the second time. IQ-DCNN = image quality deep convolutional neural network.

Table 3: Weighted κ Scores

Rater	Observer 1b	Observer 2	IQ-DCNN
Observer 1a	0.70	0.67	0.67
Observer 1b	…	0.66	0.72
Observer 2	…	…	0.74

Note.—Observer 1 viewed images twice: a denotes the first time, and b denotes the second time. IQ-DCNN = image quality deep convolutional neural network.

Table 4: Comparison of Readers and IQ-DCNN

Comparison	Mean Absolute Error	Root Mean Squared Error
Observer 1 vs Observer 2	0.42	0.57
Observer 1 vs IQ-DCNN	0.41	0.53
Observer 2 vs IQ-DCNN	0.36	0.47

Note.—IQ-DCNN = image quality deep convolutional neural network.

While the red lines represent the average of all patch grades for that specific volume and therefore the final patient grade selected by the IQ-DCNN, the green lines correspond to the grades given by the expert readers (two grades by the first expert and one by the second) to the whole volume.

Comparison with Intra- and Interobserver Variability

The intraobserver agreement is presented graphically in a Bland-Altman plot in Figure E3a (supplement). The mean intraobserver agreement was substantial (κ = 0.70). The interobserver agreement is presented in Figure E3b (supplement), where the corresponding mean interobserver agreement was substantial (κ = 0.67) (33). Krippendorff α scores were 0.86 between the first reader and the IQ-DCNN, 0.91 between the second reader and the IQ-DCNN, and 0.87 between the two human readers. For the Spearman rank correlation coefficient, the correlation between the first reader and the IQ-DCNN was 0.88, between second reader and the IQ-DCNN was 0.92, and between the two human readers was 0.88.

The correspondence between the network's grade estimation and the reference is represented as a Bland-Altman plot in Figure E3c (supplement) to allow for comparison with the intra- and interobserver agreement. The agreement between all quality assessments is summarized in Table 2 by the mean and standard deviation resulting from Bland-Altman analysis and in Table 3 by the weighted κ scores. Corresponding maps with agreement and disagreement counts are reported in Figure 4. The variability from the reference grade is depicted in Figure E3d (supplement) together with the average variability per grade obtained from the single observer on the one hand and multiple observers on the other. A plot representing the error of the network with respect to the human consensus (ie, the average of the two readers) can be found in Figure E4 (supplement). Finally, mean absolute error and root mean squared error values are reported in Table 4.

Heatmap Generation

Three axial slices of the heatmap generated using one exemplary dataset are provided in Figure 5, while the full volume is provided as an animation in the Movie (supplement). Such a heatmap shows how it is not the general blurriness or sharpness of the image that determines the final grade, but rather it seems that some specific parts of the image volumes have more influ-

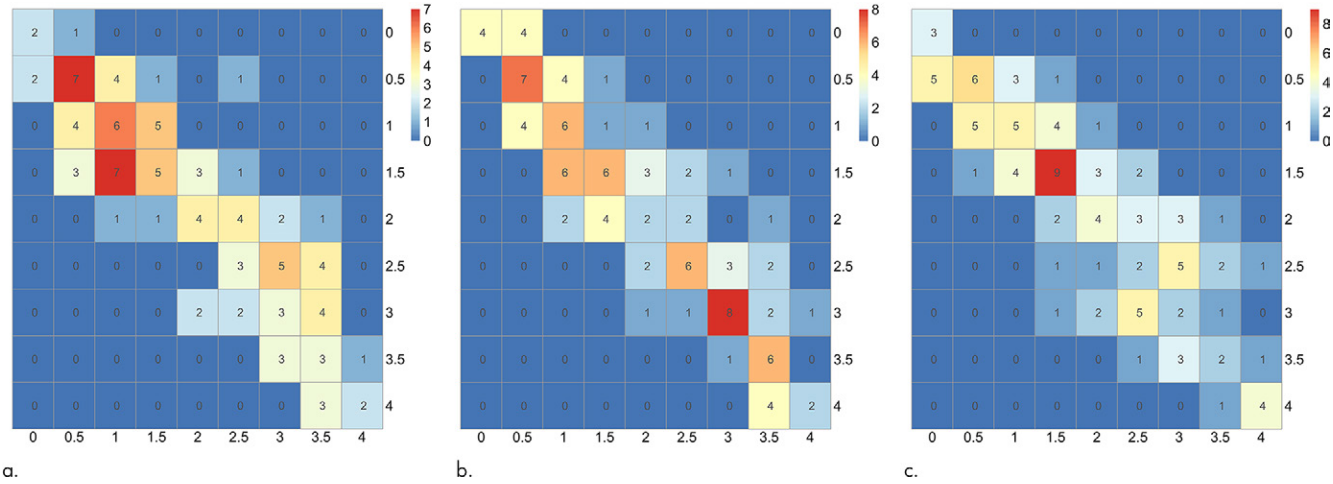

Figure 4: Maps with agreement and disagreement counts on the test set for the interobserver comparison between **(a)** the first expert reader and the IQ-DCNN, **(b)** the second expert reader and the IQ-DCNN, and **(c)** the two expert readers. It is clear from these plots how most of the counts fall along the diagonal, meaning that there is high interobserver agreement.

ence than others on the final grade. In particular, in Figure 5, we can notice how the sharpness of small vessels seems to have a major impact on the overall grade. This would suggest that deeper convolutional layers play a fundamental role in determining the response of the IQ-DCNN.

Quality Evolution in Compressed Sensing

Figure 6 represents an example dataset where the IQ-DCNN assessment was applied to images from four different respiratory phases as a function of the compressed sensing iteration (from 0 to 20). Phase 1 represents the most end-expiratory phase, while phase 4 refers to end-inspiration. Iteration 0 represents the image resulting from the first gridding step before the iterative compressed sensing optimization starts. Here, we can notice *(a)* that the improvement in image quality along with the iterations is well represented by both the objective cost function (black line, scaled to fit) and the image quality assessment (colored lines) and *(b)* that the neural network can distinguish differences in image quality among the different respiratory phases. In these datasets, a clear increase in estimated image quality is visible, and the images originating from the four respiratory phases do not converge all to the same grade. The average image quality grade for all the phases of all 69 datasets increased from 0.3 ± 0.4 (maximum of 2.5, minimum of 0.0) for the gridding step to 1.7 ± 1.1 (maximum of 3.8, minimum of 0.1) for the final iteration of the compressed sensing reconstruction. In this final iteration, the best image quality grade was assigned in 17 of 69 (25%) of the cases in phase 1, in 39 of 69 (56%) cases in phase 2, in 11 of 69 (16%) in phase 3, and two of 69 (3%) in phase 4. However, the average absolute image quality grade difference between phase 1 and phase 2 was only 0.1 ± 0.1 (range, 0.0–0.4; $P < .05$). The average grade difference between the best and worst respiratory phases obtained at the last iteration was 0.4 ± 0.3 (range, 0.1–1.4; $P < .05$). During iterative reconstruction, the image quality grade seemed to reach a plateau (grade difference < 2%) after 6 ± 2 iterations on average of the 20 total iterations evaluated. The mean image quality grade improvement from the gridding reconstruction (iteration 0) to the last iteration was 1.3 ± 0.7 (maximum of 2.7, minimum of 0.1) for all phases of all datasets. A high linear correlation ($R^2 = 0.69$) was found between the best image quality grade of the final iteration and the gridding reconstruction, as well as between the best image quality grade of the final iteration and that of the original image quality grade of the self-navigated reconstruction ($R^2 = 0.85$).

In the 16 datasets where both the expert reader and the IQ-DCNN evaluated the image quality differences between pairs of reconstructed respiratory phases, the network graded one of the two volumes to be of higher quality than the other by more than one-quarter of a grade in 21 of 96 (22%) pairwise comparisons. In these comparisons, there was 100% agreement between the human assessment and the IQ-DCNN in indicating the phase with the highest quality. Figure E5 (supplement) graphically displays the results of such comparisons for all 16 datasets.

Discussion

The use of a DCNN was proposed to automatically and quantitatively assess image quality of 3D whole-heart patient MR images mimicking human expert grading. Retraining the network several times does not yield high variability, and the algorithm tends to converge to highly equivalent solutions. The evaluation of the test set suggests that the algorithm estimates image quality with an accuracy similar to that expected from a human expert and with a precision within the bounds of the interobserver agreement between two human experts. Although the overall correlation with the expert assessment is high, very low and very high grades are slightly over- and underestimated, the variability increases for medium grades, and averaging the patch grades into a single patient grade yields better predictions by lowering the variance. In image volumes with discordant expert values (Fig 4), the estimation of quality is often less straightforward and multiple peaks occasionally exist, suggesting that different experts may have looked at different image properties. The generated heatmap suggests how in this specific case, it is not the general sharpness of the im-

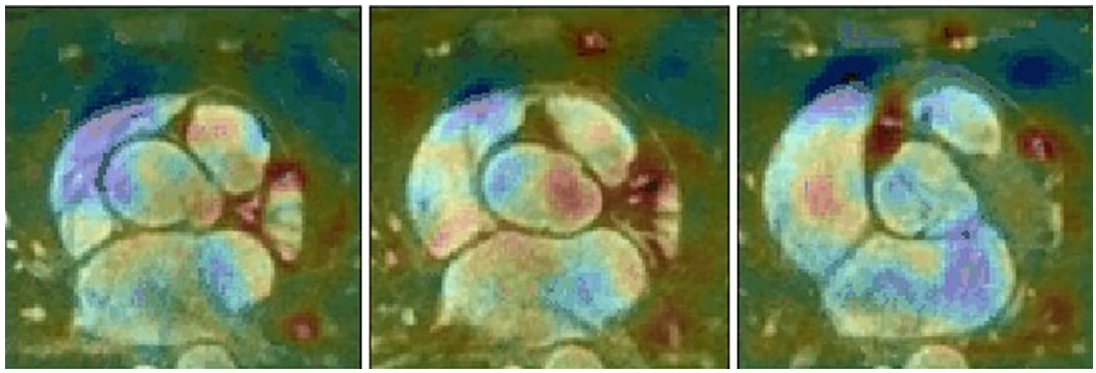

Figure 5: Three representative axial slices from the heatmap display the relative contribution of different anatomic regions to the final quality grade. Some small structures, mainly vessels and some of the edges, seem to have a major influence in the final grade. The full volume of the heatmap can be found in the Movie (supplement).

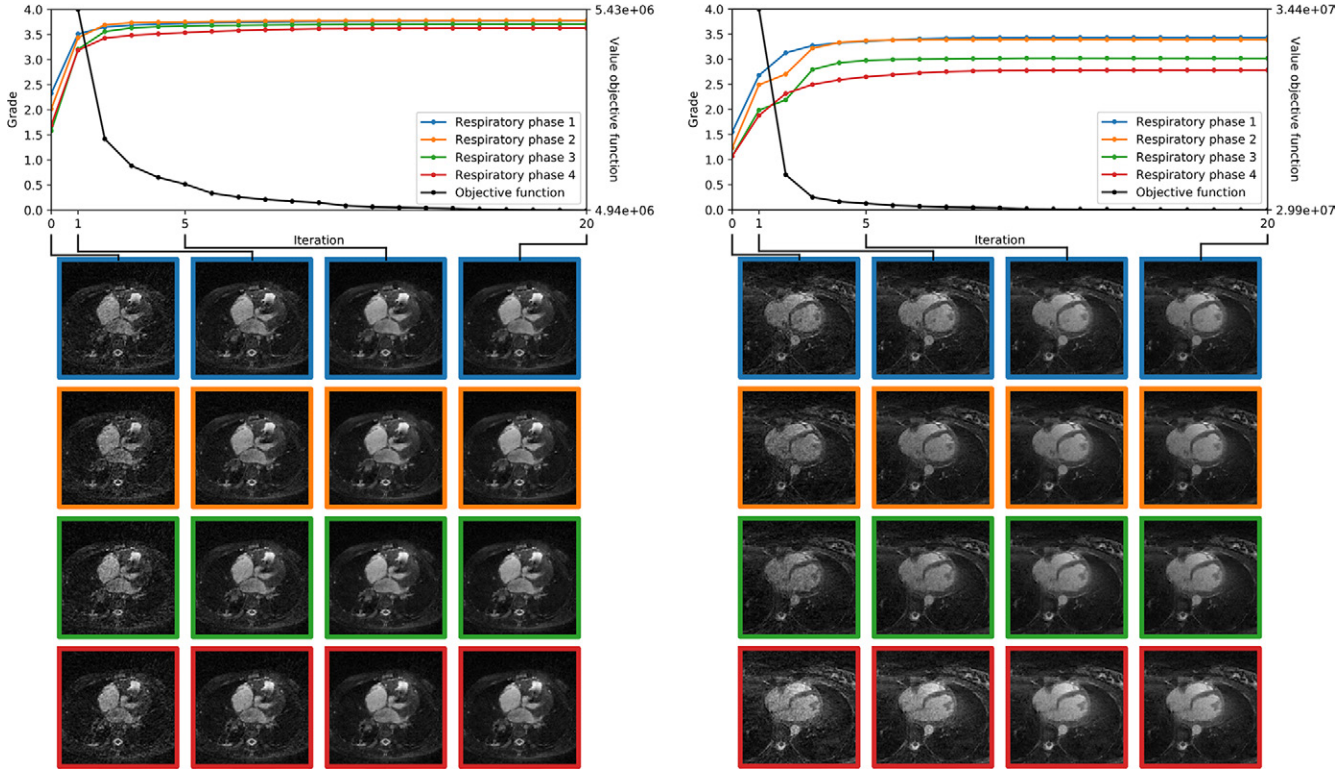

Figure 6: Two examples of automated image quality assessment during iterative compressed sensing reconstruction using the proposed deep convolutional neural network. The quality of the whole-heart image volumes corresponding to the four reconstructed respiratory phases is assessed at each iteration on the standard image quality scale from 0 (poor quality) to 4 (excellent). Phase 1 represents the most end-expiratory phase, while phase 4 refers to end-inspiration.

age that seems to determine the final grade, but rather specific parts of the anatomy. Some features seem to be more prominent than others in contributing to the final grade. Specifically, small vessels seem to have a major impact on the overall grade, which corresponds nicely with the fact that the coronary tree is a structure often looked at for determining the overall quality of a whole-heart dataset. Finally, considering the results obtained for the respiratory-resolved compressed sensing reconstructions, it may be concluded that the algorithm can differentiate image quality even though it had never previously been exposed to equivalent data during the training phase. In this scenario, the comparison of the grades at each iteration and the objective function of the iterative reconstruction can be considered as a validation of the network output. Moreover, the distribution of the best image quality grade within the four respiratory phases seems to confirm the results in Ginami et al (34), while these findings also advance the hypothesis that automated image quality assessment could directly be applied for the definition of a stopping criterion for iterative reconstruction (ie, when the quality score reaches a plateau even though the objective function keeps decreasing) and that it has a potential role and utility in comparison studies. A robust automated image quality assessment algorithm would prove very useful within the clinic. For example, an algorithm could be used to assess several different reconstruction techniques and could provide the best image quality results.

In general, clinical and technical studies may greatly benefit from an automated and quantitative determination of image quality among two or more images. Moreover, because data volumes will inevitably increase over time, an expert-based visual decision making on image quality will become a daunting task. Therefore, support mechanisms that help the expert make better informed decisions in a time-efficient way will be critically important. For instance, the IQ-DCNN algorithm can help automatically select the respiratory and cardiac phase with the best image quality (35).

This study had limitations. As the DCNN needs to be trained, the first requirement before optimization and validation of an automated image quality algorithm is the availability of human expert–established reference standard images. In this specific case, the distribution of the quality grades was highly nonuniform, and it was therefore necessary to use antibiasing methods during the optimization of the algorithm to avoid prediction bias toward low and mid grades. A variety of DCNNs exist with numerous hyperparameters. Therefore, designing the optimal network for a given task is not straightforward and, ideally, an exhaustive optimality search should be performed within the whole hyperspace of hyperparameters. Although the network used in this study represents a high-dimensional optimization problem, empirical experience in other applications of deep learning suggests that, even in these cases, stochastic gradient descent results in acceptable solutions with good error bounds on unseen data, possibly because it tends to find local minima which are reasonably flat (large basins). This is in opposition to vanilla gradient (simplest) descent which tends to converge to sharper minima and is due to the fact that stochastic gradient descent methods rely on random initialization and discrete gradient steps. In this context, flatter minima are thought to yield better generalization on unseen data than sharper minima. However, the generalization capacity of deep learning models is an area of active research with several plausible explanations (36). To understand whether the depth of the IQ-DCNN was justified and to assess what parts of the image volumes most contribute to the overall final grade, a heatmap was generated. Furthermore, an extension from 2D to 3D patches can be considered at the expense of increased computational demands. The neural network performance may also depend on the image quality criteria used by the human experts to grade the images. In the current study, general quality of the anatomic structures and coronary vessel sharpness was used to assess quality, and therefore, the reported performance of the DCNN should not be extrapolated to other image quality criteria related to features such as valves, scars, or ischemia visualization. Other limitations included that the network was trained only based on one single reader and that this was a single-center study, using a single-vendor scanner and one specific MRI sequence. Although training the network according to a consensus grading between two or more expert readers may improve the generalizability of the approach, there seems to be an intrinsically increased uncertainty for both readers (intra- and interobserver alike) and for the network when it comes to grading datasets of intermediate quality. This can be detected in the regression plots, as well as in the Bland-Altman plots and in the correlation between interobserver uncertainty and uncertainty of the IQ-DCNN with respect to a consensus grading between the two observers.

In conclusion, an automated image quality assessment algorithm employing a DCNN has been implemented, tested, and applied to clinical 3D cardiac MR images in patients for the determination of image quality. It has been shown that the algorithm is capable of estimating image quality with good agreement with respect to human expert reading. The accuracy was found to be within the bounds of expert interobserver variability. Applied to an iterative compressed sensing pipeline where different respiratory phases are reconstructed, the algorithm was not only capable of identifying improvements in image quality as a function of the increasing iterations, but also as a function of the respiratory level. In all, our findings suggest that a variant of this algorithm may be exploited to define termination criteria of an iterative process to improve time efficiency and to help identify data with the least amount of motion blurring.

Author contributions: Guarantor of integrity of entire study, D.P.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, D.P., R.D., J.S., J.R., T.K.; clinical studies, D.P., R.D., P.G.M., J.S., T.K.; experimental studies, R.D., J.H., J.Y., L.D.S.; statistical analysis, R.D., J.H., D.V.D.V.; and manuscript editing, all authors

Disclosures of Conflicts of Interest: D.P. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: employed by and has stock options in Siemens Healthcare; payment from patent with Siemens Healthcare; pending patent (Siemens gives incentive to write patent applications). Other relationships: disclosed no relevant relationships. **R.D.** disclosed no relevant relationships. **J.H.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: author's PhD studies are financially supported by Siemens Healthcare (Erlangen, Germany). Other relationships: disclosed no relevant relationships. **J.Y.** disclosed no relevant relationships. **L.D.S.** disclosed no relevant relationships. **P.G.M.** disclosed no relevant relationships. **J.S.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: institution receives unrestricted grant to support research activity of cardiac MR center. Other relationships: disclosed no relevant relationships. **D.V.D.V.** disclosed no relevant relationships. **J.R.** Activities related to the present article: 30% of author's salary paid by Siemens Healthcare Switzerland; author owns shares in Siemens Healthcare. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **T.K.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: employed by Siemens Healthcare Switzerland; participates in Siemens Healthcare Switzerland employee stock share program. Other relationships: disclosed no relevant relationships. **M.S.** Activities related to the present article: institution receives grant from Swiss National Science Foundation. Activities not related to the present article: author receives nonmonetary research support from Siemens Healthineers. Other relationships: disclosed no relevant relationships.

References

1. Martin CJ, Sharp PF, Sutton DG. Measurement of image quality in diagnostic radiology. Appl Radiat Isot 1999;50(1):21–38.
2. Chow LS, Paramesran R. Review of medical image quality assessment. Biomed Signal Process Control 2016;27:145–154.
3. Klinke V, Muzzarelli S, Lauriers N, et al. Quality assessment of cardiovascular magnetic resonance in the setting of the European CMR registry: description and validation of standardized criteria. J Cardiovasc Magn Reson 2013;15(1):55.
4. Schwitter J, Kanal E, Schmitt M, et al. Impact of the Advisa MRI pacing system on the diagnostic quality of cardiac MR images and contraction patterns of cardiac muscle during scans: Advisa MRI randomized clinical multicenter study results. Heart Rhythm 2013;10(6):864–872.

5. Schwitter J, Gold MR, Al Fagih A, et al. Image Quality of Cardiac Magnetic Resonance Imaging in Patients With an Implantable Cardioverter Defibrillator System Designed for the Magnetic Resonance Imaging Environment. Circ Cardiovasc Imaging 2016;9(5):e004025.
6. Krupinski EA. Current perspectives in medical image perception. Atten Percept Psychophys 2010;72(5):1205–1217.
7. Etienne A, Botnar RM, Van Muiswinkel AMC, Boesiger P, Manning WJ, Stuber M. "Soap-Bubble" visualization and quantitative analysis of 3D coronary magnetic resonance angiograms. Magn Reson Med 2002;48(4):658–666.
8. Rutz T, Piccini D, Coppo S, et al. Improved border sharpness of post-infarct scar by a novel self-navigated free-breathing high-resolution 3D whole-heart inversion recovery magnetic resonance approach. Int J Cardiovasc Imaging 2016;32(12):1735–1744.
9. Vincenti G, Monney P, Chaptinel J, et al. Compressed sensing single-breathhold CMR for fast quantification of LV function, volumes, and mass. JACC Cardiovasc Imaging 2014;7(9):882–892.
10. McConnell MV, Khasgiwala VC, Savord BJ, et al. Comparison of respiratory suppression methods and navigator locations for MR coronary angiography. AJR Am J Roentgenol 1997;168(5):1369–1375.
11. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–444.
12. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 2004;13(4):600–612.
13. Kryszczuk K, Drygajlo A. On face image quality measures. Proc 2nd Work Multimodal User Authentication, 2006.
14. Bosse S, Maniry D, Muller KR, Wiegand T, Samek W. Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment. IEEE Trans Image Process 2018;27(1):206–219.
15. Bianco S, Celona L, Napoletano P, Schettini R. On the use of deep learning for blind image quality assessment. Signal Image Video Process 2018;12(2):355–362.
16. Saha SK, Fernando B, Cuadros J, Xiao D, Kanagasingam Y. Automated Quality Assessment of Colour Fundus Images for Diabetic Retinopathy Screening in Telemedicine. J Digit Imaging 2018;31(6):869–878.
17. Mahapatra D, Roy PK, Sedai S, Garnavi R. Retinal image quality classification using saliency maps and CNNs. In: Wang L, Adeli E, Wang Q, Shi Y, Suk HI, eds. Machine Learning in Medical Imaging. MLMI 2016. Lecture Notes in Computer Science, vol 10019. Cham, Switzerland: Springer, 2016; 172–179.
18. Küstner T, Liebgott A, Mauch L, et al. Automated reference-free detection of motion artifacts in magnetic resonance images. MAGMA 2018;31(2):243–256.
19. Esses SJ, Lu X, Zhao T, et al. Automated image quality evaluation of T2 -weighted liver MRI utilizing deep learning architecture. J Magn Reson Imaging 2018;47(3):723–728.
20. Camps SM, Houben T, Fontanarosa D, et al. One-class Gaussian process regressor for quality assessment of transperineal ultrasound images. Int Conf Med Imaging with Deep Learn Nips, 2018.
21. Zhang L, Gooya A, Dong B, et al. Automated Quality Assessment of Cardiac MR Images Using Convolutional Neural Networks. In: Tsaftaris S, Gooya A, Frangi A, Prince J, eds. Simulation and Synthesis in Medical Imaging. SASHIMI 2016. Lecture Notes in Computer Science, vol 9968. Cham, Switzerland: Springer, 2016; 138–145.
22. Zhang L, Gooya A, Frangi AF. Semi-supervised assessment of incomplete LV coverage in cardiac MRI using generative adversarial nets. In: Tsaftaris S, Gooya A, Frangi A, Prince J, eds. Simulation and Synthesis in Medical Imaging. SASHIMI 2017. Lecture Notes in Computer Science, vol 10557. Cham, Switzerland: Springer, 2017; 61–68.
23. Piccini D, Littmann A, Nielles-Vallespin S, Zenge MO. Respiratory self-navigation for whole-heart bright-blood coronary MRI: methods for robust isolation and automatic segmentation of the blood pool. Magn Reson Med 2012;68(2):571–579.
24. Piccini D, Monney P, Sierro C, et al. Respiratory self-navigated postcontrast whole-heart coronary MR angiography: initial experience in patients. Radiology 2014;270(2):378–386.
25. Monney P, Piccini D, Rutz T, et al. Single centre experience of the application of self navigated 3D whole heart cardiovascular magnetic resonance for the assessment of cardiac anatomy in congenital heart disease. J Cardiovasc Magn Reson 2015;17:55 [Published correction appears in J Cardiovasc Magn Reson 2015;17:88.].
26. Albrecht MH, Varga-Szemes A, Schoepf UJ, et al. Coronary artery assessment using self-navigated free-breathing radial whole-heart magnetic resonance angiography in patients with congenital heart disease. Eur Radiol 2018;28(3):1267–1275.
27. He H, Garcia EA. Learning from Imbalanced Data. IEEE Trans Knowl Data Eng 2009;21(9):1263–1284.
28. Sutskever I, Hinton G, Krizhevsky A, Salakhutdinov RR. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res 2014;15(56):1929–1958.
29. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In: Bengio Y, LeCun Y, eds. 3rd International Conference on Learning Representation (ICLR) 2015, San Diego, CA, USA, May 7-9, 2015, Conf Track Proc. 2015. ArXiv 1409.1556 [preprint] http://arxiv.org/abs/1409.1556. Posted 2014.
30. Kundel HL, Polansky M. Measurement of observer agreement. Radiology 2003;228(2):303–308.
31. Piccini D, Feng L, Bonanno G, et al. Four-dimensional respiratory motion-resolved whole heart coronary MR angiography. Magn Reson Med 2017;77(4):1473–1484.
32. Feng L, Axel L, Chandarana H, Block KT, Sodickson DK, Otazo R. XD-GRASP: Golden-angle radial MRI with reconstruction of extra motion-state dimensions using compressed sensing. Magn Reson Med 2016;75(2):775–788.
33. Watson PF, Petrie A. Method agreement analysis: a review of correct methodology. Theriogenology 2010;73(9):1167–1179.
34. Ginami G, Bonanno G, Schwitter J, Stuber M, Piccini D. An iterative approach to respiratory self-navigated whole-heart coronary MRA significantly improves image quality in a preliminary patient study. Magn Reson Med 2016;75(4):1594–1604.
35. Feng L, Coppo S, Piccini D, et al. 5D whole-heart sparse MRI. Magn Reson Med 2018;79(2):826–838.
36. Neyshabur B, Bhojanapalli S, Mcallester D, Srebro N. Exploring Generalization in Deep Learning. In: Guyon I, Luxburg UV, Bengio S, et al, eds. In Advances in Neural Information Processing Systems 30. Red Hook, NY: Curran Associates, 2017; 5947–5956. http://papers.nips.cc/paper/7176-exploring-generalization-in-deep-learning.pdf. Accessed February 2020.