Journal of Imaging Informatics in Medicine (2025) 38:3444–3469
https://doi.org/10.1007/s10278-025-01462-1

ORIGINAL PAPER

A Study of Why We Need to Reassess Full Reference Image Quality Assessment with Medical Images

Anna Breger[1,2] · Ander Biguri[1] · Malena Sabaté Landman[3] · Ian Selby[4] · Nicole Amberg[5] · Elisabeth Brunner[2] · Janek Gröhl[6,7] · Sepideh Hatamikia[8,9] · Clemens Karner[2] · Lipeng Ning[10] · Sören Dittmer[1] · Michael Roberts[1] · A.I.X.-C.O.V.N.E.T. Collaboration · Carola-Bibiane Schönlieb[1]

Received: 28 October 2024 / Revised: 31 January 2025 / Accepted: 18 February 2025 / Published online: 24 March 2025

Abstract

Image quality assessment (IQA) is indispensable in clinical practice to ensure high standards, as well as in the development stage of machine learning algorithms that operate on medical images. The popular full reference (FR) IQA measures PSNR and SSIM are known and tested for working successfully in many natural imaging tasks, but discrepancies in medical scenarios have been reported in the literature, highlighting the gap between development and actual clinical application. Such inconsistencies are not surprising, as medical images have very different properties than natural images, and PSNR and SSIM have neither been targeted nor properly tested for medical images. This may cause unforeseen problems in clinical applications due to wrong judgement of novel methods. This paper provides a structured and comprehensive overview of examples where PSNR and SSIM prove to be unsuitable for the assessment of novel algorithms using different kinds of medical images, including real-world MRI, CT, OCT, X-Ray, digital pathology and photoacoustic imaging data. Therefore, improvement is urgently needed in particular in this era of AI to increase reliability and explainability in machine learning for medical imaging and beyond. Lastly, we will provide ideas for future research as well as suggest guidelines for the usage of FR-IQA measures applied to medical images.

Keywords Image quality · Quality measures · Peak signal-to-noise ratio · Structural similarity index · Pitfalls

Introduction

Advances in medical imaging technologies have been groundbreaking in the last decades, ranging from new modalities of scanners, including hardware innovations, and advances in mathematical tools for image reconstruction to the current state of the art in machine learning techniques. The overall aim is to apply novel technology in clinical scenarios to improve patient's care. In order to ensure a clinically acceptable quality of the novel imaging techniques, quantitative image quality assessment (IQA) plays an important role for quality assurance in addition to visual inspections.

Quantitative IQA can roughly be divided into three categories based on their underlying assumptions and the available information for their evaluation (cf. [83]). The first one is called full reference (FR) IQA, where a full known image is used as a reference (or ground truth) and the quality of a given image is evaluated in a comparative way that relies on a meaningful notion of distance between the two images. No reference (NR) IQA, on the other hand, is not based on a one-to-one image comparison, and instead, it aims to judge the quality of a given image solely by evaluating properties. Lastly, reduced reference (RR) IQA lays somewhere in between, for example, using specific retrieved image information such as edge information or local image characteristics as a reference. Most commonly, NR- and FR-IQA measures have been developed and used to solve quite different problems. As FR-IQA requires a reference image, it can only be used in very specific tasks. This includes the evaluation of novel (traditional or machine learning-based) imaging methods in their development stage and experiment calibration, where reference data is available. In this case, the measure is used to make conclusions about the performance of the algorithms or settings in different imaging tasks, such as image compression, denoising or reconstruction, and, consequently, to reinforce the success of the different methods. NR-IQA, on the other hand, is used to extract and judge

A list of members is provided in the acknowledgements.

Extended author information available on the last page of the article

quality information from a given image on the spot, particularly when there is no access to the reference image. This is usually the case when evaluating the quality of a given image outside of the development stage, such as real-time quality control of image acquisition in a hospital.

In this paper, we focus on FR-IQA measures that have been broadly used for the evaluation of novel image processing algorithms that are operating on medical images. Research on the development of novel methodologies, mainly driven within the areas of mathematics, computer science and engineering, often uses FR-IQA for the evaluation of the proposed algorithms. The performance of the methods under these metrics influences which methods are published; therefore, the choice of FR-IQA at the developing stage has a capital influence on the method choices that are subsequently available. However, the authors carrying out research at the development stage are not necessarily experts on particular applications and, therefore, might not take into account the specific nuisances of medical imaging data and the importance of corresponding IQA measure choice. For this reason, the field might be promoting new methodologies which are not the most suited for clinical tasks, such as diagnosis of specific data.

Some fundamental questions come alongside the task of evaluating novel methodologies and their application. Could or should the development of novel methods be decoupled from their application potential? Is it feasible to develop highly sophisticated algorithms and assess their applicability without an expert's opinion? And, on the other hand, is it feasible to assume that an application expert needs to be available every time a new methodology is developed? In the last decades, these unsolved principles have led to rather disconnected research areas of model development and eventual application. This is particularly true in the fast-advancing domain of machine learning, where publicly shared data is presumably enabling the development of novel algorithms for specific applications more easily, but where the lack of direct contact with experts on the data is hindering their real applicability, e.g. in medical imaging. In that case, accurate evaluation with IQA measures becomes even more important as visual inspection from experts is not feasible.

However, the most commonly employed FR-IQA metrics have not been developed nor broadly tested to work successfully for medical images. Usually, novel quality measures are tested by computing correlations of their outcome to the mean opinion scores of experts who have manually rated the images. In order to do so, there are several standard data bases that provide such rated data; see, e.g. the LIVE database [87] or the image database TID2012 [74]. However, these kinds of annotations are rare, as they are very time-consuming and task-dependent. For medical images, there is another layer of added difficulty: it is very hard to publish clinical images due to the sensitivity of the data. For these reasons, currently and to the best of our knowledge, there has been no publicly available database with FR ratings for medical images to assess the performance of the quality measures. Thereto, we have recently published a data set with photoacoustic images and expert annotations, which are available on Zenodo [16]. Previously published studies including medical in-house data have raised concerning results [17, 51]. The data sets include ratings by experts and radiologists, showing that the most common FR-IQA measures perform poorly for the studied tasks.

With that, this paper has the following main objectives:

1. Providing a structured collection of pitfalls of standard FR-IQA measures (namely PSNR and SSIM, as well as the more recently introduced LPIPS) when used for common medical imaging tasks. We show real-world examples in different medical imaging applications. The examples are described in detail, because without in-depth discussion the huge challenges of the problem cannot be understood appropriately.
2. Opening a discussion about the choice of existing FR-IQA measures for medical application as well as desirable properties for potential novel frameworks. They should facilitate clinical applicability and cement a more functional knowledge transfer between developers and users.
3. Suggesting general guidelines on how to use FR-IQA in the setting of medical imaging safely, as well as highlighting an existing problem on the lack of proper reporting of employed measures.

The idea of generalized image quality measures not being appropriate in medical imaging has been explored before, and some guidelines for task-adapted quality assessment exist in the literature. While a full review is out of the scope of this work, it is worth noting the research that was started by Barret et al. in Objective Assessment of Image Quality [8–10] regarding task-based assessment with observer models, relying on the specification of a task, observer and an image ensemble. The field of objective IQA for medical imaging has been active since, in more recent works also discussing various modalities, e.g. low dose CT [19], MRI [62] or even multi-modal imaging [24]. Lastly, on a related note, human perception in medical imaging has been an active field of research, where, e.g. The Handbook of Medical Image Perception and Techniques (ed.2) [82] is comprehensively discussing research on image perception, observer models and clinical relevance.

Outline

The paper is structured as follows. The "Background" section contains an overview of the most commonly used/standard FR-IQA measures and their background. In the "Examples

of Failure in Medical Imaging" section, every subsection reports the use of these FR-IQA measures in a different medical imaging modality, including a description of the corresponding reconstruction method or visualization problem and examples of failure. Finally, the "Discussion" section includes a discussion, suggestions for future directions and guidelines for task-informed usage of FR-IQA measures in the context of medical imaging.

Here, working across medical imaging domains, an international collaboration was formed that includes experts of the specific imaging techniques in order to provide insights for the different image modalities described in the "Examples of Failure in Medical Imaging" section. With that, we were able to ensure for every example to include at least one expert working in the particular field to provide the required insights for a comprehensive analysis and task-specific judgement of the obtained image qualities.

Background

The mean squared error (MSE) is a common FR-IQA metric used to measure the average squared difference between a given image and the reference, i.e. for a given reference image $I \in \mathbb{R}^{N_1 \times N_2}$ and a corresponding processed version $J \in \mathbb{R}^{N_1 \times N_2}$, the MSE is given by the Frobenius-norm, i.e.

$$\mathrm{MSE}(I, J) := \frac{1}{N_1 \cdot N_2} \|I - J\|_F^2.$$

Lower MSE values indicate that the processed image values are closer to the reference values. However, it is well known that the MSE used as a measure of image quality does not correspond well to human perception (cf. [31, 34, 57, 104]) and does not provide a consistent quantity regarding severeness of image degradation. Therefore, even in the cases where the computation of the root-MSE can serve as a useful quantity measure of deviation for some medical imaging modalities (e.g. in MRI), this would not correspond to an assessment related to a perceptual measure. A closely related measure was introduced in the early 2000s, the so-called peak signal-to-noise ratio (PSNR), which provides a re-scaled version of the MSE:

$$\mathrm{PSNR}(I, J) := 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}(I, J)}\right), \tag{1}$$

where MAX corresponds to the maximal possible intensity value, i.e. for an 8-bit image MAX $= 255$. As the PSNR solely relies on the MSE, most disadvantages and problems that are known for the MSE (e.g. same value for very different degrees of degradation) do also hold true for the PSNR.

Two decades ago, the framework of the structural similarity index (SSIM) [105] was introduced, which relies on three comparison components, luminance, contrast and structure, and can be calculated on various batches, i.e. local parts of an image. For simplicity, here, we call the batches again I and J, and then the measure is given by

$$\mathrm{SSIM}(I, J) = \frac{(2\mu_I\mu_J + c_1)(2\sigma_{IJ} + c_2)}{(\mu_I^2 + \mu_J^2 + c_1)(\sigma_I^2 + \sigma_J^2 + c_2)}, \tag{2}$$

where μ corresponds to the mean, σ to the covariance and c_1, c_2 are scaling factors. The values of SSIM theoretically range between -1 and 1, where a higher value indicates greater similarity between the images, in the sense that they are more visually alike. This measure has celebrated great success for natural images since the general framework allows for a greater insight if used appropriately. However, in practice, standard implementations only provide the mean over batches in the image as the final image quality measure. This choice increases the chance of failure in local error detection, which is often crucial in the medical settings. Limitations of the SSIM in the medical setting have been reported, e.g. in [35, 61, 71]. This is not surprising, since SSIM was not only not developed to assess the quality of medical images, but SSIM was also originally not tested for its use on medical images. Moreover, it is important to note that the SSIM framework allows to set a number of parameters, including the choice of kernel, and therefore can yield diverging results which depend on the implementation used (see the paper [101] for a detailed analysis of the variations). This means, for comparability and reproducibility, it should always be stated in detail which implementation/parameter setting of SSIM was used when applied.

Many other structural FR-IQA measures have been developed in the last decade, for example the HaarPSI [78] measure based on Haar wavelets. Most recently, LPIPS [110], a highly successful FR-IQA measure for natural images, that quantifies the perceptual similarity between two images based on features learned from deep convolutional neural networks has been suggested to be included in the evaluation for medical images (see, e.g. [80] or [91, 99]). Although LPIPS holds some properties that are beneficial in the medical setting, including the invariance to small spatial perturbations, it has not been rigorously tested nor developed for medical images. To gain more insights of that recent development, we are including the results of LPIPS (based on the default AlexNet) applied to the provided examples, where a smaller value indicates better similarity. It is important to note that the framework generally allows the development of your own learned quality measure and provides different measures based on several networks.

Although there is an enormous number of newly introduced FR-IQA measures—including methods developed for medical imaging tasks (see, e.g. [22]), and a published review from 2016 describing the current situation in medical imag-

ing [21]—the most used FR-IQA measures for assessment of classical medical imaging algorithms, such as image reconstruction, image restoration, super resolution or denoising/artefact removal, are still PSNR and SSIM. Summarizing the problem, commonly used FR-IQA measures have been developed for natural images which consist of very different properties than medical images.

Illustration of the Failure of Common FR-IQA Measures on Synthetic Image Degradations

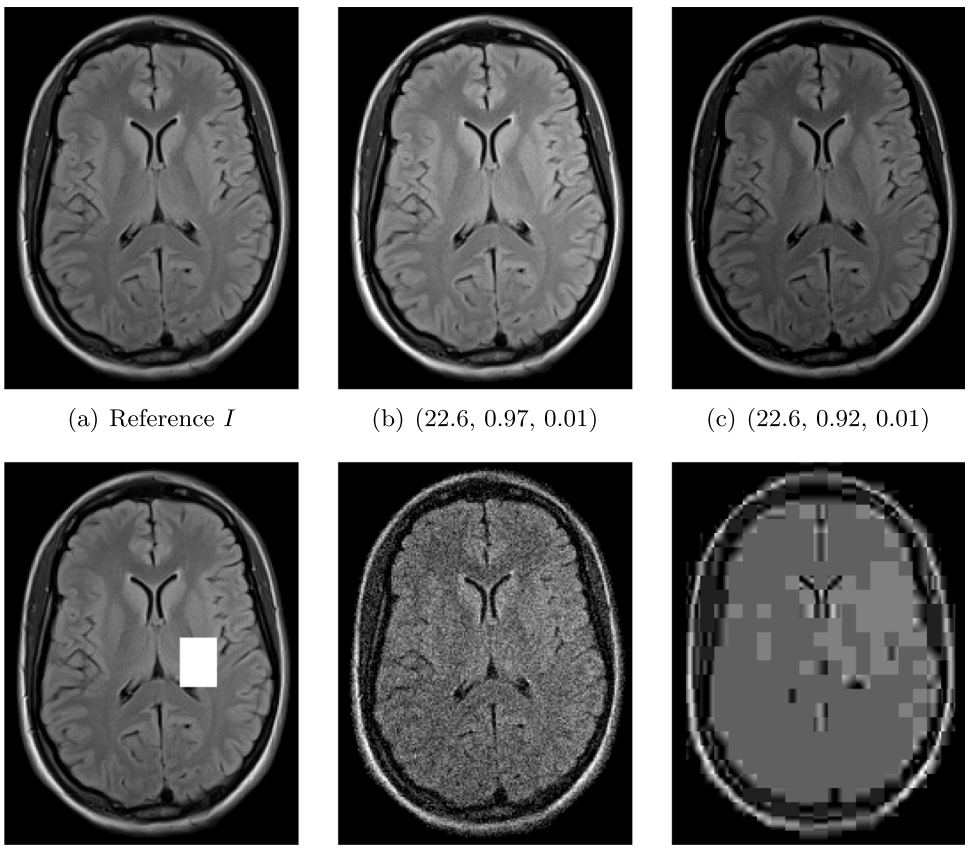
In Fig. 1, we show a toy example of misjudgments that occur when evaluating the quality of a 2D MRI scan compared to degraded versions of the same image with the selected measures. PSNR even yields the same value for all the very different degradations, and all of the tested measures fail in the judgement of massive local information loss (d), as well as in the judgement of stochastic noise (e) versus block artefacts (f). This toy example served as an inspiration to study the behavior of the standard measures in real-life medical imaging tasks.

Examples of Failure in Medical Imaging

In this section, we will present failures of PSNR/SSIM/LPIPS when applied to medical problems with real-world image data. This contains examples of tasks where the measures are currently used as standard choices in the evaluation, as well as tasks where automated objective evaluation is still an open field and urgently needed.

The structure will be as follows. In each subsection, a medical imaging modality will be shortly introduced, followed by a problem formulation in which IQA measures play an important role. Finally, a corresponding example is shown visually with a short discussion in which regards the measures act inaccurate for that problem.

The numbers provided in the subfigures correspond to the PSNR, SSIM, and LPIPS values in comparison to the reference image (a), respectively, where in-built functions of MatlabR2023a were used to compute PSNR and SSIM (namely, default settings for *psnr* and *ssim*), and for LPIPS, the Python implementation based on AlexNet provided by the authors was used [109, 110]. We computed all measures on the visualized 2D images as shown in the paper, i.e. no



(a) Reference I (b) (22.6, 0.97, 0.01) (c) (22.6, 0.92, 0.01)

(d) (22.6, 0.98, 0.08) (e) (22.6, 0.64, 0.39) (f) (22.6, 0.63, 0.27)

Fig. 1 Illustrative toy example of problems occurring when using the standard FR-IQA measures PSNR/SSIM/LPIPS for the evaluation of medical images. Degradations have been added artificially to the reference (**a**) MRI scan: contrast enhancement (**b**), brightness change (**c**), hole (**d**), Gaussian white noise (**e**), jpeg compression (**f**). PSNR yields the same value for all degradations, and SSIM and LPIPS fail to identify the hole (**d**) and misjudge the quality of (**e**) and (**f**)

further contrast/luminance adjustment was added, to ensure visual comparability to the provided numbers. The image data was first scaled according to standards in the different fields; for the CT and photoacoustic data, clipping to a predefined range was applied, and afterwards, all images have been standardized by scaling them between 0 and 1 (division by the maximum, which corresponds to the maximum of the reference image in case clipping was applied). All employed image data, besides the data from digital pathology, is originally grey-scale. Depending on the original data format, the images were saved as uint8 or uint16 images in portable network graphic (png) format.

Computed Tomography

Computed tomography (CT) is an imaging modality that aims to reconstruct 2D or 3D volumes from X-ray attenuation measurements and enables high-quality structural imaging of patients [79]. The applications of CT range from diagnostics, surgery planning and radiation therapy to image-guided interventions, making this imaging modality ubiquitous in modern medicine.

Reconstruction Problem

CT is not a direct measurement method and images need to be reconstructed by solving a large-scale system of linear equations. One of the main challenges with this task is ill-posedness, which means that in some scenarios, small perturbations on the measurements can generate large perturbations on the recovered image. Particularly problematic are limited datasets, e.g. when only limited angle or sparse full-angle tomography measurements are available, as well as the presence of noise in the measurements. In these cases, the direct and most used approach to compute a solution, i.e. the so-called filtered backprojection (FBP), can be highly corrupted by noise [12].

Different families of iterative solvers have been developed to solve a neighbouring problem that is more robust to perturbations on the measurements (see, e.g.[13, 50]). These iterative algorithms solve an optimization problem, refining the solution as they progress and allowing to incorporate prior knowledge via so-called regularization. However, different regularization techniques intrinsically rely on different assumptions on the reconstructed object (e.g. smoothness or appearance of edges), which has a direct impact on the resulting quality.

On top of that, screening requires scanning large portions of the population with harmful radiation. Therefore, taking less measurements while preserving image quality would be desirable. Classical regularization algorithms have been enhanced using data-driven methods, where some of the reconstruction steps are replaced by machine learning models. While these methods have a high success rate in perceived image quality (cf. [2, 65]), the explainability is quite low. Thereto, and to increase applicability, task-adapted reconstruction for inverse problems has been introduced into the modern data-driven pipelines (cf. [1]).

In addition to the described choice of reconstruction algorithm, the image acquisition settings (e.g. mAs and kV) as well as the geometry parameters (e.g. slice thickness) also influence the image quality of the CT reconstructions.

The following three experiments relating to the use of IQA measures in CT are presented: one on the evaluation of Krylov subspace algorithms for cone beam CT (CBCT) reconstruction, another on the evaluation of data-driven methods in lung CT reconstructions, and lastly an example on output deviations with adjusted scanner parameter settings.

Example 1: Krylov Methods in CBCT

The example presented here is taken from a study using Krylov subspace methods, a family of iterative reconstruction algorithms on CBCT data [81]. The study proposed and compared a variety of reconstruction algorithms in simulated and real CBCT problems. Here, we include an experiment involving simulated CBCT acquisitions of a head, where a mixture of Poisson and Gaussian noise is added to the measurement, to simulate realistic noise (cf. [107]). The performance of several Krylov algorithms was determined by comparing the final reconstructions to the ground truth (cf. Fig. 2).

FR-IQA Mismatches

The reconstructions in Fig. 2 g and h contain pixel-wise noise and some undesired stripe artefacts in the lower section of the head, which is not unexpected for reconstructions based on ABBA-GMRES (cf. [42]). In comparison, the other methods do not produce such artefacts and do consist of a more uniform tissue value. However, in Fig. 2, we can see that the computed IQA values do not penalize the loss of detailed information, and in fact, PSNR/LPIPS suggest that the reconstruction in e is significantly worse than the ABBA-GMRES methods g and h, which contradicts the visual perception in these regards. SSIM on the other hand struggles here to penalize blur strong enough and gives the low-quality image in b and a higher rating than h.

Quantitative assessment of novel CBCT reconstruction methods is highly needed and also encouraged to be reported when publishing a novel method. In this example, we can see that the suggested measures do not yield consistent results, and more complex image quality metrics would be required to capture both local and non-local effects appropriately.

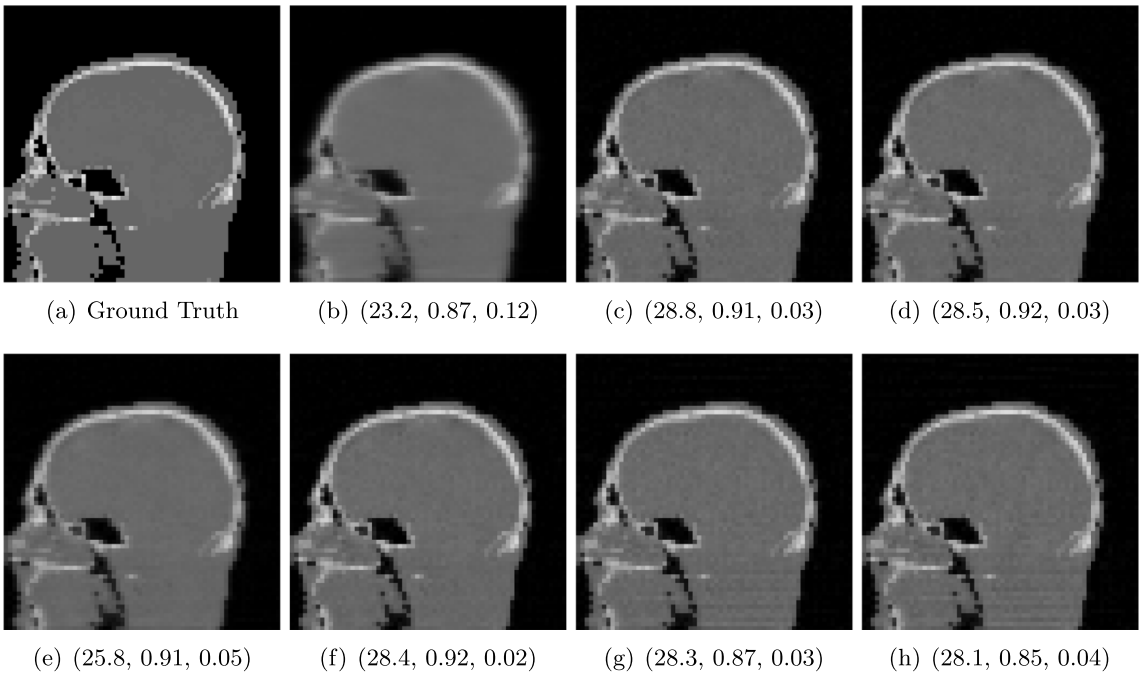


Fig. 2 CBCT reconstructions (**b**–**h**) of phantom head data using different Krylov methods (cf. [81]) and PSNR/SSIM/LPIPS compared to the ground truth (**a**). The overall visual appearance is misjudged here by all three measures, e.g. PSNR in (**g**), SSIM in (**e**) and LPIPS in (**g**)

Example 2: Data-Driven Reconstruction Methods in Lung CT Screening

There is sufficient evidence that screening for certain tumours using CT images may improve the prognosis of cancer survivability [14]. As mentioned above, in order to gain better image quality with less X-ray dose, many enhanced regularization techniques with integrated machine learning steps have been suggested for CT reconstruction, and in a full reference setting, they are commonly evaluated by applying PSNR and SSIM (see, e.g. [2, 45, 98]). As CT images are generally taken to perform a clinical task, they are not the final step of a medical process but often the initial one. Therefore, the definition of what makes a good image heavily depends on the task in hand, and for prognosis-related cancer, the identification of tumours is of upmost importance.

In ongoing research on photon counting detector types and screening procedures for lung cancer (EPSCR grant: EP/W004445/1), an experiment was conducted testing enhanced reconstruction algorithms. Simulations using less than 10% of a clinical X-ray dose were performed to investigate if data-driven methods could sufficiently enhance the images to clearly see the tumours in the lungs while providing a very low amount of dosage to the patients. The corresponding data was a CT-dose simulation, using images from the open LIDC-IDRI dataset [3] as references, as well as simulated and reconstructed images with in-house software. Figure 3 shows the results of the experiment. We show the reference image used as a basis for the simulation, together with five different reconstruction algorithms. The first is an iterative solver, a gradient descend algorithm with TV minimization [90], and c–f correspond to machine learning methods: FBPConvnet is a denoising algorithm that cleans the bad image [48], LPD is an iterative unrolled method that combines traditional solvers with machine learning [2], Noise2Inverse is a self-supervised learning method (i.e. does not require ground truth data) [45] and ItNet is another iterative unrolled method, the best-performing winner of the AAPM DL-Sparse-View CT challenge [33]. ItNet is also judged here as the best result according to PSNR, SSIM and LPIPS.

FR-IQA Mismatches

This experiment was performed to evaluate the quality of different kinds of CT reconstruction, especially the lung tumour detection capabilities thereof. The best result according to the chosen IQA measures is given by ItNet in Fig. 3f, which performs visually poorly. Not only the tumour (zoomed-in white circle) is significantly less visible in the reconstruction, but ItNet also produces structures in the lung that are different than the ones in the reference image; it blurs and lengthens much of the soft tissue present in the lungs and it also created structure from noise in some places. Moreover, the image is overly smooth. Comparing the other reconstruction algorithms, it seems that FBPConvnet Fig. 3c is the one

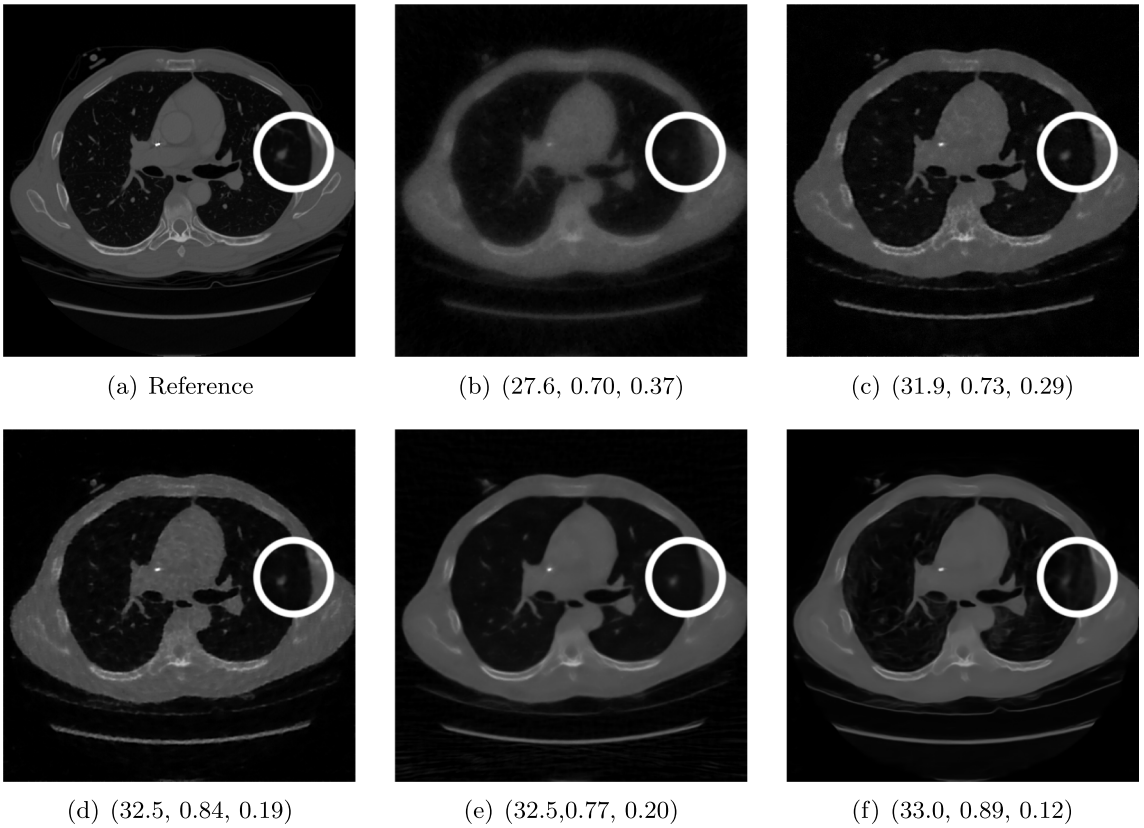


Fig. 3 Reference image (**a**) and outputs of different reconstruction methods (**b–f**) applied to dose simulated data. PSNR/SSIM/LPIPS are unable to identify the best reconstruction (**c**), where also the tumour is visualized well

performing best at preserving the shape of the lung nodule, even when the resulting image contains enhanced pixel-level noise.

We can see here that the qualitative findings strongly contradict the numbers provided by the selected measures. The reconstruction of ItNet, Fig. 3f, outperforms the other reconstructions in regard to the measures, and the qualitative winner FBPConvnet, Fig. 3c, is judged as second worst by the same measures. This experiment suggests that the discussed measures are not a good choice for that kind of CT reconstruction application and are yielding misleading results.

While pixel-independent random noise may be a worse effect in a natural image than a slightly oversmooth reconstruction, this is not true in CT images, where small structures may disappear if smoothing is promoted against edge preservation. In iterative reconstruction algorithms, such choices are explicitly made by choosing the prior appropriately, and in data-driven models, the researcher has limited control on the type of implicit priors the algorithm learns from the data, i.e. model builders do not know what the algorithms choose to learn from the ground truth. In these cases, appropriate evaluation would be even more important to ensure the described quality properties.

Example 3: Scanner Settings Impact in IQ

Changing CT scanner settings, like tube voltage or reconstruction geometry, has a direct impact in the noise distribution of the data and thus in the quality of the reconstructed images. Here, we show an example of quality differences with acquired CT data from a realistic silicone phantom fabricated with multi-material extrusion 3D printing technology [44]. The phantom model was derived from an abdominal CT and was fabricated with realistic radio density values which could mimic imaging properties of soft tissues in CT.

For the reference image, the anatomical phantom was scanned with the standard clinical CT protocol from SOMATOM Definition AS scanner, Siemens Healthineers, Erlangen, Germany (tube current time product 70 mAs for samples and 150 mAs for anatomical phantom, tube voltage 120 kVp, slice thickness 0.60 mm, pixel spacing 0.77 mm, iterative reconstruction kernel J30s). Additional scans with varying kVp values (80/100/120) as well as varying slice thickness (0.6/2 mm) were also performed to assess the effect of the parameters on the image quality. We observed that changing kVp and slice thickness resulted in different image quality, where higher kVp and smaller slice thickness give the best visual result.

FR-IQA Mismatches

Although all IQA measures yield a better value for the image shown in Fig. 4c, a higher visual correspondence with the reference image can be seen in Fig. 4b despite the black shadow in the bottom left corner. The image in Fig. 4c with lower kVp yields a result that is too smooth in comparison to the reference. This yields another CT example where the IQA measures have been misled by quality properties that are not relevant for the clinical application.

MRI

Magnetic resonance imaging (MRI) is a non-invasive medical imaging modality that provides excellent image quality tissue structure without ionizing radiation, but on the other hand is relatively slow. The acquired 3D data, sampled in the k-space domain, corresponds to the Fourier transform of the spatial-domain MR image. To reconstruct an accurate MR image, sampling theory indicates the total number of k-space data that must be acquired to avoid artefacts in the reconstruction. As this number is relatively large and cannot be arbitrarily reduced, shortening the total scan time compromises the image quality [15, 69].

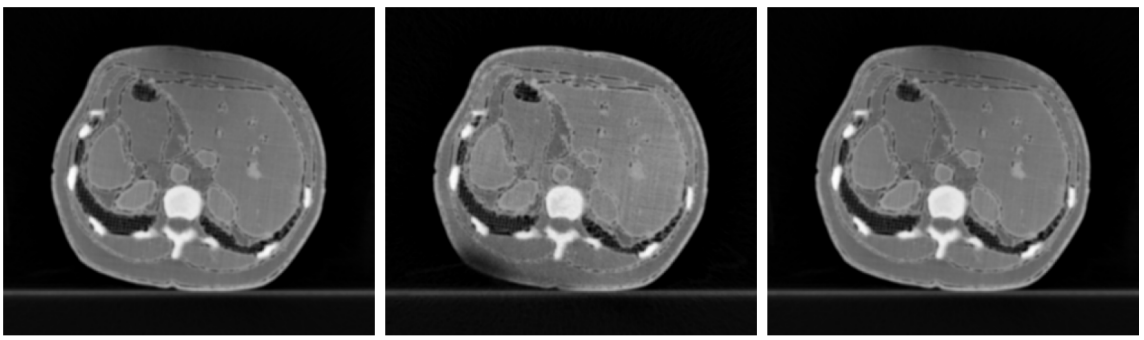
Reconstruction Problem

MRI requires long acquisition times, directly related to the final resolution and tissue contrast. For many clinical applications, faster data acquisition is necessary to minimize the stress on the patient, and moreover, it is important to reduce physiological motion as much as possible since this causes artefacts in the images. In order to fasten acquisition, but still receive reasonable image quality, several approaches have been introduced (see [66]). Most of these techniques acquire less data than theoretically required. To avoid low quality due to less sampled data, techniques such as parallel imaging [37, 75] and compressed sensing [58] have been successfully employed in the past decades. More recently, aiming for even more advancement, machine learning methods have demonstrated promising results. The goal is to achieve a high acceleration factor while preserving the imaging quality. The acceleration factor is given by the ratio of the amount of k-space data required for a fully sampled image to the amount collected in an accelerated acquisition. The outputs of such methods are usually evaluated with PSNR and SSIM (see, e.g. [53, 111]).

Example 1: Scan Acceleration

For this example, the data is obtained from the publicly available fastMRI brain dataset [108], which consists in total of 6405 T1, T2 and FLAIR 3D k-space volumes. The fastMRI challenge series provided MRI datasets to foster the development of accelerated reconstruction algorithms. The series consists of a knee MRI dataset and challenge in 2019 [52], of a brain dataset and challenge in 2020 [64], and of a prostate dataset in 2023 [95]. The winners of the challenges were selected by the comparison of the provided reference images, created by the rSOS of the fully sampled data, to the image outputs of the proposed method via the SSIM, and the highest ranked results were submitted to receive experts' opinions.

We show here images obtained from two machine learning reconstruction algorithms that took part in the fastMRI multi-coil brain dataset challenge in 2020, namely the end-to-end variational network *E2E-VarNet* [92] and *XPDNet* [77]. *XPDNet* was among the top three submissions of the challenge and both algorithms perform very well on the corresponding public leaderboard [68], that allows comparison of algorithms submitted after the challenge deadline. The authors of the *XPDNet* algorithm provided two distinct models for different acceleration factors. Here, we employ



(a) Reference (b) (26.3, 0.88, 0.07) (c) (28.2, 0.95, 0.03)

Fig. 4 Comparison of image acquisition settings, (**a**) reference image with best-chosen parameter setting (0.6 mm and 120 kVp), (**b**) preserves more detail (0.6 mm and 80 kVp) than **c** which is more smoothed (2 mm and 100 kVp). PSNR/SSIM misjudge the visual quality, and LPIPS yields reasonable quality scores here

the neural network provided for acceleration factor 4. The reconstructions in Fig. 5 were obtained by the application of *E2E-VarNet*, Fig. 5 b, c, e, and f, and *XPDNet*, Fig. 5 a and d, to sub-sampled data with random masks (acceleration factor between 1 to 5) in the frequency domain.

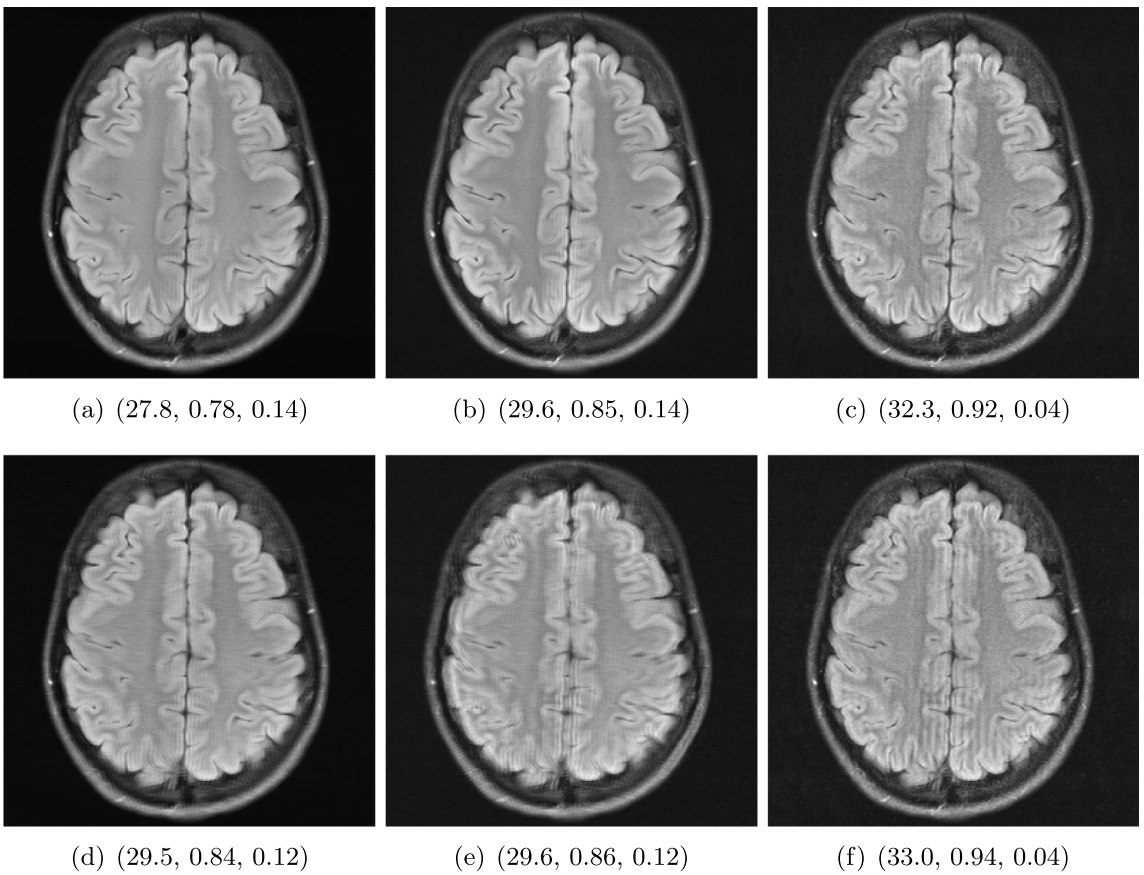
FR-IQA Mismatch

We can see in Fig. 5 that the visual quality of the obtained images does not correspond to the numbers provided by PSNR/SSIM/LPIPS, since the images with better numbers (bottom row) suffer from information loss due to blur and ringing. This is not surprising as some challenges with SSIM as a performance metric have already been discussed and shown in the official results paper of the fastMRI challenge [63]. Here, we complement with examples where the visual results also ask for a different judgement in a non-local manner. Curiously, the degraded images e and f do hold quite higher numbers in comparison to a which is nearly noise-free.

Example 2: Diffusion-Weighted MRI (dMRI)

dMRI is an important MRI technique to study the neural architecture and connectivity of the brain. It is based on obtaining multiple 3-dimensional diffusion-weighted images to investigate the water diffusivity along various directions, being clinically important especially for the investigation of brain disorders (see, e.g. [88]). However, low signal-to-noise ratio and acquisition time limit the spatial resolution of dMRI, and therefore, its usage is currently mainly restricted to medium-to-large white matter structures, whereas very small cortical or sub-cortical regions cannot be traced accurately. To overcome this, several methods for increasing the spatial resolution of dMRI have been introduced (see, e.g. [27, 36, 100]).

Here, we study image data from an acquisition and reconstruction scheme for obtaining high spatial resolution dMRI images using multiple low-resolution images (cf. [67]). The suggested method combines the concepts of compressed sensing and super-resolution to reconstruct high-resolution



(a) (27.8, 0.78, 0.14) (b) (29.6, 0.85, 0.14) (c) (32.3, 0.92, 0.04)

(d) (29.5, 0.84, 0.12) (e) (29.6, 0.86, 0.12) (f) (33.0, 0.94, 0.04)

Fig. 5 Reconstruction outputs of accelerated FLAIR MRI data from the algorithms *Xpdnet*(**a**, **d**) and *E2varnet* (**b**, **c**, **e**, **f**). The bottom images (**d**–**f**) are judged by PSNR/SSIM/LPIPS as better reconstructions than the respective image above them (**a**–**c**), although they contain stronger blur and contain more ringing artefacts

diffusion data while allowing faster scan time. The data is visualized via the fractional anisotropy (FA) measures computed using diffusion tensor imaging [11].

The data from a human subject was acquired from a MGH connectome 3T scanner. Three thick-slice diffusion-weighted imaging (DWI) volumes with voxel size $0.9 \times 0.9 \times 2.7mm^3$, TE/TR = $84/7600ms$ and 60 gradient directions at $b = 2000s/mm^2$. A separate low-resolution isotropic DWI with a spatial resolution of $1.8 \times 1.8 \times 1.8mm^3$ and with 60 gradient directions at $b = 2000s/mm^2$.

The super-resolution image in Fig. 6a, serving as a reference here, was obtained using the super-resolution reconstruction technique that combines multiple thick-slice DWI with all 60 diffusion directions into a high-resolution image (cf. [76]). This technique yields a high-quality image with good detail preservation, but takes a much longer scan time than the standard upsampling method in Fig. 6c, where the FA map of the low-resolution data was up-sampled using *3DSlicer* [29] to the higher resolution.

The image in Fig. 6b (cf. [67]) was obtained using a combined super-resolution reconstruction, compressive sensing, and spatial regularization techniques with thick-slice images, where each thick-slice DWI has a different set of 20 diffusion gradient directions, saving indispensable scan time. The advanced method yields a much higher visual quality image than Fig. 6c, preserving more anatomical details.

FR-IQA Mismatch

We can see in Fig. 6 that PSNR and SSIM misjudge the visual quality of the high-resolution reconstruction in b in comparison to the up-sampled image in c. The image is per default more blurry and does not provide sufficient anatomical details and therefore offers worse visual quality than the reconstruction in b. LPIPS yields more sufficient results in this example and correctly attributes c a higher-quality error.

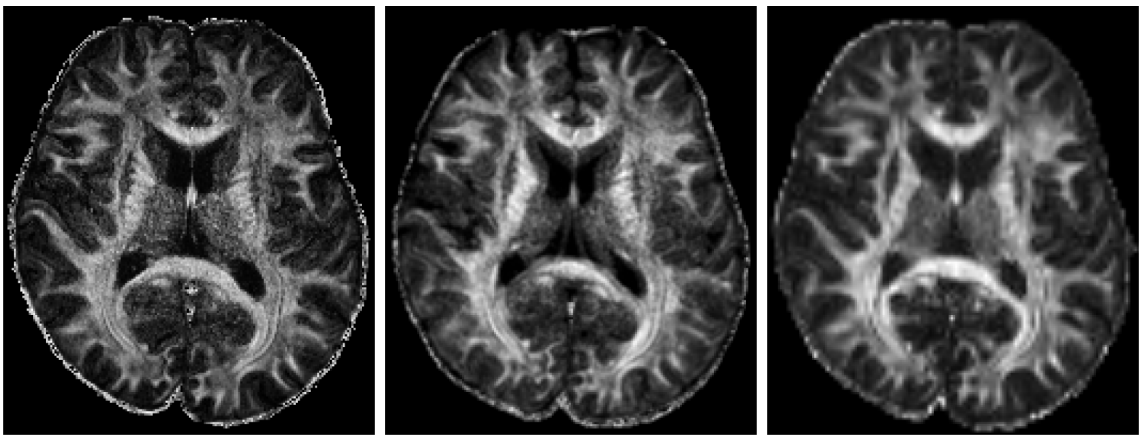
In this example, it has to be noted that the computed IQA numbers are generally quite low, because the resulting FA images do not necessarily have the same range or distribution as the reference image. Therefore, in order to compare the reconstruction quality directly, this task generally benefits from NR-IQA evaluation.

X-ray

X-ray imaging is a fundamental form of radiography. Reducing radiation dose while maintaining image quality is a key principle in radiology known as ALARA (as low as reasonably achievable) [84]. New technologies and imaging techniques, such as post-processing by artificial intelligence (AI) [43], may allow diagnostic objectives to be achieved with lower radiation doses. Furthermore, advancements in X-ray have also the potential to influence and enhance computed tomography (CT) [30]. Whereas CT requires complex imaging reconstruction algorithms, X-ray is more straightforward, employing post-processing for high-quality and detailed imaging, being crucial for clinical assessment of anatomical structures and potential pathologies.

Post-processing Problem

The raw data captured during digital radiography reflects the pattern of X-ray attenuation by different tissues. The digital signal is then processed to create a grey-scale image, where each shade corresponds to the radiodensity of the tissues, ranging from black for air through varying shades of grey for soft tissue and white for bone. Post-processing software refines the raw image to enhance clarity and diagnostic



(a) Reference (b) (17.62, 0.57, 0.19) (c) (18.23, 0.58, 0.26)

Fig. 6 Visualized FA images obtained from diffusion MRI with super-resolution reconstructions. The up-sampled image (**c**) with lower resolution is wrongly judged to have better quality than the high-resolution reconstruction (**b**) by PSNR and SSIM, LPIPS judges this task correctly

utility [86]. This may involve adjusting parameters such as brightness and contrast, applying filters for noise reduction or using algorithms for edge enhancement [54]. The aim is to produce an image that provides the best possible diagnostic information while adhering to the ALARA principle.

Different quality properties may be desired depending on the purpose of the X-ray. For example, when visualizing the lung tissue, adjustments are made to the brightness and contrast to best highlight the anatomy and common abnormalities while minimizing noise. However, noise is less important when aiming to confirm the position of a line, tube, or foreign object. In this case, an image with edge enhancement and adjusted brightness levels may be desirable to amplify the distinction between the dense material of the object and the surrounding soft tissue.

Quality control for the provided default post-processing is usually made by the manufacturers themselves, therefore not accessible to the end user, and may be divergent to the IQ needed for clinical images [97]. After the machine has been placed in the hospital environment, personalized post-processing settings are often determined by subjective visual inspection. Objective evaluation would help to find an optimal post-processing type for visualization, allowing faster and consistent evaluation. Beyond this setting, objective IQA is important for the development and testing of machine learning algorithms on chest X-ray images including super-resolution, denoising or inpainting methods, where PSNR and SSIM are currently the standard choice for quality assessment (see e.g. [47, 60, 94, 96, 103]).

Data

Posteroanterior chest radiographs were acquired on two imaging systems (both Discovery XR656 HD models, GE Healthcare, USA) at Cambridge University Hospitals NHS Trust. Each scanner was being set up in the hospital with different post-processing parameters (chosen by the operating radiologists), which are used here as reference images (see Fig. 7 and 8a). Additional images, serving as real-life examples of lower quality, were produced for each radiographic exposure using multiple different post-processing settings. The post-processing was applied in the hospital directly on the scanner itself by adjusting parameters in the provided framework.

FR-IQA Mismatches

In Fig. 7, contrast deviation and edge enhancement were reduced in b, but increased in c, the noise reduction algorithm was removed in both. The brightness was increased in both images but more so in c, and low-contrast enhancement was removed in b. The result is that b has relatively low contrast in the lungs compared to the reference (a) and radiograph (c). In Fig. 8, edge enhancement has been dramatically increased in b, while the contrast deviation and tissue contrast have been reduced. In c, the brightness, tissue contrast and edge enhancement have been slightly increased. Consequently, b provides low contrast in the lungs with excessively prominent lung markings and vasculature which make it harder to detect abnormalities such as pneumonia.

All of the chosen FR-IQA metrics wrongly judge b as the better image in the first example Fig. 7, and the results in b and c of the second example in Fig. 8 are quite close, where PSNR and SSIM are also providing the wrong order. The tested measures are not suitable to evaluate the quality of data sets with X-ray images that have large variations regarding contrast, luminance and sharpness.

(a) Reference (b) (21.1, 0.90, 0.11) (c) (19.5, 0.88, 0.16)

Fig. 7 Chest X-ray scans with different kinds of post-processing: (**a**) serves as a reference, and (**b**) is wrongly judged as better visualization by PSNR/SSIM/LPIPS

(a) Reference (b) (22.8, 0.97, 0.08) (c) (21.2, 0.96, 0.07)

Fig. 8 Chest X-ray scans with different kinds of post-processing: (**a**) serves as a reference, and (**b**) is wrongly judged as better visualization by PSNR and SSIM; LPIPS gives a slightly worse evaluation for (**b**)

OCT

Optical coherence tomography (OCT) is a well-established imaging technique based on low-coherence interferometry that enables volumetric imaging of biological tissue at high resolution [73]. Light is split into a reference and a sample arm, recombined after being backreflected by a mirror and backscattered at different depths of the sample, in the respective paths. Using Fourier domain OCT [32], a tomographic image of the sample is reconstructed by spectral analysis of the resulting interference patterns. In order to achieve optimal axial resolution, it is essential to match dispersion between the two arms of the interferometer [26]. This can be achieved on the one hand by carefully matching the length of the arms and the corresponding dispersive materials and on the other hand by numerical methods in the reconstruction process [106]. OCT was originally developed for imaging of the retina, and while there are many other applications of OCT available, its highest impact is to this day in ophthalmology, where this technology plays a critical role in correct diagnosis [73].

Reconstruction Problem

Most commonly, OCT processing algorithms compute intensity images from the recorded spectral data. Standard algorithms include a step that performs numerical dispersion compensation. Methods are available which automatically determine the correct dispersion compensation parameter [106], but the stability of automatic numerical dispersion compensation methods under varying imaging conditions is not yet fully understood. Besides the dispersion parameter, there exist also algorithms that provide geometrical corrections within the reconstruction process (see, e.g. [89]), namely the correction of rotation introduced by eye motion and correction of the curvature of the retina. Automatic parameter selection for the geometric corrections is a challenging task which has so far not been addressed, and therefore, these parameters are usually set manually, which is very time-consuming especially in the case of clinical studies where often a large number of patients and imaging locations are included.

Data

We employ image data obtained using an adaptive optics (AO) supported spectral domain OCT system [18], where cross-sectional images (B-scans) were retrieved from two AO-OCT volumes recorded in a young healthy volunteer with a field of view of approximately $4° \times 4°$ (corresponding to $1.4 \times 1.4mm$ on the retina). Different imaging locations and focus settings were considered. One data set was recorded in the fovea with the focus of the imaging beam set to the posterior retina and one data set close to the optic disc with the focus shifted to the anterior retina. An algorithm including dispersion compensation and geometrical corrections was employed [89] for the reconstruction. The reference images, Figs. 9a and 10a, were obtained by manually optimizing over the parameters that define the compensation of dispersion, rotation and curvature, respectively. The examples in Figs. 9 and 10 compare the chosen reference to three sub-optimal reconstructions, where b had a bad choice for the rotation correction parameter, c for the curvature correction parameter and d for the dispersion compensation parameter.

FR-IQA Mismatches

Good dispersion compensation should provide images with a depth resolution that is optimized for the system at hand. In the ophthalmic application of AO-OCT, this high axial resolution allows for the visualization and identification of the different retinal layers [18, 89, 106], different retinal lay-

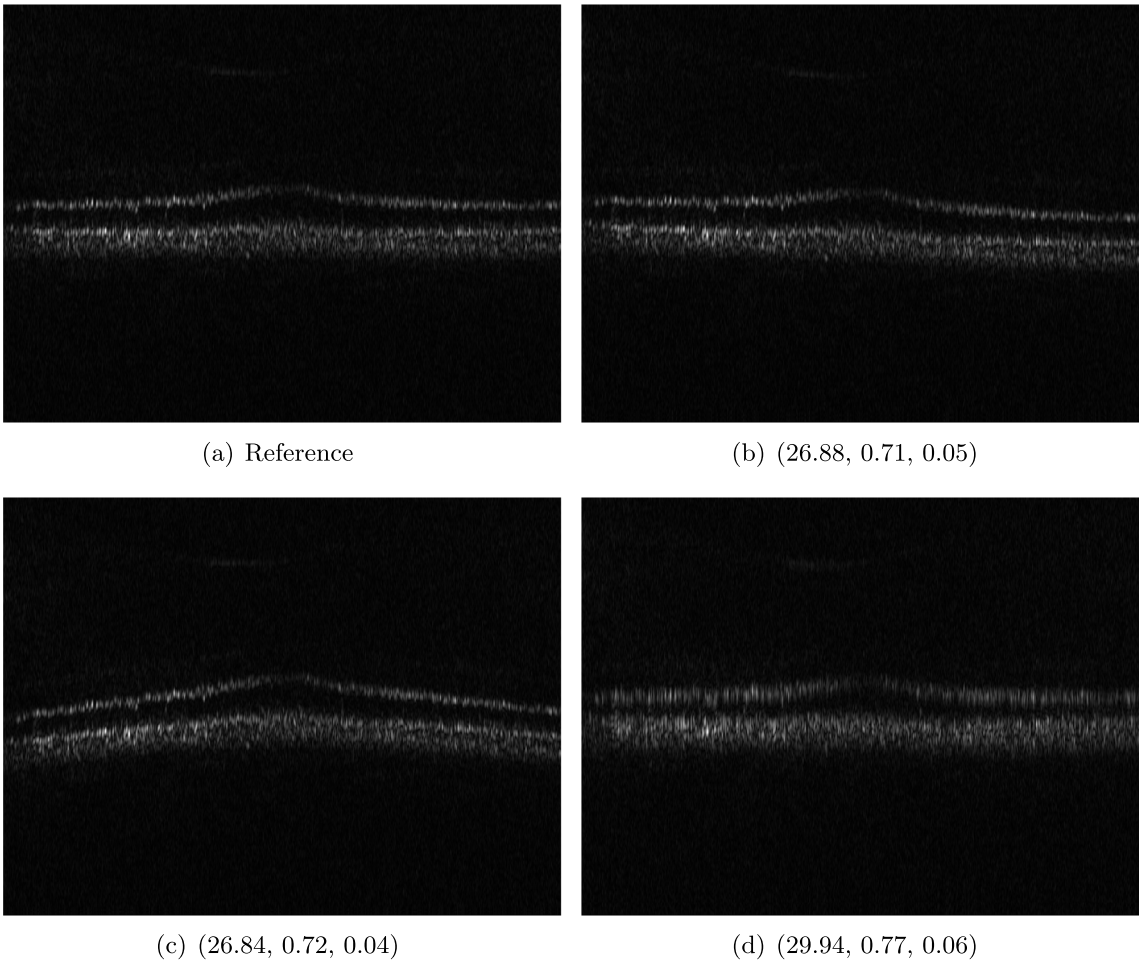


Fig. 9 OCT reference reconstruction (**a**) and reconstructions with sub-optimal parameters (**b**–**d**) leading to geometric deviations (**b**, **d**) and low resolution (**d**). Here, (**d**) is wrongly judged as the best reconstruction by SSIM and PSNR, and LPIPS is able to ignore the small spatial deviations

ers and structures, such as blood vessels. In the first example (Fig. 9), the clear separation of the different layers in the posterior retina, such as the photoreceptor bands and the retinal pigment epithelium, is crucial for clinicians/researchers who investigate the structure and function of the healthy and the diseased human retina [49]. Therefore, in this example, the reconstruction shown in Fig. 2d should have been rated the lowest as the axial resolution is the lowest because of faulty dispersion compensation, which cannot be fixed by further post-processing. In the second example (Fig. 10), a cross-sectional view of three retinal vessels which are embedded in the nerve fibre layer is given. Changes in the thickness of vessel walls are an important early biomarker for retinal diseases such as diabetic retinopathy (cf. [7]). Again, the reconstruction with faulty dispersion compensation shown in Fig. 10d should have been rated the lowest. The loss in axial resolution worsens the visualization of the vessel walls and would lead to inaccurate measurements of the vessel wall thickness.

In the current version of the algorithm, the parameters which determine the amount of dispersion, rotation and curvature compensation applied by the reconstruction algorithm have to be set manually. Therefore, automated evaluation would be very helpful to fasten the process. PSNR and SSIM are not suitable to assess the problem correctly, as they penalize the spatial deviations in b and c strongly. The geometric deviations are not beneficial, but these errors could be corrected in an additional post-processing step unlike the deteriorated axial resolution due to the wrong dispersion compensation parameter (d). LPIPS is able to ignore these small spatial deviations.

Digital Pathology

Digital pathology integrates the acquisition, management and interpretation of pathology information generated from digitalized tissue stainings present on glass slides [46]. The process starts with high-throughput scanning of glass slides

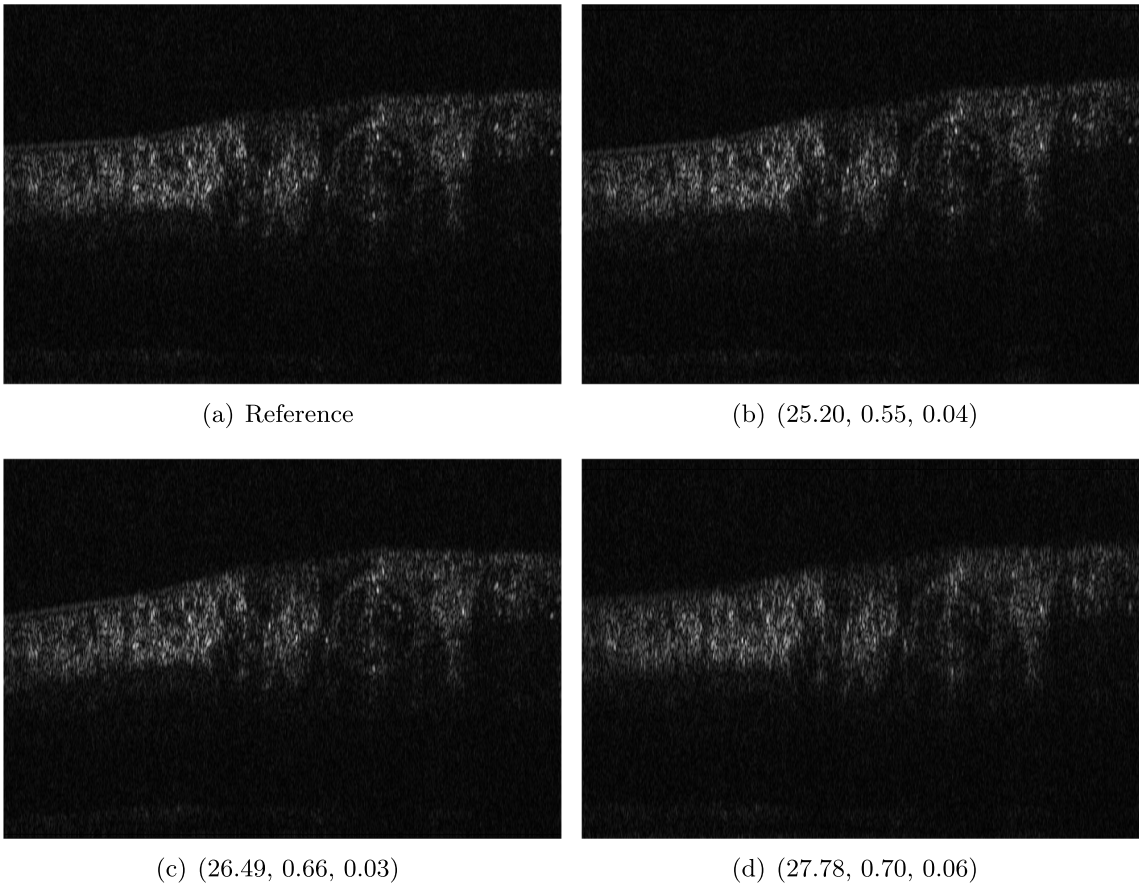


Fig. 10 OCT reference reconstruction (**a**) and reconstructions with sub-optimal parameters (**b**–**d**) leading to geometric deviations (**b**, **c**) and low resolution (**d**). Here, (**d**) is wrongly judged as the best reconstruction by PSNR and SSIM, and LPIPS is able to ignore the small spatial deviations

on dedicated slide scanners. The obtained images can be further used for diagnostics, web-based consultations with other expert pathologists in tumour boards, quantitative analysis and secure archival of pathology data as well as for the development of machine learning tools for tumour classification. However, the application of the listed operations is only valuable upon high image quality and a cost-effective scanning process. Image quality is highly dependent on scanner type and scan settings. Furthermore, image size is considerably large, with one image usually comprising 1–2 GB, such that data storage and data access represent additional challenges for digital pathology. In this light, optimizing the digitization process through the establishment of quantitative criteria for image quality standards would greatly contribute to efficient workflows and manageable image usage.

Visualization Problem

Currently, the slide scanner operators personally set the parameters on the machine to optimize the scanning procedure. At first, the scanning device performs a low-resolution overview scan in an automated manner. The operator is required then to select a dedicated field of imaging for high-resolution imaging with a 40X objective, followed by the application of focus points to the selected scan area. The distribution of these focus points can either be performed in an automated manner by the scanner software (up to 9 focus points), or manually by the operator (if more than 9 focus points are desired). The choice of how many focus points are used is usually based on the operator's personal preference, where setting focus points manually is highly time-consuming. Furthermore, each focus point increases the required scanning time for slide digitization, thus adding to the time investment of pathologists at the machine while at the same time limiting the number of slides that can be scanned within 1 day. However, image quality correlates with the ability of the pathologists to provide an accurate diagnosis to their patients. The diagnosis can have life-impacting consequences, such as the choice of the best treatment options for tumour patients based on tumour subtype classification by the histopathological evaluation of tissue sections. Thus, automated IQA for different choices of focus points would be helpful to design a standardized, cost- and time-effective workflow while providing medical experts with images of reasonable quality for accurate diagnostics.

Data

The data presented in this study have been acquired from digital images of immunohistochemistry stainings that were performed on archival tissue obtained from the neurobiobank of the Division of Neuropathology and Neurochemistry at the Medical University of Vienna. Stainings have been performed according to standard procedures [41, 85]. Figure 11 a–c shows a tumour biopsy of a gliosarcoma patient stained for the astrocyte marker GFAP (brown signal, cytoplasmic localization) and counterstained with Hematoxylin (blue signal, nuclear localization). Figure 3 d and e shows fetal cerebellar tissue stained for the epigenetic mark H3K27me3 (brown signal, nuclear localization) and counterstained with Hematoxylin (blue signal, nuclear localization). The stained sections have been digitalized on a NanoZoomer 2.0-HT digital slide scanner C9600 (Hamamatsu Photonics, Hamamatsu, Japan). The corresponding software NPD.Viewer2 was used to export the scanned images to tiff files. Here, we performed individual scans of a selected imaging area with different numbers of focus points. We chose either 1, 3 or 9 focus points while not changing the spatial settings for the selected field of interest. The image with 9 focus points, allowing the highest resolution, serves as the reference image.

FR-IQA Mismatches

Although the spatial settings for the selected scan area of interest were not changed during the experiment, the physical performance of the scanner showed slight spatial deviations of the selected area between individual scans and thus did not allow for high spatial accuracy during re-scanning processes. PSNR and SSIM fail to correctly assess the images in Fig. 11 as they are very sensitive to that kind of spatial misalignments. Whereas the scan with 3 focus points corresponds much better to the higher-quality reference as the blurred scan with 1 focus point (see b and e versus c and f), both measures incorrectly judge the blurred scan as the better one. This wrong judgement due to a tiny spatial change is very problematic in the respective framework as it is impossible to guarantee completely exact spatial alignment, even if no other settings had been changed during the scanning process. LPIPS, not being so sensitive to small spatial deviations, is able to correctly judge the rank of quality here.

Photoacoustic Data

Photoacoustic imaging (PAI) is an emerging medical imaging modality with important clinical applications such as inflammatory bowel disease and cardiovascular diseases [4, 6]. PAI combines ultrasound with optical imaging to break through the optical diffusion limit, enabling imaging at depths beyond the reach of conventional optical methods. The inherent optical contrast offers valuable insights into tissue composition and function, enabling enhanced visualization of anatomical structures and pathological abnormalities in vascularization [93] and blood oxygenation [38].

Photoacoustic Inverse Problems

The inverse problems of PAI pertain to the task to accurately visualize molecular distributions and determine functional tissue information from acquired PA time series signals [25]. It can be broadly divided into two main components: (1) the reconstruction of images from time series measurements (acoustic inverse problem) and (2) the correction for the nonlinear light fluence (optical inverse problem). Progress has been made towards both inverse problems, but solutions typically involve forward simulations that rely on assumptions that may not hold for a given clinical application [39]. This makes the reconstruction of images from photoacoustic measurements challenging, especially for medical applications, and an active field of research. To ensure a suitable quality of reconstructed PA images, objective measures would be needed.

In the field of PAI, it is extremely difficult to compare new methods to the state of the art, as many algorithms are not available open source and thus have to be reimplemented from scratch. Furthermore, as no standards exist yet, the field uses standard IQA measures, such as PSNR and SSIM, over measures tailored for PA reconstruction problems [5]. Standardized, objective measures are therefore highly needed, especially for clinical applications to ensure high visual quality of the data used for diagnosis and prognosis.

Data

As examples for PA image reconstructions, we show in Fig. 12 reconstructed images containing estimated distributions of the optical absorption coefficient from cross-sectional photoacoustic images of piecewise-constant test objects (phantoms) (cf. [40]). These images were the result of a two-stage process to solve the PA inverse problems. PA data was acquired with a commercial photoacoustic imaging system (MSOT InVision 256-TF, iThera Medical GmbH, Munich, Germany) and processed with three different algorithms (referred to as *Alg1, Alg2, Alg3*). For the visualization and assessment, the outputs of the algorithms were clipped with the reference image's maximum.

Alg1 corrects a reconstructed PA image by using the light fluence obtained from simulations based on the reference measurements. *Alg2* and *Alg3* are deep learning algorithms that are trained to directly estimate the absorption coefficient from the reconstructed image. *Alg2* was trained with simulated data, and *Alg3* was trained with experimental data. The reference absorption coefficients are obtained by using

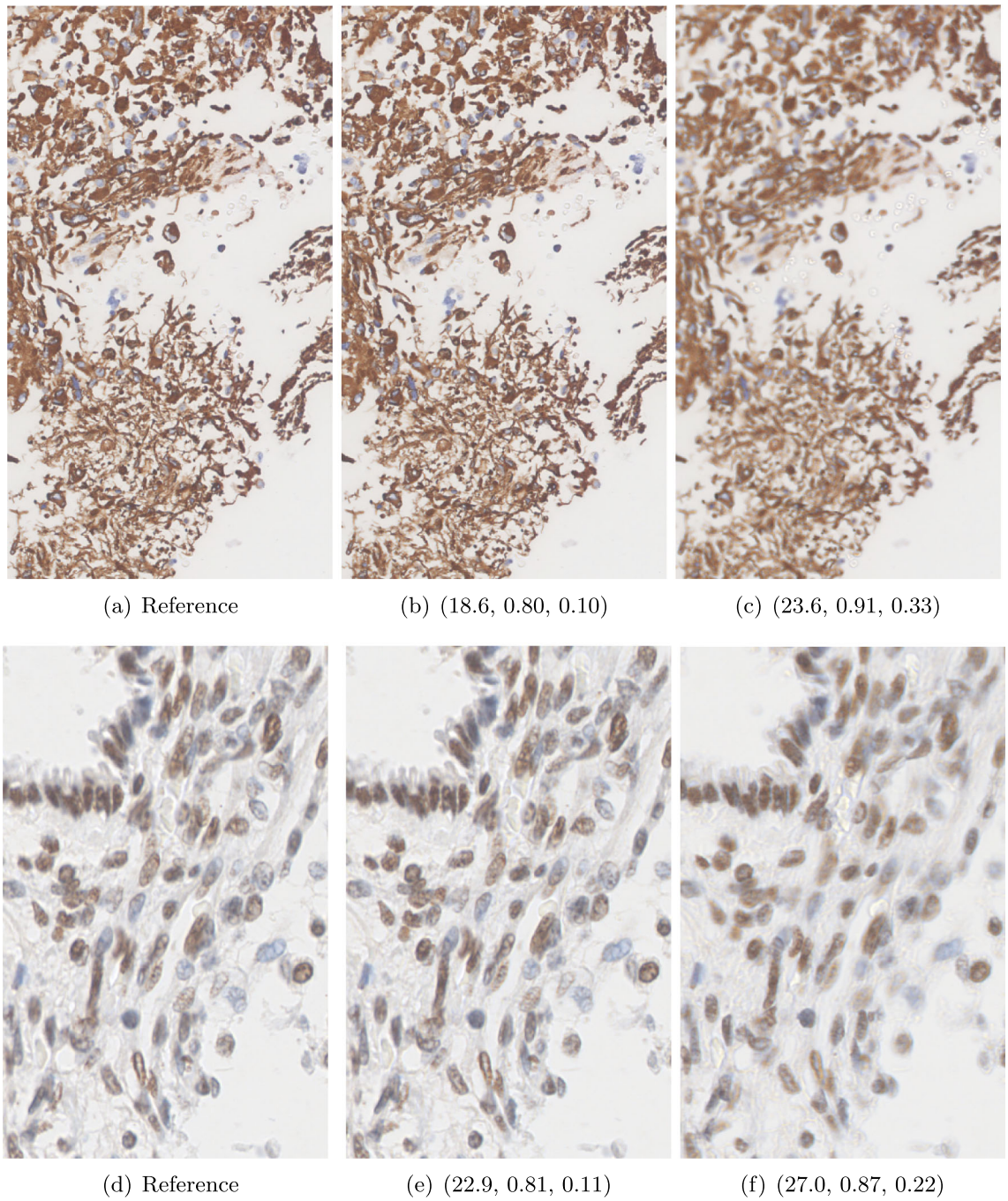


Fig. 11 Data acquired with a slide scanner and 9 (**a**, **d**), 3 (**b**, **e**) and 1 (**c**, **f**) focus points. The image with 9 focus points serves as a reference here. PSNR and SSIM misjudge the tiny spatial misalignment and therefore favour the blurry images with 1 focus point. LPIPS is able to ignore these spatial misalignments

a double-integrating sphere [72] setup as a complementary measurement system, which only yields point estimates for homogeneous material samples. Because of the piecewise-constant nature of the used phantoms, one can fabricate an additional batch of the material used for the test object, measure it and relate the calculated properties to the test object. For any complicated objects or in vivo images, this process would be unfeasible.

FR-IQA Mismatches

Photoacoustic images are typically sparse, which complicates the application of image-based quality measures. In the original paper [40], the qualitative assessment was conducted manually. In that assessment, *Alg3* should perform best, *Alg1* second best and *Alg2* worst across the test images. As this kind of manual assessment is not feasible for larger

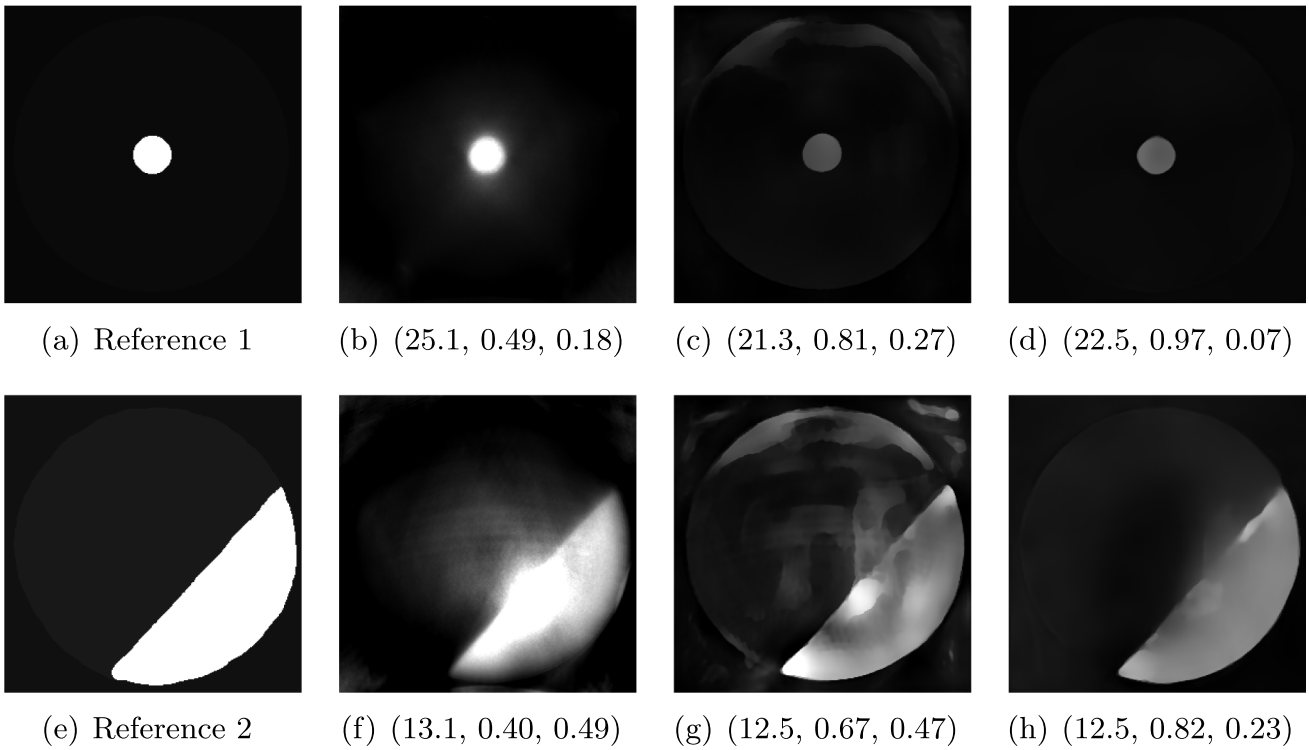


Fig. 12 Two examples of photoacoustic images, ground truth images (**a**, **e**) and comparisons of the reconstructions with *Alg1* (**b**, **f**), *Alg2* (**c**, **g**) and *Alg3* (**d**, **h**). None of the measures was able to assess the relevant quality properties for both examples correctly

data sets and may naturally also introduce inconsistencies as well as biases, automation is needed. However, when PSNR and SSIM are applied to the images, these intended results cannot be reproduced. SSIM wrongly assesses that *Alg1* performs worst throughout both images (Fig. 12), LPIPS for the second example. PSNR wrongly judges *Alg1* as the best rather than *Alg3* and does not give much lower results for *Alg2* than *Alg3*, despite the introduction of artefacts that lead to significant degradation of image quality.

Depending on the goal of the data analysis, desired quality properties may differ here. Artefacts might introduce structures into the image that could be mistaken for regions of interest. On the other hand, an inaccurate estimation of the absolute absorption coefficient might lead to errors in the estimation of functional tissue parameters, such as blood oxygenation, that are important to assess the health status of a patient. Therefore, targeted IQA measures that are indicative of success given a desired use case are required for the objective assessment of algorithms for quantitative PAI and would allow fast advances in the still-evolving field.

Discussion

The results presented in this paper give a collection of examples in diverse medical imaging tasks illustrating the most common pitfalls of the standard FR-IQA measures. Some of the problems associated with these measures have been discussed in isolated studies before. This paper aims to present them in a summarized and structured manner, highlighting the most relevant aspects in the context of medical imaging.

We showed examples of failure when using the standard FR-IQA measures PSNR and SSIM, as well as the more recently introduced LPIPS (based on AlexNet), which provides some beneficial properties for problems occurring in medical imaging (in particular, invariance to small spatial deviations), but still does not show sufficient performance in many tasks. The occurring problems for the employed measures across the medical imaging modalities are manifold and include the following:

- Penalization of task-irrelevant perceptual information (Fig. 4, Fig. 7, Fig. 8 and Fig. 12)
- Misjudgment of broad artefacts leading to information loss (Fig. 2 and Fig. 5)
- Inability to detect local errors and structural details (Fig. 3)
- For PSNR/SSIM, undesired sensitivity to small spatial changes and geometric deviations (Fig. 9, Fig. 10 and Fig. 11)

The evidence in this paper suggests that there is a need for a comprehensive discussion on standard FR-IQA measures and their applicability in medical imaging tasks. Based on the

known struggles of FR-IQA measures and the fact that many non-trivial choices have to be made when applying such measures in medical imaging (this includes the choice of image visualization strategy, which strongly influences the result, e.g. cropping and standardization, as well as implementation details), FR-IQA measures should be acknowledged to not being a straightforward evaluation choice for medical imaging. Moreover, comprehensive reporting about the made choices is needed to ensure reproducibility.

NR-IQA measures are by design tailored to specific problems and therefore can be applied and interpreted in a more direct way. Unfortunately, such measures or biomarkers are often only known within groups of experts and not always easily accessible for non-experts working outside of the medical field. Recently, research in this direction has evolved; see, e.g. the work on NR-IQA measures for different MRI modalities including the contribution of public data with ratings [28]. Other attempts of clinically tailored NR-IQA measures for MRI include [56, 69, 70], and for CT data [20, 23, 102].

Nonetheless, there are medical settings in which FR-IQA measures can be a meaningful addition to NR-IQA measures, e.g. for quantification purposes in reconstruction algorithms or specific perceptual properties that are needed for the subsequent usage. In order to employ perceptual FR evaluation, more extensive research is highly needed that investigates diverse medical imaging modalities and characterizes targets that can be evaluated well. Such extensive studies should be published independently of final applications in research papers introducing novel methods. In order to enable proper evaluation of the suitability of FR-IQA measures for clinical tasks, medical image data has to be shared, together with expert ratings, such as for NR-IQA the recently introduced CT IQA challenge data set [55].

On a related note, a very recent publication [59] provides a comprehensive framework for the choice of imaging analysis metrics for segmentation, detection and classification. Such contributions are very important to bridge the gap between application and methodology through transdisciplinary research, and would also be highly beneficial for IQA measures.

Advancing Task-Informed Medical FR-IQA

From these thoughts, we can derive an expandable list of research directions that help to advance sufficient evaluation with FR-image quality metrics for medical images:

1. Studies on existing FR-IQA measures applied to medical data that identify which quality measures work well, and in which specific settings. The used data and distortions should come from a realistic clinical context, i.e. degradations should be obtained from medical scanners directly or modelled in a realistic manner. The results may include the realization that some standard measures are working well in some specific tasks; these tasks have to be identified and comprehensively discussed.
2. Published medical image data with manual expert ratings for specific tasks. Relevant acquisition information should be added, including which kind of visualization and standardization strategies of the images were used in order to obtain the ratings. Did the experts use the same screens? Were they allowed to change the luminance/contrast during the process?
3. Development of novel task-dependent FR-IQA measures that are tailored and extensively tested for particular medical imaging tasks.
4. Identification of perceptual quality properties that are needed for specific medical imaging tasks, in connection with perceptual metrics.
5. Development of standard frameworks/platforms to annotate and test FR-IQA measures in a medical setting.

It is not possible to have one quality measure that serves it all. The usage of task- and image-dependent well-tested measures, e.g. identified through (1.) and (3.), is indispensable for sufficient evaluation schemes. However, comparability of results is also of major importance when novel methodologies are introduced. Therefore, in addition, more general measures are needed for the broader evaluation of novel methods.

A way to approach this could be the creation of a carefully curated set of basis measures that correspond to different perceptual properties, e.g. luminance, contrast, structure and texture. This idea is related to the SSIM framework, which relies on three properties, but eventually combines these to one number. Identified needed perceptual quality metrics (4.) then give subsets for certain tasks. Such a basis set can be used for imaging problems arising from several acquisition modalities and may also provide deeper insights into bigger data sets by identifying which kind of quality was preserved well and which has not.

Most importantly, measures that have not been assessed for medical imaging at all should not be used for such tasks. In order to allow assessment of measures for medical imaging tasks, available data (2.) is crucial: without shared data, it is not possible to conduct evaluations. A standardized framework (5.) would support the implementation of annotation studies.

Lastly, it should be acknowledged that using perceptual quality measures on acquired medical data is a two-stage problem. Firstly, the question of visualisation has to be tackled, and secondly, the identification of a suitable perceptual measure. Both of the stages are not trivial for medical images.

Guidelines for the Choice, Description and Application of FR-IQA Measures

A set of IQA measures should be reported that includes well-tested task-specific FR- and NR-IQA measures as well as visual quality assessment by experts. The discussion of suitability of the employed IQA measures should be a comprehensive and important part of a method paper. Whenever possible, qualitative evaluation by experts should be included. Moreover, it should not be encouraged to use an IQA measure as a loss function and at the same time as an evaluation metric of a method, as this introduces bias in the optimization/evaluation steps. If not avoidable, the employed measure should be treated and reported as part of the validation but not of the testing phase.

Guidelines—Application of FR-IQA Measures

In this section, we will suggest guidelines for the usage of FR-IQA measures for the output's evaluation of image processing algorithms with reference data.

0. **Suitability of IQA**: Check if IQA is the correct way to evaluate the image, e.g. if the image is the final outcome of a process to be visualized or if is it just an intermediate result. Verify that the value range corresponds to an image that can be visualized.
1 **Locality**: Discuss if the measure should be employed on the entire images or if there are specific regions that would ask for a distinctive evaluation. Identify a reasonable region of interest and cropping, being aware that huge black background areas influence the outcome.
2 **Choice of measure set**: Verify if the IQA measure is properly tested for the intended kinds of images and tasks. If not, check what specific characteristics of the measure are needed for the evaluation task and if there are other more suitable measures that can be employed. Note if the measure was part of the method optimization. If yes, report it separately as a validation measure and add more measures for testing. If possible, add NR-IQA measures that are suitable for the problem and data. If not, specify why.
3 **Discuss and argue your choice**: Identify properties for your task that shall be tackled by the quality measure, including structure, texture, contrast/brightness, shift (in)variance and colour scheme.
4 **State exact implementation of the measure**: If a measure provides a framework with parameters and/or different implementations (e.g. SSIM, see introduction), report which parameters and implementations have been used. State on what kind of data type the measure operates on (e.g. uint, float, etc) and how that fits your data.
5 **Data standardization**: Report conducted data standardization, e.g. how have the images been visualized and what kind of standardization was necessary to visualize the images in a meaningful way. Check if the visualization corresponds to the evaluated FR-IQA quantification. If not, and a perceptual measure was employed, specify why.
6 **Share your code**: Make your method as well as your evaluation schemes publicly available to ensure reproducibility.

Limitations of the Study

In this paper, we have gathered examples of different medical imaging modalities and identified tasks that commonly use or would benefit from using FR-IQA measures. The choice of examples was made based on relevance, but also guided by the available fields of expertise of the people who span the present collaborative network. This collection is by no means complete and could be widely extended.

We included the two most commonly used FR-IQA measures for the evaluation of computational medical imaging tasks (PSNR and SSIM) and focused on exemplifying their failures in such settings. There are many tasks in which these measures perform very well, especially for natural images, but evidence is given here that in various medical imaging tasks more extensive testing regarding suitability is necessary. The novel IQA measure LPIPS was also included because there is an ongoing trend to suggest this measure for medical imaging tasks; the message of this paper is to apply caution and conduct further research regarding its broad applicability. The list of existing FR-IQA measures is long, and it is out of the scope of this paper to provide an extensive evaluation scheme of quality measures. Instead, we want to provide examples of failure, highlighting the need for further research regarding the applicability of IQA measures.

Summary

This study shall serve as a first structured collection of pitfalls in medical imaging tasks when employing the most commonly used FR-IQA measures for evaluation, namely PSNR and SSIM, as well as the more recently introduced deep learning-based measure LPIPS. The collection includes examples analyzed by several experts from real-world medical imaging problems with CT, MRI, X-ray, photoacoustic and pathological image data, where failures may ultimately affect clinical tasks conducted on reconstructed images. Moreover, we formulate guidelines for the application of FR-IQA measures in medical settings and provide suggestions for future research directions. We hope that this

critical review will encourage more researchers to actively participate in the research field of evaluation methods and their constraints, in particular regarding FR-IQA. Concluding, this opens up fantastic opportunities for new impactful interdisciplinary research directions where several research communities can benefit from.

Acknowledgements We also want to acknowledge and thank the members of the AIX-COVNET collaboration: Michael Roberts[1,2], Sören Dittmer[1] Ian Selby[4,5], Anna Breger[1,6], Matthew Thorpe[7], Julian Gilbey[1], Jonathan R. Weir-McCall[4,5,8], Judith Babar[4,5], Effrossyni Gkrania-Klotsas[2,4], Jacobus Preller[2], Lorena Escudero Sánchez[5,9], Anna Korhonen[10] Emily Jefferson[11], Georg Langs[12], Helmut Prosch[12], Guang Yang[13], Xiaodan Xing[13], Yang Nan[13], Ming Li[13], Jan Stanczuk[1], Jing Tang[14], Tolou Shadbahr[14], Philip Teare[15], Mishal Patel[15,16], Marcel Wassink[17], Markus Holzer[17], Eduardo González Solares[18], Nicholas Walton[18], Pietro Lió[19], James H. F. Rudd[2,4], John A.D. Aston[15], Evis Sala[5,9,21,22] and Carola-Bibiane Schönlieb[1].
[1]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK; [2]School of Clinical Medicine, University of Cambridge, Cambridge, UK; [3]ZeTeM, University of Bremen, Bremen, Germany; [4]Addenbrooke's Hospital, Cambridge University Hospitals NHS Trust, Cambridge, UK; [5]Department of Radiology, University of Cambridge, Cambridge, UK; [6]Center of Medical Physics and Biomedical Engineering, Medical University of Vienna, Austria; [7]Department of Mathematics, University of Manchester, Manchester, UK; [8]Royal Papworth Hospital, Cambridge, Royal Papworth Hospital NHS Foundation Trust, Cambridge, UK; [9]CRUK Cambridge Institute, Cambridge, UK; [10]Language Technology Laboratory, University of Cambridge, Cambridge, UK; [11]Population Health and Genomics, School of Medicine, University of Dundee, Dundee, UK; [12]Department of Biomedical Imaging and Image-guided Therapy, Computational Imaging Research Lab Medical University of Vienna, Vienna, Austria; [13]National Heart and Lung Institute, Imperial College London, London, UK; [14]Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Helsinki, Finland; [15]Data Science & Artificial Intelligence, AstraZeneca, Cambridge, UK; [16]Clinical Pharmacology & Safety Sciences, AstraZeneca, Cambridge, UK; [17]contextflow GmbH, Vienna, Austria; [18]Institute of Astronomy, University of Cambridge, Cambridge, UK; [19]Department of Computer Science and Technology, University of Cambridge, Cambridge, UK; [20]Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge, UK; [21]Advanced Radiodiagnostics Centre, Fondazione Policlinico Universitario Agostino Gemelli, Rome, Italy; [22]Università Cattolica del Sacro Cuore, Rome, Italy.

Funding The authors wish to acknowledge support from the EU/EFPIA Innovative Medicines Initiative 2 Joint Undertaking - DRAGON (101005122) (A.Br., I.S., M.R., S.D. C.B.S., AIX-COVNET); the Austrian Science Fund (FWF) through project T1307 (A.Br., C.K.) and P33217 (A.Br.); and through SFB 10.55776/F68,'Tomography Across The Scales', project F6807-N36 (E.B.) and project V1041 (N.A.); the OEAW/L'oreal Austria through the fellowship 'FOR WOMEN IN SCIENCE' (A.Br.); the U.S. National Science Foundation through grant DMS-2208294 (M.S.L); the Accelerate Programme for Scientific Discovery and EPRSC grant EP/W004445/1 (A.Bi.); the German Research Foundation through grant GR 5824/1 (J.G.); the National Institute for Health and Care Research (NIHR) Cambridge Biomedical Research Centre (BRC-1215-20014) (I.S.); the EPSRC Cambridge Mathematics of Information in Healthcare Hub EP/T017961/1 (M.R., C.B.S.); and the Trinity Challenge BloodCounts! project (M.R., C.B.S.). C.B.S additionally acknowledges support from the Philip Leverhulme Prize, the Royal SocietyWolfson Fellowship, the EPSRC advanced career fellowship EP/V029428/1, the EP-SRC programme grant EP/V026259/1, and the EPSRC grants EP/S026045/1 and EP/T003553/1, EP/N014588/1, the Wellcome Innovator Awards 215733/Z/19/Z and 221633/Z/20/Z, the European Union Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreementNo. 777826 NoMADS, the Cantab Capital Institute for the Mathematics of Information and the Alan Turing Institute. This research was supported by the NIHR Cambridge Biomedical Research Centre (NIHR203312).

Data Availability Not applicable.

Declarations

Ethics Approval and Consent to Participate The Brent Research Ethics Committee provided ethical approval for our retrospective X-ray study (IRAS ID: 282705, REC No.: 20/HRA/2504, R&D No.: A095585). Informed consent was not required as data was pseudonymized.

Conflict of Interest The authors declare no competing interests.

References

1. Adler, Lunz, Verdier, Schönlieb, and Ozan Öktem. Task adapted reconstruction for inverse problems. *Inverse Problems*, 38, 2022.
2. Jonas Adler and Ozan Öktem. Learned primal-dual reconstruction. *IEEE Transactions on Medical Imaging*, 37 (6):1322–1332, 2018. https://doi.org/10.1109/TMI.2018.2799231.
3. Samuel G 3rd Armato, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, Ella A Kazerooni, Heber MacMahon, Edwin J R Van Beeke, David Yankelevitz, Alberto M Biancardi, Peyton H Bland, Matthew S Brown, Roger M Engelmann, Gary E Laderach, Daniel Max, Richard C Pais, David P Y Qing, Rachael Y Roberts, Amanda R Smith, Adam Starkey, Poonam Batrah, Philip Caligiuri, Ali Farooqi, Gregory W Gladish, C Matilda Jude, Reginald F Munden, Iva Petkovska, Leslie E Quint, Lawrence H Schwartz, Baskaran Sundaram, Lori E Dodd, Charles Fenimore, David Gur, Nicholas Petrick, John Freymann, Justin Kirby, Brian Hughes, Alessi Vande Casteele, Sangeeta Gupte, Maha Sallamm, Michael D Heath, Michael H Kuhn, Ekta Dharaiya, Richard Burns, David S Fryd, Marcos Salganicoff, Vikram Anand, Uri Shreter, Stephen Vastagh, and Barbara Y Croft. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Med Phys*, 38(2):915–931, Feb 2011. ISSN 0094-2405

(Print); 0094-2405 (Electronic); 0094-2405 (Linking)https://doi.org/10.1118/1.3528204.

4. Hisham Assi, Rui Cao, Madhura Castelino, Ben Cox, Fiona J. Gilbert, Janek Gröhl, Kurinchi Gurusamy, Lina Hacker, Aoife M. Ivory, James Joseph, Ferdinand Knieling, Martin J. Leahy, Ledia Lilaj, Srirang Manohar, Igor Meglinski, Carmel Moran, Andrea Murray, Alexander A. Oraevsky, Mark D. Pagel, Manojit Pramanik, Jason Raymond, Mithun Kuniyil Ajith Singh, William C. Vogt, Lihong Wang, Shufan Yang, Members of IPASC, and Sarah E. Bohndiek. A review of a strategic roadmapping exercise to advance clinical translation of photoacoustic imaging: From current barriers to future adoption. *Photoacoustics*, 32:100539, 2023a. ISSN 2213-5979. https://www.sciencedirect.com/science/article/pii/S2213597923000927.
5. Hisham Assi, Rui Cao, Madhura Castelino, Ben Cox, Fiona J. Gilbert, Janek Gröhl, Kurinchi Gurusamy, Lina Hacker, Aoife M. Ivory, James Joseph, Ferdinand Knieling, Martin J. Leahy, Ledia Lilaj, Srirang Manohar, Igor Meglinski, Carmel Moran, Andrea Murray, Alexander A. Oraevsky, Mark D. Pagel, Manojit Pramanik, Jason Raymond, Mithun Kuniyil Ajith Singh, William C. Vogt, Lihong Wang, Shufan Yang, Members of IPASC, and Sarah E. Bohndiek. A review of a strategic roadmapping exercise to advance clinical translation of photoacoustic imaging: From current barriers to future adoption. *Photoacoustics*, 32:100539, 2023b. ISSN 2213-5979. https://www.sciencedirect.com/science/article/pii/S2213597923000927.
6. Amalina Binte Ebrahim Attia, Ghayathri Balasundaram, Mohesh Moothanchery, U S Dinish, Renzhe Bi, Vasilis Ntziachristos, and Malini Olivo. A review of clinical photoacoustic imaging: Current and future trends. *Photoacoustics*, 16:100144, Dec 2019. ISSN 2213-5979 (Print); 2213-5979 (Electronic); 2213-5979 (Linking)https://doi.org/10.1016/j.pacs.2019.100144.
7. Elise Bakker, Felix Anne Dikland, Roan van Bakel, Danilo Andrade De Jesus, Luisa Sánchez Brea, Stefan Klein, Theo van Walsum, Florence Rossant, Daniela Castro Farías, Kate Grieve, and Michel Paques. Adaptive optics ophthalmoscopy: a systematic review of vascular biomarkers. *Survey of Ophthalmology*, 67(2):369–387, 2022. ISSN 0039-6257. https://doi.org/10.1016/j.survophthal.2021.05.012. https://www.sciencedirect.com/science/article/pii/S0039625721001363.
8. Harrison H Barrett. Objective assessment of image quality: effects of quantum noise and object variability. *JOSA A*, 7(7):1266–1278, 1990.
9. Harrison H Barrett, JL Denny, Robert F Wagner, and Kyle J Myers. Objective assessment of image quality. ii. fisher information, fourier crosstalk, and figures of merit for task performance. *JOSA A*, 12(5):834–852, 1995.
10. Harrison H Barrett, Craig K Abbey, and Eric Clarkson. Objective assessment of image quality. iii. roc metrics, ideal observers, and likelihood-generating functions. *JOSA A*, 15(6):1520–1535, 1998.
11. P J Basser, J Mattiello, and D LeBihan. Estimation of the effective self-diffusion tensor from the nmr spin echo. *J Magn Reson B*, 103(3):247–254, Mar 1994. ISSN 1064-1866 (Print); 1064-1866 (Linking). https://doi.org/10.1006/jmrb.1994.1037.
12. Marcel Beister, Daniel Kolditz, and Willi A Kalender. Iterative reconstruction methods in x-ray ct. *Phys Med*, 28(2):94–108, Apr 2012. ISSN 1724-191X (Electronic); 1120-1797 (Linking). https://doi.org/10.1016/j.ejmp.2012.01.003.
13. Ander Biguri, Manjit Dosanjh, Steven Hancock, and Manuchehr Soleimani. Tigre: a matlab-gpu toolbox for cbct image reconstruction. *Biomedical Physics & Engineering Express*, 2 (5):055010, 2016.
14. Phillip M Boiselle. Computed tomography screening for lung cancer. *JAMA*, 309(11):1163–1170, Mar 2013. ISSN 1538-3598 (Electronic); 0098-7484 (Linking). https://doi.org/10.1001/jama.2012.216988.
15. Anna Breger, Gabriel Ramos Llorden, Gonzalo Vegas Sanchez-Ferrero, W. Scott Hoge, Martin Ehler, and Carl-Fredrik Westin. Orthogonal projections for image quality analyses applied to mri. *PAMM*, 20(1):e202000159, 2021. https://doi.org/10.1002/pamm.202000159. https://onlinelibrary.wiley.com/doi/abs/10.1002/pamm.202000159.
16. Anna Breger, Janek Gröhl, and Thomas Else. Photoacoustic data annotations supplementing the paper: "a study on the adequacy of common iqa measures for medical images", September 2024a. https://doi.org/10.5281/zenodo.13325197.
17. Anna Breger, Clemens Karner, Ian Selby, Janek Gröhl, Sören Dittmer, Edward Lilley, Judith Babar, Jake Beckford, Timothy J Sadler, Shahab Shahipasand, Arthikkaa Thavakumar, Michael Roberts, and Carola-Bibiane Schönlieb. A study on the adequacy of common iqa measures for medical images. In *Springer Lecture Notes in Electrical Engineering*, page accepted, 2024b.
18. Elisabeth Brunner, Julia Shatokhina, Muhammad Faizan Shirazi, Wolfgang Drexler, Rainer Leitgeb, Andreas Pollreisz, Christoph K. Hitzenberger, Ronny Ramlau, and Michael Pircher. Retinal adaptive optics imaging with a pyramid wavefront sensor. *Biomed. Opt. Express*, 12(10):5969–5990, Oct 2021. https://doi.org/10.1364/BOE.438915. https://opg.optica.org/boe/abstract.cfm?URI=boe-12-10-5969.
19. Jianmei Cai, Xiaogang Chen, Wuhao Huang, and Xuanqin Mou. Image quality assessment on ct reconstruction images: Task-specific vs. general quality assessment. *Fully3D*, 2017.
20. Yuan Cheng, Ehsan Abadi, Taylor Brunton Smith, Francesco Ria, Mathias Meyer, Daniele Marin, and Ehsan Samei. Validation of algorithmic ct image quality metrics with preferences of radiologists. *Med Phys*, 46(11):4837–4846, Nov 2019. ISSN 2473-4209 (Electronic); 0094-2405 (Linking). https://doi.org/10.1002/mp.13795.
21. Li Sze Chow and Raveendran Paramesran. Review of medical image quality assessment. *Biomedical Signal Processing and Control*, 27: 145–154, 2016. ISSN 1746-8094. https://doi.org/10.1016/j.bspc.2016.02.006. https://www.sciencedirect.com/science/article/pii/S1746809416300180.
22. Li Sze Chow and Heshalini Rajagopal. Modified-brisque as no reference image quality assessment for structural mr images. *Magn Reson Imaging*, 43:74–87, Nov 2017. ISSN 1873-5894 (Electronic); 0730-725X (Linking). https://doi.org/10.1016/j.mri.2017.07.016.
23. Minsoo Chun, Jin Hwa Choi, Sihwan Kim, Chulkyun Ahn, and Jong Hyo Kim. Fully automated image quality evaluation on patient ct: Multi-vendor and multi-reconstruction study. *PLoS One*, 17(7):e0271724, 2022. ISSN 1932-6203 (Electronic); 1932-6203 (Linking). https://doi.org/10.1371/journal.pone.0271724.
24. Eric Clarkson, Matthew A Kupinski, Harrison H Barrett, and Lars Furenlid. A task-based approach to adaptive and multimodality imaging. *Proceedings of the IEEE*, 96(3):500–511, 2008.
25. Ben Cox, Jan G Laufer, Simon R Arridge, and Paul C Beard. Quantitative spectroscopic photoacoustic imaging: a review. *J Biomed Opt*, 17(6):061202, Jun 2012. ISSN 1560-2281 (Electronic); 1083-3668 (Linking). https://doi.org/10.1117/1.JBO.17.6.061202.
26. W. Drexler, Y. Chen, A. Aguirre, B. Považay, A. Unterhuber, and J. G. Fujimoto. *Ultrahigh Resolution Optical Coherence Tomography*, pages 239–279. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-77550-8. https://doi.org/10.1007/978-3-540-77550-8_8.
27. Tim B. Dyrby, Henrik Lundell, Mark W. Burke, Nina L. Reislev, Olaf B. Paulson, Maurice Ptito, and Hartwig R. Siebner. Interpolation of diffusion weighted imaging datasets. *NeuroImage*, 103:202–213, 2014. ISSN 1053-8119. https://doi.org/

10.1016/j.neuroimage.2014.09.005. https://www.sciencedirect.com/science/article/pii/S1053811914007472.
28. Oscar Esteban, Ross W. Blair, Dylan M. Nielson, Jan C. Varada, Sean Marrett, Adam G. Thomas, Russell A. Poldrack, and Krzysztof J. Gorgolewski. Crowdsourced mri quality metrics and expert quality annotations for training of humans and machines. *Scientific Data*, 6(1):30, 2019. https://doi.org/10.1038/s41597-019-0035-4.
29. Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, John Buatti, Stephen Aylward, James V Miller, Steve Pieper, and Ron Kikinis. 3d slicer as an image computing platform for the quantitative imaging network. *Magn Reson Imaging*, 30(9):1323–1341, Nov 2012. ISSN 1873-5894 (Electronic); 0730-725X (Print); 0730-725X (Linking). https://doi.org/10.1016/j.mri.2012.05.001.
30. Joelle A. Feghali, Greg Chambers, Julie Delépierre, Sophie Chapeliere, Inès Mannes, and Catherine Adamsbaum. New image quality and dose reduction technique for pediatric digital radiography. *Diagnostic and Interventional Imaging*, 102 (7):463–470, 2021. ISSN 2211-5684. https://doi.org/10.1016/j.diii.2021.01.009. https://www.sciencedirect.com/science/article/pii/S2211568421000309.
31. Xuan Fei, Liang Xiao, Yubao Sun, and Zhihui Wei. Perceptual image quality assessment based on structural similarity and visual masking. *Signal Processing: Image Communication*, 27 (7):772–783, 2012. ISSN 0923-5965. https://doi.org/10.1016/j.image.2012.04.005. https://www.sciencedirect.com/science/article/pii/S092359651200094X.
32. A.F. Fercher, C.K. Hitzenberger, G. Kamp, and S.Y. El-Zaiat. Measurement of intraocular distances by backscattering spectral interferometry. *Optics Communications*, 117(1):43–48, 1995. ISSN 0030-4018. https://doi.org/10.1016/0030-4018(95)00119-S. https://www.sciencedirect.com/science/article/pii/003040189500119S.
33. Martin Genzel, Jan Macdonald, and Maximilian März. Aapm dl-sparse-view ct challenge submission report: Designing an iterative network for fanbeam-ct with unknown geometry. arXiv:2106.00280, 2021.
34. B. Girod. Psychovisual aspects of image processing: What's wrong with mean squared error? In *Proceedings of the Seventh Workshop on Multidimensional Signal Processing*, pages P.2–P.2, 1991. https://doi.org/10.1109/MDSP.1991.639240.
35. Daniel Gourdeau, Simon Duchesne, and Louis Archambault. On the proper use of structural similarity for the robust evaluation of medical image synthesis models. *Med Phys*, 49(4):2462–2474, Apr 2022. ISSN 2473-4209 (Electronic); 0094-2405 (Linkin). https://doi.org/10.1002/mp.15514.
36. S M Grieve and J J Maller. High-resolution diffusion imaging: ready to become more than just a research tool in psychiatry? *Molecular Psychiatry*, 22(8):1082–1084, 2017. https://doi.org/10.1038/mp.2016.170.
37. Mark A Griswold, Peter M Jakob, Robin M Heidemann, Mathias Nittka, Vladimir Jellus, Jianmin Wang, Berthold Kiefer, and Axel Haase. Generalized autocalibrating partially parallel acquisitions (grappa). *Magn Reson Med*, 47(6):1202–1210, Jun 2002. ISSN 0740-3194 (Print); 0740-3194 (Linking). https://doi.org/10.1002/mrm.10171.
38. Janek Gröhl, Thomas Kirchner, Tim J. Adler, Lina Hacker, Niklas Holzwarth, Adrián Hernández-Aguilera, Mildred A. Herrera, Edgar Santos, Sarah E. Bohndiek, and Lena Maier-Hein. Learned spectral decoloring enables photoacoustic oximetry. *Scientific Reports*, 11(1):6565, 2021a. https://doi.org/10.1038/s41598-021-83405-8.
39. Janek Gröhl, Melanie Schellenberg, Kris Dreher, and Lena Maier-Hein. Deep learning for biomedical photoacoustic imaging: A review. *Photoacoustics*, 22:100241, 2021b. ISSN 2213-5979. https://doi.org/10.1016/j.pacs.2021.100241. https://www.sciencedirect.com/science/article/pii/S2213597921000033.
40. Janek Grohl, Thomas R Else, Lina Hacker, Ellie V Bunce, Paul W Sweeney, and Sarah E Bohndiek. Moving beyond simulation: data-driven quantitative photoacoustic imaging using tissue-mimicking phantoms. *IEEE Trans Med Imaging*, PP, Nov 2023. ISSN 1558-254X (Electronic); 0278-0062 (Linking). https://doi.org/10.1109/TMI.2023.3331198.
41. Yong Guo, Verena Endmayr, Anastasia Zekeridou, Andrew McKeon, Frank Leypoldt, Katharina Hess, Alicja Kalinowska-Lyszczarz, Andrea Klang, Akos Pakozdy, Elisabeth Höftberger, Simon Hametner, Carmen Haider, Désirée De Simoni, Sönke Peters, Ellen Gelpi, Christoph Röcken, Stefan Oberndorfer, Hans Lassmann, Claudia F Lucchinetti, and Romana Höftberger. New insights into neuropathology and pathogenesis of autoimmune glial fibrillary acidic protein meningoencephalomyelitis. *Acta Neuropathol*, 147(1):31, Feb 2024. ISSN 1432-0533 (Electronic); 0001-6322 (Print); 0001-6322 (Linking).https://doi.org/10.1007/s00401-023-02678-7.
42. Per Christian Hansen, Ken Hayami, and Keiichi Morikuni. Gmres methods for tomographic reconstruction with an unmatched back projector. *Journal of Computational and Applied Mathematics*, 413:114352, 2022.
43. Hugh Harvey and Eric J Topol. More than meets the ai: refining image acquisition and resolution. *The Lancet*, 396(10261):1479, 2020. ISSN 0140-6736. https://doi.org/10.1016/S0140-6736(20)32284-4. https://www.sciencedirect.com/science/article/pii/S0140673620322844.
44. Sepideh Hatamikia, Laszlo Jaksa, Gernot Kronreif, Wolfgang Birkfellner, Joachim Kettenbach, Martin Buschmann, and Andrea Lorenz. Silicone phantoms fabricated with multi-material extrusion 3d printing technology mimicking imaging properties of soft tissues in ct. *Zeitschrift für Medizinische Physik*, 2023. ISSN 0939-3889. https://doi.org/10.1016/j.zemedi.2023.05.007. https://www.sciencedirect.com/science/article/pii/S0939388923000764.
45. Allard Adriaan Hendriksen, Daniël Maria Pelt, and K. Joost Batenburg. Noise2inverse: Self-supervised deep convolutional denoising for tomography. *IEEE Transactions on Computational Imaging*, 6: 1320–1335, 2020. https://doi.org/10.1109/TCI.2020.3019647.
46. Stephan W Jahn, Markus Plass, and Farid Moinfar. Digital pathology: Advantages, limitations and emerging perspectives. *J Clin Med*, 9(11), Nov 2020. ISSN 2077-0383 (Print); 2077-0383 (Electronic); 2077-0383 (Linking). https://doi.org/10.3390/jcm9113697.
47. Xiaoben Jiang, Yu Zhu, Bingbing Zheng, and Dawei Yang. Images denoising for covid-19 chest x-ray based on multi-resolution parallel residual cnn. *Mach Vis Appl*, 32(4):100, 2021. ISSN 0932-8092 (Print); 1432-1769 (Electronic); 0932-8092 (Linking). https://doi.org/10.1007/s00138-021-01224-3.
48. Kyong Hwan Jin, Michael T. McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26 (9):4509–4522, 2017. https://doi.org/10.1109/TIP.2017.2713099.
49. Ravi S. Jonnal, Omer P. Kocaoglu, Robert J. Zawadzki, Sang-Hyuck Lee, John S. Werner, and Donald T. Miller. The Cellular Origins of the Outer Retinal Bands in Optical Coherence Tomography Images. *Investigative Ophthalmology & Visual Science*, 55 (12):7904–7918, 12 2014. ISSN 1552-5783. https://doi.org/10.1167/iovs.14-14907.

50. Barbara Kaltenbacher, Andreas Neubauer, and Otmar Scherzer. *Iterative Regularization Methods for Nonlinear Ill-Posed Problems*. De Gruyter, Berlin, New York, 2008. ISBN 9783110208276. https://doi.org/10.1515/9783110208276.
51. Sergey Kastryulin, Jamil Zakirov, Nicola Pezzotti, and Dmitry V. Dylov. Image quality assessment for magnetic resonance imaging. *IEEE Access*, 11:14154–14168, 2023. https://doi.org/10.1109/ACCESS.2023.3243466.
52. Florian Knoll, Tullie Murrell, Anuroop Sriram, Nafissa Yakubova, Jure Zbontar, Michael Rabbat, Aaron Defazio, Matthew J Muckley, Daniel K Sodickson, C Lawrence Zitnick, et al. Advancing machine learning for mr image reconstruction with an open competition: Overview of the 2019 fastmri challenge. *Magnetic resonance in medicine*, 84(6): 3054–3070, 2020.
53. Yilmaz Korkmaz, Tolga Cukur, and Vishal M. Patel. Self-supervised mri reconstruction with unrolled diffusion models. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 491–501, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43999-5.
54. Elizabeth A Krupinski, Mark B Williams, Katherine Andriole, Keith J Strauss, Kimberly Applegate, Margaret Wyatt, Sandra Bjork, and J Anthony Seibert. Digital radiography image quality: image processing and display. *J Am Coll Radiol*, 4(6):389–400, Jun 2007. ISSN 1558-349X (Electronic); 1546-1440 (Linking). https://doi.org/10.1016/j.jacr.2007.02.001.
55. Wonkyeong Lee, Fabian Wagner, and et al. Low-dose computed tomography perceptual image quality assessment. *Medical Image Analysis*, 99:103343, 2025. ISSN 1361-8415. https://doi.org/10.1016/j.media.2024.103343. https://www.sciencedirect.com/science/article/pii/S1361841524002688.
56. Ke Lei, Ali B. Syed, Xucheng Zhu, John M. Pauly, and Shreyas S. Vasanawala. Artifact- and content-specific quality assessment for mri with image rulers. *Medical Image Analysis*, 77:102344, 2022. ISSN 1361-8415. https://doi.org/10.1016/j.media.2021.102344. https://www.sciencedirect.com/science/article/pii/S1361841521003893.
57. Weisi Lin and C.-C. Jay Kuo. Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, 22(4):297–312, 2011. ISSN 1047-3203. https://doi.org/10.1016/j.jvcir.2011.01.005. https://www.sciencedirect.com/science/article/pii/S1047320311000204.
58. Michael Lustig, David L. Donoho, Juan M. Santos, and John M. Pauly. Compressed sensing mri. *IEEE Signal Processing Magazine*, 25(2): 72–82, 2008. https://doi.org/10.1109/MSP.2007.914728.
59. Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A. Riegler, Manuel Wiesenfarth, A. Emre Kavur, Carole H. Sudre, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, Tim Rädsch, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Arriel Benis, Matthew B. Blaschko, M. Jorge Cardoso, Veronika Cheplygina, Beth A. Cimini, Gary S. Collins, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram van Ginneken, Robert Haase, Daniel A. Hashimoto, Michael M. Hoffman, Merel Huisman, Pierre Jannin, Charles E. Kahn, Dagmar Kainmueller, Bernhard Kainz, Alexandros Karargyris, Alan Karthikesalingam, Florian Kofler, Annette Kopp-Schneider, Anna Kreshuk, Tahsin Kurc, Bennett A. Landman, Geert Litjens, Amin Madani, Klaus Maier-Hein, Anne L. Martel, Peter Mattson, Erik Meijering, Bjoern Menze, Karel G. M. Moons, Henning Müller, Brennan Nichyporuk, Felix Nickel, Jens Petersen, Nasir Rajpoot, Nicola Rieke, Julio Saez-Rodriguez, Clara I. Sánchez, Shravya Shetty, Maarten van Smeden, Ronald M. Summers, Abdel A. Taha, Aleksei Tiulpin, Sotirios A. Tsaftaris, Ben Van Calster, Gaël Varoquaux, and Paul F. Jäger. Metrics reloaded: recommendations for image analysis validation. *Nature Methods*, 21(2):195–212, 2024. https://doi.org/10.1038/s41592-023-02151-z.
60. Ivica Mandić, Hajdi Peić, Jonatan Lerga, and Ivan Štajduhar. Denoising of x-ray images using the adaptive algorithm based on the lpa-rici algorithm. *Journal of Imaging*, 4(2), 2018. ISSN 2313-433X. https://doi.org/10.3390/jimaging4020034. https://www.mdpi.com/2313-433X/4/2/34.
61. Sho Maruyama. Properties of the ssim metric in medical image assessment: correspondence between measurements and the spatial frequency spectrum. *Physical and Engineering Sciences in Medicine*, 46 (3):1131–1141, 2023. https://doi.org/10.1007/s13246-023-01280-1.
62. Allister Mason, James Rioux, Sharon E Clarke, Andreu Costa, Matthias Schmidt, Valerie Keough, Thien Huynh, and Steven Beyea. Comparison of objective image quality metrics to expert radiologists' scoring of diagnostic quality of mr images. *IEEE transactions on medical imaging*, 39 (4):1064–1072, 2019.
63. Matthew J Muckley, Bruno Riemenschneider, Alireza Radmanesh, Sunwoo Kim, Geunu Jeong, Jingyu Ko, Yohan Jun, Hyungseob Shin, Dosik Hwang, Mahmoud Mostapha, Simon Arberet, Dominik Nickel, Zaccharie Ramzi, Philippe Ciuciu, Jean-Luc Starck, Jonas Teuwen, Dimitrios Karkalousos, Chaoping Zhang, Anuroop Sriram, Zhengnan Huang, Nafissa Yakubova, Yvonne W Lui, and Florian Knoll. Results of the 2020 fastmri challenge for machine learning mr image reconstruction. *IEEE Trans Med Imaging*, 40(9):2306–2317, Sep 2021a. ISSN 1558-254X (Electronic); 0278-0062 (Print); 0278-0062 (Linking). https://doi.org/10.1109/TMI.2021.3075856.
64. Matthew J Muckley, Bruno Riemenschneider, Alireza Radmanesh, Sunwoo Kim, Geunu Jeong, Jingyu Ko, Yohan Jun, Hyungseob Shin, Dosik Hwang, Mahmoud Mostapha, et al. Results of the 2020 fastmri challenge for machine learning mr image reconstruction. *IEEE transactions on medical imaging*, 40 (9):2306–2317, 2021b.
65. Subhadip Mukherjee, Sören Dittmer, Zakhar Shumaylov, Sebastian Lunz, Ozan Öktem, and Carola-Bibiane Schönlieb. Learned convex regularizers for inverse problems, 2021.
66. Camila Munoz, Anastasia Fotaki, René M. Botnar, and Claudia Prieto. Latest advances in image acceleration: All dimensions are fair game. *Journal of Magnetic Resonance Imaging*, 57 (2):387–402, 2023. https://doi.org/10.1002/jmri.28462
67. Lipeng Ning, Kawin Setsompop, Oleg Michailovich, Nikos Makris, Martha E Shenton, Carl-Fredrik Westin, and Yogesh Rathi. A joint compressed-sensing and super-resolution approach for very high-resolution diffusion imaging. *Neuroimage*, 125:386–400, Jan 2016. ISSN 1095-9572 (Electronic); 1053-8119 (Print); 1053-8119 (Linking). https://doi.org/10.1016/j.neuroimage.2015.10.061.
68. Facebook AI NYU Langone Health. Public leaderboard fastmri challenge. https://web.archive.org/web/20220321054325mp_/https://fastmri.org/leaderboards.
69. Rafał Obuchowicz, Adam Piórkowski, Andrzej Urbanik, and Michał Strzelecki. Influence of acquisition time on mr image quality estimated with nonparametric measures based on texture features. *Biomed Res Int*, 2019:3706581, 2019. ISSN 2314-6141 (Electronic); 2314-6133 (Print). https://doi.org/10.1155/2019/3706581.
70. Rafał Obuchowicz, Mariusz Oszust, Marzena Bielecka, Andrzej Bielecki, and Adam Piórkowski. Magnetic resonance image quality assessment by using non-maximum suppression and entropy analysis. *Entropy (Basel)*, 22(2), Feb 2020. ISSN 1099-4300 (Electronic); 1099-4300 (Linking). https://doi.org/10.3390/e22020220.

71. J. Pambrun and R. Noumeir. Limitations of the ssim quality metric in the context of diagnostic imaging. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 2960–2963, 2015.https://doi.org/10.1109/ICIP.2015.7351345.
72. John W. Pickering, Scott A. Prahl, Niek van Wieringen, Johan F. Beek, Henricus J. C. M. Sterenborg, and Martin J. C. van Gemert. Double-integrating-sphere system for measuring the optical properties of tissue. *Appl. Opt.*, 32(4):399–410, Feb 1993. https://doi.org/10.1364/AO.32.000399. https://opg.optica.org/ao/abstract.cfm?URI=ao-32-4-399.
73. Michael Pircher. Optical coherence tomography and its application to imaging of skin and retina. In Bob D. Guenther and Duncan G. Steel, editors, *Encyclopedia of Modern Optics (Second Edition)*, pages 155–167. Elsevier, Oxford, second edition edition, 2018. ISBN 978-0-12-814982-9. https://doi.org/10.1016/B978-0-12-803581-8.09787-3. https://www.sciencedirect.com/science/article/pii/B9780128035818097873.
74. Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. Image database tid2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, 2015. ISSN 0923-5965. https://doi.org/10.1016/j.image.2014.10.009. https://www.sciencedirect.com/science/article/pii/S0923596514001490.
75. K P Pruessmann, M Weiger, M B Scheidegger, and P Boesiger. Sense: sensitivity encoding for fast mri. *Magn Reson Med*, 42(5):952–962, Nov 1999. ISSN 0740-3194 (Print); 0740-3194 (Linking).
76. Gabriel Ramos-Llordén, Lipeng Ning, Congyu Liao, Rinat Mukhometzianov, Oleg Michailovich, Kawin Setsompop, and Yogesh Rathi. High-fidelity, accelerated whole-brain submillimeter in vivo diffusion mri using gslider-spherical ridgelets (gslider-sr). *Magn Reson Med*, 84(4):1781–1795, Oct 2020. ISSN 1522-2594 (Electronic); 0740-3194 (Print); 0740-3194 (Linking). https://doi.org/10.1002/mrm.28232.
77. Zaccharie Ramzi, Philippe Ciuciu, and Jean-Luc Starck. XPDNet for MRI Reconstruction: an application to the 2020 fastMRI challenge. In *ISMRM*, pages 1–4, 2021.
78. Rafael Reisenhofer, Sebastian Bosse, Gitta Kutyniok, and Thomas Wiegand. A haar wavelet-based perceptual similarity index for image quality assessment. *Signal Process. Image Commun.*, 61:33–43, 2018. https://doi.org/10.1016/j.image.2017.11.001.
79. Geoffrey D Rubin. Computed tomography: revolutionizing the practice of medicine for 40 years. *Radiology*, 273(2 Suppl):S45–74, Nov 2014. ISSN 1527-1315 (Electronic); 0033-8419 (Linking). https://doi.org/10.1148/radiol.14141356.
80. Branimir Rusanov, Ghulam Mubashar Hassan, Mark Reynolds, Mahsheed Sabet, Jake Kendrick, Pejman Rowshanfarzad, and Martin Ebert. Deep learning methods for enhancing cone-beam ct image quality toward adaptive radiation therapy: A systematic review. *Med Phys*, 49(9):6019–6054, Sep 2022. ISSN 2473-4209 (Electronic); 0094-2405 (Print); 0094-2405 (Linking). https://doi.org/10.1002/mp.15840.
81. Malena Sabaté Landman, Ander Biguri, Sepideh Hatamikia, Richard Boardman, John Aston, and Carola-Bibiane Schönlieb. On krylov methods for large-scale cbct reconstruction. *Physics in Medicine and Biology*, 2022.
82. Ehsan Samei and Elizabeth A. Krupinski. *The Handbook of Medical Image Perception and Techniques*. Cambridge University Press, Cambridge, 2018. ISBN 9781107194885. https://doi.org/10.1017/9781108163781. https://www.cambridge.org/core/product/D25C3FA40B8BC2521301F9DBA119F747.
83. Umme Sara, Morium Akter, and Mohammad Shorif Uddin. Image quality assessment through fsim, ssim, mse and psnr—a comparative study. *Journal of Computer and Communications*, 07:8–18, 01 2019. https://doi.org/10.4236/jcc.2019.73002.
84. Alan Schiska. Teaching radiography students the alara principle. *Radiol Technol*, 93(2):228–231, Nov 2021. ISSN 1943-5657 (Electronic); 0033-8397 (Linking).
85. Carmen Schwaiger, Thomas Haider, Verena Endmayr, Tobias Zrzavy, Victoria E Gruber, Gerda Ricken, Anika Simonovska, Simon Hametner, Jan M Schwab, and Romana Höftberger. Dynamic induction of the myelin-associated growth inhibitor nogo-a in perilesional plasticity regions after human spinal cord injury. *Brain Pathol*, 33(1):e13098, Jan 2023. ISSN 1750-3639 (Electronic); 1015-6305 (Print); 1015-6305 (Linking). https://doi.org/10.1111/bpa.13098.
86. Euclid Seeram and David Seeram. Image postprocessing in digital radiology—a primer for technologists. *Journal of Medical Imaging and Radiation Sciences*, 39 (1):23–41, 2008. ISSN 1939-8654. https://doi.org/10.1016/j.jmir.2008.01.004. https://www.sciencedirect.com/science/article/pii/S1939865408000052.
87. H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. Live image quality assessment database release 2. http://live.ece.utexas.edu/research/quality
88. M E Shenton, H M Hamoda, J S Schneiderman, S Bouix, O Pasternak, Y Rathi, M-A Vu, M P Purohit, K Helmer, I Koerte, A P Lin, C-F Westin, R Kikinis, M Kubicki, R A Stern, and R Zafonte. A review of magnetic resonance imaging and diffusion tensor imaging findings in mild traumatic brain injury. *Brain Imaging Behav*, 6(2):137–192, Jun 2012. ISSN 1931-7565 (Electronic); 1931-7557 (Print); 1931-7557 (Linking). https://doi.org/10.1007/s11682-012-9156-5.
89. Muhammad Faizan Shirazi, Elisabeth Brunner, Marie Laslandes, Andreas Pollreisz, Christoph K. Hitzenberger, and Michael Pircher. Visualizing human photoreceptor and retinal pigment epithelium cell mosaics in a single volume scan over an extended field of view with adaptive optics optical coherence tomography. *Biomed. Opt. Express*, 11(8):4520–4535, Aug 2020. https://doi.org/10.1364/BOE.393906. https://opg.optica.org/boe/abstract.cfm?URI=boe-11-8-4520.
90. Emil Y Sidky, Jakob H Jørgensen, and Xiaochuan Pan. Convex optimization problem prototyping for image reconstruction in computed tomography with the chambolle-pock algorithm. *Phys Med Biol*, 57(10):3065–3091, May 2012. ISSN 1361-6560 (Electronic); 0031-9155 (Print); 0031-9155 (Linking). https://doi.org/10.1088/0031-9155/57/10/3065.
91. Youssef Skandarani, Pierre-Marc Jodoin, and Alain Lalande. Gans for medical image synthesis: An empirical study. *Journal of Imaging*, 9(3), 2023. ISSN 2313-433X. https://doi.org/10.3390/jimaging9030069. https://www.mdpi.com/2313-433X/9/3/69.
92. Anuroop Sriram, Jure Zbontar, Tullie Murrell, Aaron Defazio, C Lawrence Zitnick, Nafissa Yakubova, Florian Knoll, and Patricia Johnson. End-to-end variational networks for accelerated mri reconstruction. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*, pages 64–73. Springer, 2020.
93. Paul W. Sweeney, Lina Hacker, Thierry L. Lefebvre, Emma L. Brown, Janek Gröhl, and Sarah E. Bohndiek. Segmentation of 3d blood vessel networks using unsupervised deep learning. *bioRxiv*, 2023. https://doi.org/10.1101/2023.04.30.538453. https://www.biorxiv.org/content/early/2023/04/30/2023.04.30.538453.
94. Dang Thanh, Kalavathi Palanisamy, Le Thanh, and Surya Prasath. Chest x-ray image denoising using nesterov optimization method with total variation regularization. *Procedia Computer Science*, 171:1961–1969, 06 2020. https://doi.org/10.1016/j.procs.2020.04.210.
95. Radhika Tibrewala, Tarun Dutt, Angela Tong, Luke Ginocchio, Mahesh B Keerthivasan, Steven H Baete, Sumit Chopra, Yvonne W Lui, Daniel K Sodickson, Hersh Chandarana, et al. Fastmri prostate: A publicly available, biparametric mri dataset

to advance machine learning for prostate cancer imaging. arXiv:2304.09254, 2023.
96. Minh-Trieu Tran, Soo-Hyung Kim, Hyung-Jeong Yang, and Guee-Sang Lee. Deep learning-based inpainting for chest x-ray image. In *The 9th International Conference on Smart Media and Applications*, SMA 2020, pages 267–271, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450389259. https://doi.org/10.1145/3426020.3426088.
97. Ioannis A. Tsalafoutas, Shady AlKhazzam, Virginia Tsapaki, and Mohammed Hassan Kharita. Automatic image quality evaluation in digital radiography using for-processing and for-presentation images. *Journal of Applied Clinical Medical Physics*, n/a (n/a):e14285, 2024. https://doi.org/10.1002/acm2.14285. https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/acm2.14285.
98. Mehmet Ozan Unal, Metin Ertas, and Isa Yildirim. Self-supervised training for low-dose ct reconstruction. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 69–72, 2021. https://doi.org/10.1109/ISBI48211.2021.9433944.
99. Mehmet Ozan Unal, Metin Ertas, and Isa Yildirim. Proj2proj: self-supervised low-dose ct reconstruction. *PeerJ Comput Sci*, 10:e1849, 2024. ISSN 2376-5992 (Electronic); 2376-5992 (Linking).https://doi.org/10.7717/peerj-cs.1849.
100. Gwendolyn Van Steenkiste, Ben Jeurissen, Jelle Veraart, Arnold J den Dekker, Paul M Parizel, Dirk H J Poot, and Jan Sijbers. Super-resolution reconstruction of diffusion parameters from diffusion-weighted images with different slice orientations. *Magn Reson Med*, 75(1):181–195, Jan 2016. ISSN 1522-2594 (Electronic); 0740-3194 (Linking). https://doi.org/10.1002/mrm.25597.
101. Abhinau K. Venkataramanan, Chengyang Wu, Alan Conrad Bovik, Ioannis Katsavounidis, and Zafar Shahid. A hitchhiker's guide to structural similarity. *IEEE Access*, 9:28872–28896, 2021. https://doi.org/10.1109/ACCESS.2021.3056504.
102. F.R. Verdun, D. Racine, J.G. Ott, M.J. Tapiovaara, P. Toroi, F.O. Bochud, W.J.H. Veldkamp, A. Schegerer, R.W. Bouwman, I. Hernandez Giron, N.W. Marshall, and S. Edyvean. Image quality in ct: From physical measurements to model observers. *Physica Medica*, 31(8):823–843, 2015. ISSN 1120-1797.https://doi.org/10.1016/j.ejmp.2015.08.007. https://www.sciencedirect.com/science/article/pii/S1120179715003294.
103. Jie Wang, Huaiwei Cong, Xin Wei, Baolian Qi, Jinpeng Li, and Ting Cai. X-ray image blind denoising in hybrid noise based on convolutional neural networks. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI-IAT '21 Companion)*,, pages 203–212, 12 2021. https://doi.org/10.1145/3498851.3498952.
104. Zhou Wang, Alan C. Bovik, and Ligang Lu. Why is image quality assessment so difficult? In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–3313–IV–3316, 2002. https://doi.org/10.1109/ICASSP.2002.5745362.
105. Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13 (4):600–612, 2004. https://doi.org/10.1109/TIP.2003.819861.
106. Maciej Wojtkowski, Vivek J. Srinivasan, Tony H. Ko, James G. Fujimoto, Andrzej Kowalczyk, and Jay S. Duker. Ultrahigh-resolution, high-speed, fourier domain optical coherence tomography and methods for dispersion compensation. *Opt. Express*, 12(11):2404–2422, 2004. https://doi.org/10.1364/OPEX.12.002404. https://opg.optica.org/oe/abstract.cfm?URI=oe-12-11-2404.
107. Jingyan Xu and Benjamin M. W. Tsui. Electronic noise modeling in statistical iterative reconstruction. *IEEE Transactions on Image Processing*, 18 (6):1228–1238, 2009. https://doi.org/10.1109/TIP.2009.2017139.
108. Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J. Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael Rabbat, Pascal Vincent, Nafissa Yakubova, James Pinkerton, Duo Wang, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastmri: An open dataset and benchmarks for accelerated mri, 2019.
109. Richard Zhang. Github repository of lpips. https://github.com/richzhang/PerceptualSimilarity.
110. Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. https://doi.org/10.1109/CVPR.2018.00068.
111. Bo Zhou, Jo Schlemper, Neel Dey, Seyed Sadegh Mohseni Salehi, Kevin Sheth, Chi Liu, James S. Duncan, and Michal Sofka. Dual-domain self-supervised learning for accelerated non-cartesian mri reconstruction. *Medical Image Analysis*, 81:102538, 2022. ISSN 1361-8415. https://doi.org/10.1016/j.media.2022.102538. https://www.sciencedirect.com/science/article/pii/S1361841522001852.

Authors and Affiliations

Anna Breger[1,2] · Ander Biguri[1] · Malena Sabaté Landman[3] · Ian Selby[4] · Nicole Amberg[5] · Elisabeth Brunner[2] · Janek Gröhl[6,7] · Sepideh Hatamikia[8,9] · Clemens Karner[2] · Lipeng Ning[10] · Sören Dittmer[1] · Michael Roberts[1] · A.I.X.-C.O.V.N.E.T. Collaboration · Carola-Bibiane Schönlieb[1]

✉ Anna Breger
ab2864@cam.ac.uk

1 Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK

2 Center of Medical Physics and Biomedical Engineering, Medical University of Vienna, Vienna, Austria

3 Department of Mathematics, Emory University, Atlanta, USA

4 Department of Radiology, University of Cambridge, Cambridge, UK

5 Department of Neurology, Medical University of Vienna, Vienna, Austria

6 Department of Physics, University of Cambridge, Cambridge, UK

7 Cancer Research UK, Cambridge Institute, University of Cambridge, London, UK

8 Research Center for Clinical AI-Research in Omics and Medical Data Science (CAROM), Department of Medicine, Krems an der Donau, Austria

9 Austrian Center for Medical Innovation and Technology, Wiener Neustadt, Austria

10 Harvard Medical School, Brigham and Women's Hospital, Boston, MA, USA