A STUDY ON THE ADEQUACY OF COMMON IQA MEASURES FOR MEDICAL IMAGES

Anna Breger[1,2*] *Clemens Karner*[1,2] *Ian Selby*[3,4] *Janek Gröhl*[5] *Sören Dittmer*[1]
Edward Lilley[2] *Judith Babar*[3,4] *Jake Beckford*[4] *Timothy J Sadler*[4]
Shahab Shahipasand[4] *Arthikkaa Thavakumar*[4] *Michael Roberts*[1] *Carola-Bibiane Schönlieb*[1]

[1]University of Cambridge, Department of Applied Mathematics and Theoretical Physics, Cambridge, United Kingdom
[2]Medical University of Vienna, Center of Medical Physics and Biomedical Engineering, Vienna, Austria
[3]University of Cambridge, Department of Radiology, Cambridge, United Kingdom
[4]Cambridge University Hospitals NHS Trust, Department of Radiology, Cambridge, United Kingdom
[5]University of Cambridge, Department of Physics, Cambridge, United Kingdom

ABSTRACT

Image quality assessment (IQA) is standard practice in the development stage of novel machine learning algorithms that operate on images. The most commonly used IQA measures have been developed and tested for natural images, but not in the medical setting. Reported inconsistencies arising in medical images are not surprising, as they have different properties than natural images. In this study, we test the applicability of common IQA measures for medical image data by comparing their assessment to manually rated chest X-ray (5 experts) and photoacoustic image data (1 expert). Moreover, we include supplementary studies on grayscale natural images and accelerated brain MRI data. The results of all experiments show a similar outcome in line with previous findings for medical imaging: PSNR and SSIM in the default setting are in the lower range of the result list and HaarPSI outperforms the other tested measures in the overall performance. Also among the top performers in our medical experiments are the full reference measures DISTS, FSIM, LPIPS and MS-SSIM. Generally, the results on natural images yield considerably higher correlations, suggesting that the additional employment of tailored IQA measures for medical imaging algorithms is needed.

1. INTRODUCTION

Advances in medical imaging technologies have been groundbreaking in the last decades, including the rapid development of deep learning techniques. To ensure the quality of novel image processing methodologies, quantitative image quality assessment (IQA) plays an important role in quality assurance in addition to visual inspection or even serves as the main assessment criterion when no experts' opinion is available. Quantitative IQA can roughly be divided into three categories based on their underlying assumptions and the information available for their evaluation [1, 2]. The first one is called full reference (FR) IQA, where a known full image is used as a reference 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, aims to judge the quality without a reference based on pre-defined properties. Reduced reference (RR) IQA uses specific retrieved image information of reference data.

Most commonly used IQA measures have been developed for natural images and tested for specific tasks on a small amount of publicly available rated data sets. It is unknown how well these measures can be expanded to medical images since they have different properties including a different target space (color versus grayscale). Little research has been performed on the applicability of common IQA measures to medical imaging data, see e.g. [3] for a recent overview. Many prior applicability studies have limitations in the study design, including non-expert ratings (see e.g. [4]), non-realistic distortions (such as Gaussian additive noise, see e.g. [5]) and a very limited choice of IQA measures (see e.g. [6]). The research field is suffering from the lack of publicly available data sets and expert annotations for reproducible comparison studies, as well as non-public code of introduced IQA measures.

Recently, in [3], the first extensive study on IQA measures comparing MRI outputs of image restoration models with expert ratings has been published. The results suggest that the most widely used measures PSNR [7] and SSIM [1] are not a good choice for the tested MRI tasks. Here, we will build on that study and assess the adequacy of common IQA measures for medical imaging in 4 different experimental setups and 5 data sets, including expert ratings for photoacoustic and chest X-ray data. For the comparisons we have included the following IQA measures: peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), multi-scale SSIM (MS-SSIM) [8], information con-

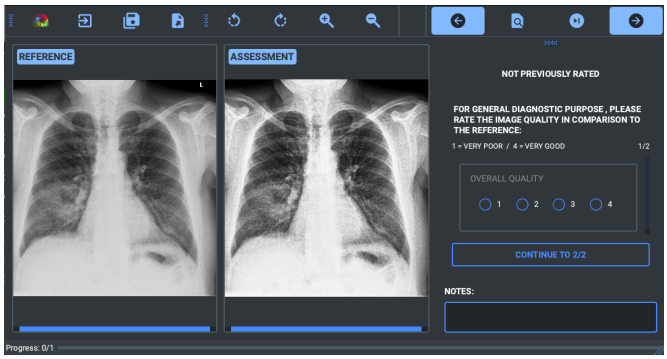

Fig. 1: The speedyIQA annotation app allows setting a task and rating categories for manual image ratings.

tent weighted SSIM (IW-SSIM) [9], deep image structure and texture similarity (DISTS) [10], DCT subband similarity (DSS) [11], feature similarity index (FSIM) [12], gradient magnitude similarity deviation (GMSD) [13], Haar wavelet-based perceptual similarity (HaarPSI) [14], learned perceptual image patch similarity (LPIPS) [15], mean deviation similarity (MDSI) [16], visual information fidelity (VIF) [17], visual saliency-induced index (VSI) [18], in the FR setting; blind/referenceless image spatial quality evaluator BRISQUE [19], from patches to pictures (PaQ-2-PIQ) [20], natural image quality evaluator (NIQE) [21] in the NR setting.

NO Reference

2. METHODS

The IQA measures were computed with the implementations provided by their authors, either in MATLAB or Python, or both. An exception is the VIF measure for which we used the PyTorch Lightning implementation because the code by the authors is not publicly available. Reporting the implementation used is of utmost importance because IQA results may differ substantially, see, e.g. [22]).

When novel manual image ratings were obtained (Experiment 1, 3, and 4), the publicly available speedyIQA annotation app [23] was used. The software asks the user to set a task and the rating categories, see Figure 1. Obtained ratings have been saved as a CSV file.

For the evaluation of the IQ measure performance we employed the Spearman Rank Correlation Coefficient (SRCC) and the Kendall Rank Correlation Coefficient (KRCC), which assess the ordinal association (rank) between the quality measures and the manual ratings. To account for the different scoring between multiple graders, we apply the z-score to the raw rating data of each grader and afterwards compute the mean, cf. [24]. The absolute values are stated.

2.1. Experiment 1 - Grayscale LIVE data

The two data sets in the first experiment correspond to commonly used natural imaging quality assessment databases, namely the LIVE Image Quality Assessment Database Release 2 and the LIVE Multiply Distorted Image Quality Database [25, 26], which we transformed to grayscale images with the in-built MATLAB function *mat2gray*. The data sets contain respectively 982 and 405 images, including the degradations of Gaussian noise, jpeg compression, and blurring. Five volunteers were asked to rate all degraded images of both data sets from 1 (very poor), 2 (poor), 3 (good) to 4 (very good) regarding the ability to identify the detailed image content in comparison to the given reference image. Note that we did not use the originally available color image ratings because of the change in target space, see e.g. [14] for an example of inconsistencies related to grayscale versus color image ratings) and moreover, here, we asked the volunteers to rate the quality regarding detailed image content. Usually medical images are grayscale and detailed quality is of upmost importance, therefore this experiment is supplementing the tasks with medical data. For the evaluation, the SRCC and KRCC between the z-score of the image ratings and the IQ values are computed. The annotations are planned to be made available.

2.2. Experiment 2 - MRI acceleration

The MRI data set was retrieved from the publicly available fastMRI brain dataset [27], which contains in total 6405 T1, T2, and FLAIR 3D k-space volumes. The fastMRI challenge series provided MRI datasets to foster the development of accelerated reconstruction algorithms. In [3] data from the fastMRI data set has been used for a comprehensive analysis of different degradations with expert annotations. Here, we use a subset of 4742 reference image slices (created by the root sum of squares, rSOS, of the fully sampled data) and around $151k$ corresponding accelerated image reconstructions obtained from two machine learning algorithms that took part in the fastMRI multi-coil brain dataset challenge in 2020, namely the end-to-end variational network *E2E-VarNet* [28] and *XPDNet* [29]. *XPDNet* was among the top three submissions of the challenge and both algorithms perform very well on the corresponding public leaderboard. The reconstructions were obtained by the application of the machine learning models to the fully sampled and accelerated data, where we designed masks with acceleration factor 1 to 16 to yield decreasing visual image quality in the reconstructions, see Figure 2.

The created data set serves as a sanity check for the identification of decreasing image quality. We evaluate the performance of the IQA measures in two ways: First, the SRCC and KRCC are computed for each image and the corresponding quality decreases, where the acceleration factor serves as the image quality category, and secondly, we will plot the mean IQA value for each acceleration class and measure. For this illustration, all measures are linearly scaled so that lower is always worse, starting with the IQA value of acceleration

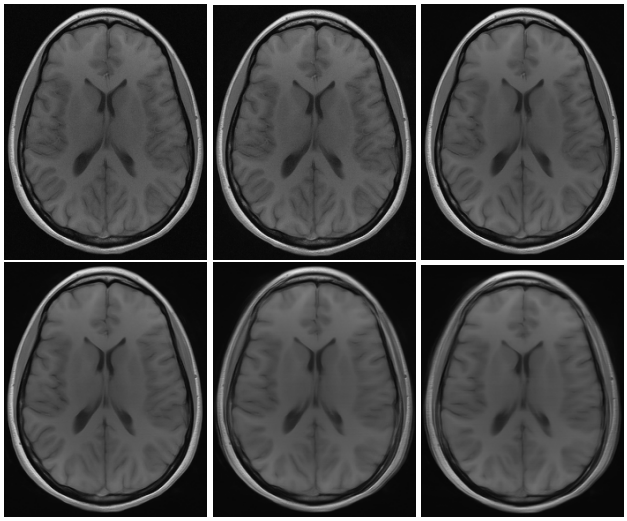

Fig. 2: Reconstructed MRI brain data from the fastMRI data set obtained with *E2E-VarNet* on the sub-sampled data with acceleration factors $1, 4, 8, 12$ and 16. The left top image corresponds to the reference image obtained via the rSOS of the fully sampled data. The visual quality decreases with the increased acceleration factor.

factor 1 as 100%. The purpose of this experiment is to test the IQA measure's ability to flawlessly detect distinct quality decrease in medical images.

2.3. Experiment 3 - Photoacoustic reconstruction

Photoacoustic (PA) imaging is an emerging medical imaging modality with important clinical applications such as inflammatory bowel disease and cardiovascular diseases [30]. . The inverse problems of PAI pertain to accurately visualizing molecular distributions and determining functional tissue information from acquired PA time series signals [31]. We use a previously published data set that consists of reconstructed images containing estimated distributions of the optical absorption coefficient from cross-sectional photoacoustic images of piecewise constant test objects (phantoms) [32]. The PA data was acquired with a commercial photoacoustic imaging system. The 378 reference images are obtained using a double-integrating sphere [33] setup as a complementary measurement system, which 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. Unfortunately, this process is unfeasible for complicated objects or in vivo images.

Here, 1134 reconstructed images, corresponding to the outputs of 3 reconstruction methods (see Figure 3), have been annotated by one expert. The expert was asked to rate the images from 1 (very poor), 2 (poor), 3 (good) to 4 (very good) regarding overall quality in comparison to the reference image without changing the contrast or luminance. For visualization and assessment, the outputs of the algorithms were clipped with the reference image's maximum. For the evaluation, the SRCC and KRCC between the expert's rating and the IQ values are computed.

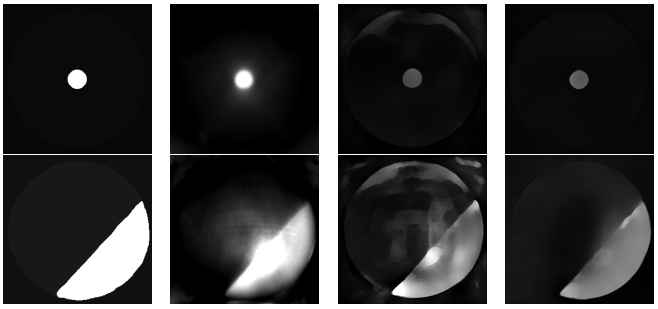

Fig. 3: Two examples of the photoacoustic images used, ground truth (left) and reconstructions from three algorithms. The first algorithm corrects a reconstructed PA image by using the light fluence obtained from simulations. The second and third algorithms are deep-learning models trained to estimate the absorption coefficient.

2.4. Experiment 4 - Chest X-Ray post-processing

For the last experiment, we used posteroanterior chest radiographs that were acquired on two imaging systems (both Discovery XR656 HD models, GE Healthcare, USA) at Cambridge University Hospitals NHS Trust. Each scanner had previously been set up with different default post-processing parameters (chosen by local radiologists following a subjective assessment), yielding the reference images. Additional images, serving as real-life examples of lower quality, were produced for each radiographic exposure using multiple different post-processing settings, see examples in Figure 4. The post-processing was applied in the hospital directly on the scanner itself by adjusting parameters in the framework provided, including brightness, overall and tissue contrast, edge enhancement, noise reduction, and local contrast enhancement. In total, the data set contains 444 reference images and 2018 post-processed images that were rated by 3 consultant radiologists, 1 trainee radiologist, and 1 senior reporting radiographer. Each expert was asked to rate all post-processed images from 1 (very poor), 2 (poor), 3 (good) to 4 (very good) for general diagnostic purposes in comparison to the reference without changing the contrast or luminance of the displayed image. The image data and expert annotations are planned to be managed and made available through the hospital's clinical informatics unit. For the evaluation, the SRCC and KRCC between the z-score of the image ratings and the IQ values are computed.

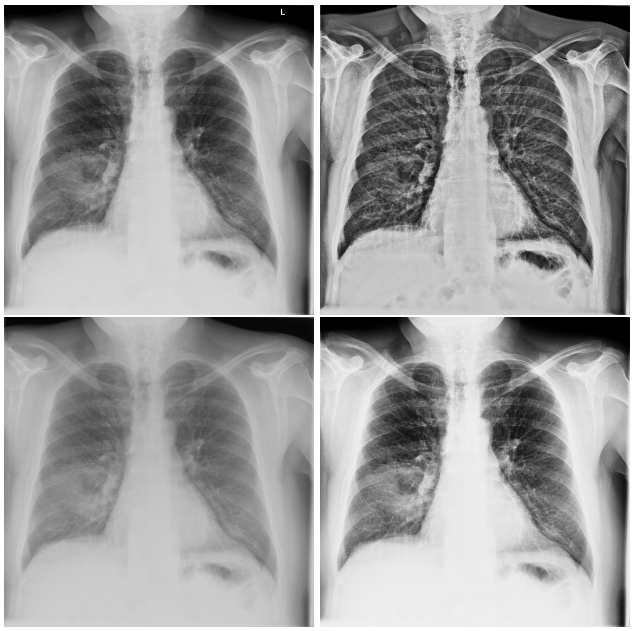

Fig. 4: Chest X-ray scans with different kinds of post-processing, the image in the top left serves as the reference, the other images show lower visual quality.

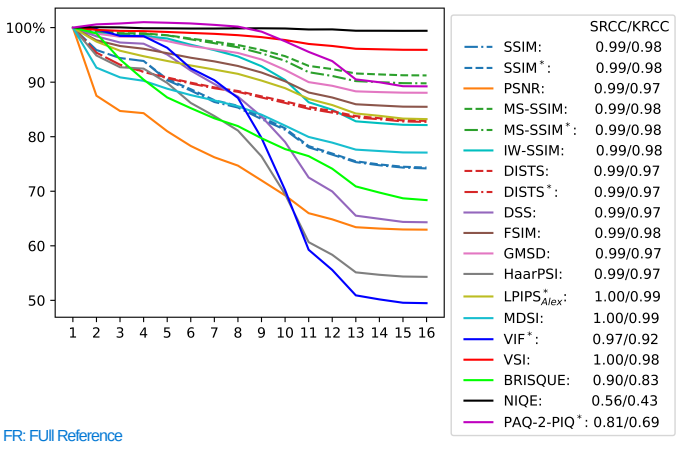

Fig. 5: IQA comparison of decreasing MRI reconstruction quality through an increase in acceleration factor (1 to 16). All tested FR measures correctly identify a decrease in quality, two tested NR measures (NIQE and PAQ-2-PIQ) struggle to identify the quality loss accurately. The SRCC/KRCC values between the measures and the acceleration categories show corresponding behavior.

3. RESULTS AND DISCUSSION

An important problem regarding fairness in machine learning is the adequacy of employed assessment methods. In the context of IQA measures, this relates to the question of generalizability across different datasets and image target spaces. To tackle this question, we conducted 4 independent experiments, where on the one hand 4 novel data sets were obtained with manual expert ratings and on the other hand 1 data set was designed to test the ability to identify obvious quality decrease in medical images. For the comparison we chose IQA measures that are commonly used across image modalities and tasks. The results of Experiment 2 (see Figure 5) show that all tested FR measures were able to pass the sanity check, i.e. they were able to identify the distinct decrease in quality of the MRI images, confirmed through the descending mean quality value as well as the high SRCC/KRCC values related to the acceleration categories. Two of the tested NR measures, NIQE and PAQ-2-PIQ, struggled to correctly identify the decreasing image quality.

In Table 1 we show the SRCC and KRCC values for the tested IQA measures and the 4 manually rated data sets of Experiments $1, 3$ and 4. The commonly used measures PSNR and SSIM yield relatively low correlation values in all tasks, especially regarding the medical imaging data sets. Outstanding behavior is shown by HaarPSI which is among the top 3 performers for all tested data sets. Generally, the correlation values for the natural images are higher than for the medical images, indicating that improvement is still needed in the medical domain and tailored available measures should be employed for specified tasks. The reasons for that are manifold. On the one hand, most of the tested measures have been developed and calibrated for natural images, and on the other hand, medical imaging tasks are often very complex or ask for specific quality information. Strongly dependent on the task, different image features might be more or less important.

The tested FR measures yield higher correlation coefficients than the NR measures, which is not surprising as FR assessment is using more information. In the NR setting, recently, measures tailored towards assessment of specific medical images have been introduced, see e.g. [34,35], and it is advisable to employ such measures in medical imaging tasks, in addition to more generalizable measures. Using unsuitable IQA measures might give incorrect conclusions about novel introduced algorithms, not necessarily favoring the most adequate methods.

4. CONCLUSION

We have tested the ability of common IQA measures to assess the quality of medical images in 4 experiments. To that end, we employed experts to rate a total of around 3000 chest X-ray and photoacoustic images to provide a comparison to the assessment of common FR and NR IQA measures. Additionally, we conducted supplementary experiments with around 1500 rated grayscale natural images and $151k$ MRI reconstructions with IQ categories through varying acceleration factors. The results show that most measures succeed in these simpler experiments, but struggle in the more complicated medical imaging tasks. In particular, the overall correlations

	Grayscale Natural Images		Medical Images	
Full-Reference	LIVE	LIVE_{Multi}	Photoacoustic	Chest X-ray
PSNR	0.87 / 0.71	0.74 / 0.56	0.51 / 0.39	0.66 / 0.48
SSIM	0.88 / 0.72	0.67 / 0.49	0.66 / 0.53	0.70 / 0.50
SSIM*	0.88 / 0.71	0.67 / 0.49	0.67 / 0.54	0.70 / 0.50
MS-SSIM	0.91 / 0.77	0.88 / 0.70	**0.71 / 0.57**	0.80 / 0.58
MS-SSIM*	0.91 / 0.76	0.88 / 0.71	0.69 / 0.56	0.79 / 0.57
IW-SSIM	0.92 / 0.79	**0.93 / 0.77**	0.64 / 0.51	0.72 / 0.52
DISTS	0.91 / 0.76	0.75 / 0.56	**0.70 / 0.56**	0.77 / 0.54
DISTS*	0.91 / 0.76	0.74 / 0.56	0.69 / 0.55	0.77 / 0.55
DSS	0.92 / 0.78	0.91 / 0.74	0.67 / 0.53	0.68 / 0.50
FSIM	**0.93 / 0.80**	0.92 / 0.75	**0.70 / 0.56**	0.79 / 0.56
GMSD	**0.92 / 0.79**	0.91 / 0.74	0.66 / 0.53	0.82 / 0.61
HaarPSI	**0.93 / 0.79**	**0.92 / 0.76**	**0.74 / 0.60**	**0.83 / 0.61**
LPIPS^*_{Alex}	0.90 / 0.75	0.77 / 0.59	0.69 / 0.56	**0.82 / 0.62**
MDSI	0.92 / 0.78	**0.92 / 0.76**	0.54 / 0.42	0.76 / 0.53
VIF*	0.85 / 0.68	0.90 / 0.72	0.16 / 0.12	0.63 / 0.43
VSI	0.91 / 0.77	0.89 / 0.71	0.03 / 0.01	**0.83 / 0.62**
No-Reference				
BRISQUE	0.92 / 0.78	0.46 / 0.33	0.55 / 0.43	0.05 / 0.03
NIQE	0.88 / 0.71	0.75 / 0.57	0.49 / 0.38	0.44 / 0.30
PAQ-2-PIQ*	0.76 / 0.57	0.86 / 0.68	0.26 / 0.19	0.59 / 0.41

Table 1: SRCC/KRCC of all tested IQA measures and the mean of the rated images' z-scores, described in Section 2 in Experiment $1, 3, 4$. The top 3 performers have been printed in bold for each data set. Measures marked with * have been computed with implementations provided by the authors in Python, for all other measures the provided MATLAB implementations were used. PSNR and HaarPSI provided in both implementations identical results.

between the tested IQA measures and the ratings of the natural images are higher than the correlations with the ratings of the medical images, emphasizing the need to employ specifically task-targeted IQ measures.

The commonly employed FR measures PSNR and SSIM yield relatively low results, confirming previous studies that these two measures are not beneficial to broadly assess medical imaging tasks without further testing or adaption. On the other hand, HaarPSI showed exceptional behavior regarding generalizability, suggesting it acts as a robust measure of quality in addition to specifically tailored medical imaging IQ measures. FSIM, GMSD and MS-SSIM also showed relatively robust successful behavior across the data sets; LPIPS and DISTS succeeded in most tasks.

In summary, this is a further study to assess the adequacy of common IQA measures in medical imaging tasks with a specific focus on clinical chest X-ray scans and photoacoustic image reconstructions. We plan to extend this study and hope to provoke more research in this direction.

5. REFERENCES

[1] Z. Wang et al. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[2] S. Athar and Z. Wang. A comprehensive performance evaluation of image quality assessment algorithms. *IEEE Access*, 7:140030–140070, 09 2019.

[3] S. Kastryulina et al. Image quality assessment for magnetic resonance imaging. *Elsevier Medical Image Analysis*, 2022.

[4] L. S. Chow et al. Correlation between subjective and objective assessment of magnetic resonance (mr) images. *Magn Reson Imaging*, 34(6):820–831, Jul 2016.

[5] K. Ohashi et al. Applicability evaluation of full-reference image quality assessment methods for computed tomography images. *Journal of Digital Imaging*, 36(6):2623–2634, 2023.

[6] G. P. Renieblas et al. Structural similarity index family for image quality assessment in radiological images. *J Med Imaging (Bellingham)*, 4(3):035501, Jul 2017.

[7] B. Girod. Psychovisual aspects of image processing: What's wrong with mean squared error? In *Proceedings of the Seventh Workshop on Multidimensional Signal Processing*, pp. P.2–P.2, 1991.

[8] E. P. S. Zhou Wang and A. C. Bovik. Multi-scale structural similarity for image quality assessment. In *Proceedings of the 37th IEEE Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA*, 2003.

[9] Z. Wang and Q. Li. Information content weighting for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 20(5):1185–1198, 2011.

[10] K. Ding et al. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022.

[11] A. Balanov et al. Image quality assessment based on dct subband similarity. In *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 2105–2109, 2015.

[12] L. Zhang et al. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011.

[13] W. Xue et al. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, 23(2):684–695, 2014.

[14] R. Reisenhofer et al. A haar wavelet-based perceptual similarity index for image quality assessment. *Signal Process. Image Commun.*, 61:33–43, 2018.

[15] R. Zhang et al. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

[16] H. Ziaei Nafchi et al. Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator. *IEEE Access*, 4:5579–5590, 2016.

[17] H. Sheikh and A. Bovik. A visual information fidelity approach to video quality assessment. *The First International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2005.

[18] L. Zhang et al. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23(10):4270–4281, 2014.

[19] A. Mittal et al. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

[20] Z. Ying et al. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. pp. 3572–3582, 06 2020.

[21] A. Mittal et al. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013.

[22] A. K. Venkataramanan et al. A hitchhiker's guide to structural similarity. *IEEE Access*, 9:28872–28896, 2021.

[23] I. Selby. Github repository speedyiqa, March 2024.

[24] H. Sheikh et al. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006.

[25] H. R. Sheikh et al. Live image quality assessment database release 2, http://live.ece.utexas.edu/research/quality.

[26] D. Jayaraman et al. Objective quality assessment of multiply distorted images. In *2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pp. 1693–1697, 2012.

[27] J. Zbontar et al. fastmri: An open dataset and benchmarks for accelerated mri, 2019.

[28] A. Sriram et al. 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*, pp. 64–73. Springer, 2020.

[29] Z. Ramzi et al. XPDNet for MRI Reconstruction: an application to the 2020 fastMRI challenge. In *ISMRM*, pp. 1–4, 2021.

[30] H. Assi et al. A review of a strategic roadmapping exercise to advance clinical translation of photoacoustic imaging: From current barriers to future adoption. *Photoacoustics*, 32:100539, 2023.

[31] B. Cox et al. Quantitative spectroscopic photoacoustic imaging: a review. *J Biomed Opt*, 17(6):061202, Jun 2012.

[32] J. Grohl et al. Moving beyond simulation: data-driven quantitative photoacoustic imaging using tissue-mimicking phantoms. *IEEE Trans Med Imaging*, PP, Nov 2023.

[33] J. W. Pickering et al. Double-integrating-sphere system for measuring the optical properties of tissue. *Appl. Opt.*, 32(4):399–410, Feb 1993.

[34] K. Lei et al. Artifact- and content-specific quality assessment for mri with image rulers. *Medical Image Analysis*, 77:102344, 2022.

[35] M. Chun et al. Fully automated image quality evaluation on patient ct: Multi-vendor and multi-reconstruction study. *PLoS One*, 17(7):e0271724, 2022.