



# An Unsupervised Approach for Artifact Severity Scoring in Multi-Contrast MR Images

**Savannah P. Hays**<sup>1</sup> 

SHAYS6@JHU.EDU

<sup>1</sup> *Department of Electrical and Computer Engineering, Johns Hopkins University, USA*

**Lianrui Zuo**<sup>2</sup> 

LIANRUI.ZUO@VANDERBILT.EDU

<sup>2</sup> *Department of Electrical and Computer Engineering, Vanderbilt University, USA*

**Blake E. Dewey**<sup>3</sup> 

BLAKE.DEWEY@JHU.EDU

<sup>3</sup> *Department of Neurology, Johns Hopkins School of Medicine, USA*


**Samuel W. Remedios**<sup>4</sup> 

SREMEDI1@JHU.EDU

<sup>4</sup> *Department of Computer Science, Johns Hopkins University, USA*

**Jinwei Zhang**<sup>1</sup>

JWZHANG@JHU.EDU

**Ellen M. Mowry**<sup>3</sup> 

EMOWRY1@JHMI.EDU

**Scott D. Newsome**<sup>3</sup> 

SNEWSOM2@JHMI.EDU

**Aaron Carass**<sup>1</sup> 

AARON\_CARASS@JHU.EDU

**Jerry L. Prince**<sup>1,4</sup> 

PRINCE@JHU.EDU

**Editors:** Under Review for MIDL 2025

## Abstract

Quality assurance (QA) in magnetic resonance (MR) imaging is critical but remains a challenging and time-intensive process, particularly when working with large-scale, multi-site imaging datasets. Manual QA methods are subjective, prone to inter-rater variability, and impractical for high-throughput workflows. Existing automated QA methods often lack generalizability to diverse datasets or fail to provide interpretable insights into the causes of poor image quality. To address these limitations, we introduce an unsupervised and interpretable QA framework for multi-contrast MR images that quantifies artifact severity. By assigning a numerical score to each image, our method enables objective, consistent evaluation of image quality and highlights specific levels of artifact presence that can impair downstream analysis. Our framework employs an unsupervised contrastive learning approach, leveraging simulated artifact transformations, including random bias, noise, anisotropy, and ghosting, to train the model without requiring manual labels or preprocessing. A margin-based contrastive loss further enables differentiation between varying levels of artifact severity. We validate our framework using simulated artifacts on a public dataset and real artifacts on a private clinical dataset, demonstrating its robustness and generalizability for automatic MR image QA. By efficiently evaluating image quality and identifying artifacts prior to data processing, our approach streamlines QA workflows and enhances the reliability of subsequent analyses in both research and clinical settings.

**Keywords:** MRI, Quality Assurance, Artifact Detection

## 1. Introduction

Magnetic resonance (MR) imaging is a cornerstone of medical diagnostics and research, offering unparalleled insights into the structure and function of tissues (Bernstein et al., 2004). However, the quality of MR images can be significantly compromised by various artifacts, including bias field inhomogeneities, noise, motion, anisotropic resolution, and ghosting (Zaitsev et al., 2015). These artifacts not only degrade image interpretability but also affect downstream analyses, potentially leading to erroneous conclusions in both clinical and research settings (Zaitsev et al., 2015). Consequently, robust quality assurance (QA) of MR images is essential to ensure the reliability of data and analyses.

Artifacts in MR images can have far-reaching implications. For instance, in clinical settings, poor image quality can lead to misdiagnoses or necessitate costly and time-consuming rescans. In research, artifacts can bias analyses, particularly in multi-site studies where variability in image quality is compounded by differences in scanner hardware, acquisition protocols, and patient populations. Advanced techniques are particularly vulnerable to artifact-induced errors, underscoring the need for automated QA systems.

Existing MR image QA methods often involve manual inspection, which is labor-intensive, subjective, and prone to inter-rater variability (Esteban et al., 2017; Alfaro-Almagro et al., 2018). Automated methods frequently rely on preprocessing steps such as image registration or supervised training paradigms, which require large, labeled datasets and are susceptible to biases inherent in the training data (Esteban et al., 2017). Traditional quality control (QC) metrics like signal-to-noise ratio (SNR) and contrast-to-noise ratio (CNR) are limited in their ability to capture complex artifacts. Moreover, these approaches fail to provide interpretable outputs that can guide users in identifying corrupted images within their dataset.

MRIQC (Esteban et al., 2017) is a widely used tool that reports a range of image quality metrics (IQMs) for structural, functional, and diffusion-weighted MR images. For structural images, it supports both  $T_1$ -weighted ( $T_1$ -w) and  $T_2$ -weighted ( $T_2$ -w) modalities and provides quantitative assessments based on noise, entropy, and contrast-related measurements. While MRIQC generates detailed reports, interpreting these IQMs can be challenging, especially for large datasets. It remains difficult to establish consistent thresholds and optimal combinations of IQMs to determine whether an image should pass or fail QA.

Among MRIQC’s structural IQMs, we identified four key metrics that are relevant to our work: coefficient of joint variation (CJV), contrast-to-noise ratio (CNR), entropy focus criterion (EFC), and foreground-background energy ratio (FBER). CJV (Ganzetti et al., 2016) measures intensity variability between gray matter (GM) and white matter (WM) and is sensitive to head motion and intensity non-uniformity artifacts. CNR (Magnotta and Friedman, 2006) assesses how well the intensity distributions of GM and WM are separated. EFC (Atkinson et al., 1997) is based on Shannon entropy and quantifies the amount of ghosting and blurring induced by motion artifacts. FBER (Shehzad et al., 2015) measures the mean energy of voxel intensities within the brain relative to areas outside the brain.

Although MRIQC provides useful quantitative measures, determining a single threshold or combination of IQMs that universally defines poor-quality images remains a challenge. The complexity of these metrics and their dataset dependence further motivate the need for a more interpretable and automated QA framework.

In this work, we propose an automatic method for QA of MR images that addresses these limitations. Our approach generates interpretable artifact severity scores, where higher scores correspond to worse image quality and lower scores indicate better image quality. These scores can be used to set thresholds for automatic QA, enabling users to exclude poor-quality images prior to downstream processing. Our method does not require image preprocessing steps, making it computationally efficient and broadly applicable across different datasets. It can be directly used on Nifti files of many MR image modalities.

Central to our approach is an unsupervised training framework based on contrastive learning inspired by that of (Zuo et al., 2023a,b). By leveraging a data loader that applies a diverse set of realistic transformations—including random bias, noise, anisotropy, and ghosting—we simulate a wide range of artifact severities. This enables the model to learn discriminative features that correlate with image quality without the need for labeled training data. Contrastive learning, which pulls similar samples closer in feature space while pushing dissimilar ones apart, is particularly well-suited for this task as it enables the model to generalize to unseen artifact types. Furthermore, our method is capable of recognizing and penalizing poor resolution, a critical but often overlooked aspect of image quality.

Our approach is designed to address the growing reliance on large, multi-site datasets where artifact heterogeneity poses significant challenges. By providing interpretable artifact severity scores, our method promotes reproducibility and reliability in medical imaging studies. Moreover, it can be seamlessly integrated into automated pipelines for preprocessing large-scale datasets, significantly reducing the burden of manual QC and enabling more efficient use of resources.

We validate our approach by comparing it with MRIQC (Esteban et al., 2017), demonstrating its effectiveness in accurately identifying low-quality images and providing actionable insights for improving data quality. Our method offers a scalable and interpretable solution for automatic MR image QA, paving the way for more reliable and reproducible analyses in medical imaging. Our model is open source and is publicly available from: <https://github.com/UponAcceptance>.

## 2. Methods

### 2.1. Overview

Our model assigns an artifact score to an MR image, quantifying the level of artifacts present. The model was trained in an unsupervised fashion using a diverse set of simulated artifacts. It leverages triplet loss and the L2 loss to learn meaningful representations of image quality. The overall training and inference workflow is illustrated in Fig. 1.

### 2.2. Training Dataset

We trained our model using 297 structural MR volumes from the TRaditional vs. Early Aggressive Therapy for Multiple Sclerosis (TREAT-MS) pragmatic, clinical trial (NCT03500328). These scans were acquired from seven different imaging sites and included multiple contrasts: T<sub>1</sub>-w, T<sub>2</sub>-w, T<sub>2</sub>-w FLAIR, and proton density (PD) images. To improve generalization across scanners and imaging conditions, only high-quality images were included in the training dataset. Prior to training, the images were N4 bias field corrected (Tustison et al., 2010) to

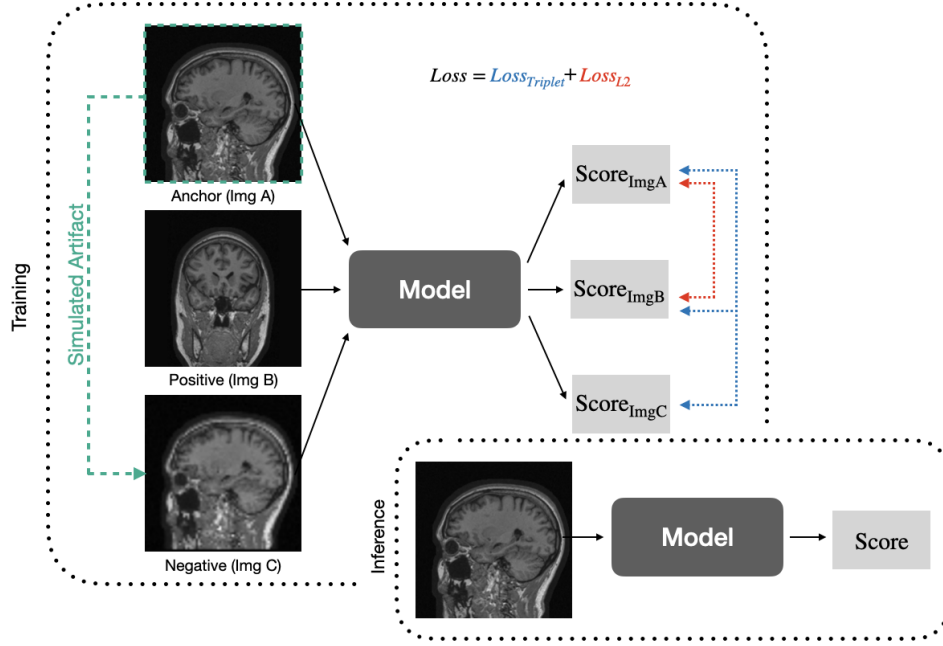


Figure 1: Training and inference workflow for our model. During training, three images are used to calculate the total loss. Img A and Img B are two different, clean image slices, while Img C is Img A with a randomly simulated artifact.

address intensity inhomogeneities and 2D acquisitions were super-resolved (Remedios et al., 2023) to ensure consistent resolution. These preprocessing steps were necessary for artifact simulation but are not required during inference, ensuring the model remains applicable to diverse datasets without additional preprocessing.

### 2.3. Model Architecture

Our model operates on 2D MR image slices, making it computationally efficient while allowing for a larger number of training samples. The architecture consists of two key components: a custom convolutional block and a primary encoder. The convolutional block is composed of two convolutional layers with  $3 \times 3$  kernels, each followed by instance normalization and a LeakyReLU activation function, allowing the network to effectively capture spatial features. The encoder stacks two of the custom convolutional blocks with increasing channel dimensions, followed by a single convolutional layer to reduce the feature dimensions and an adaptive average pooling operation to reduce the output to the desired dimension. An absolute value function is applied to enforce non-negative scores for interpretability. During inference, an MR volume is assigned a single score by averaging scores across the middle 60% of slices.

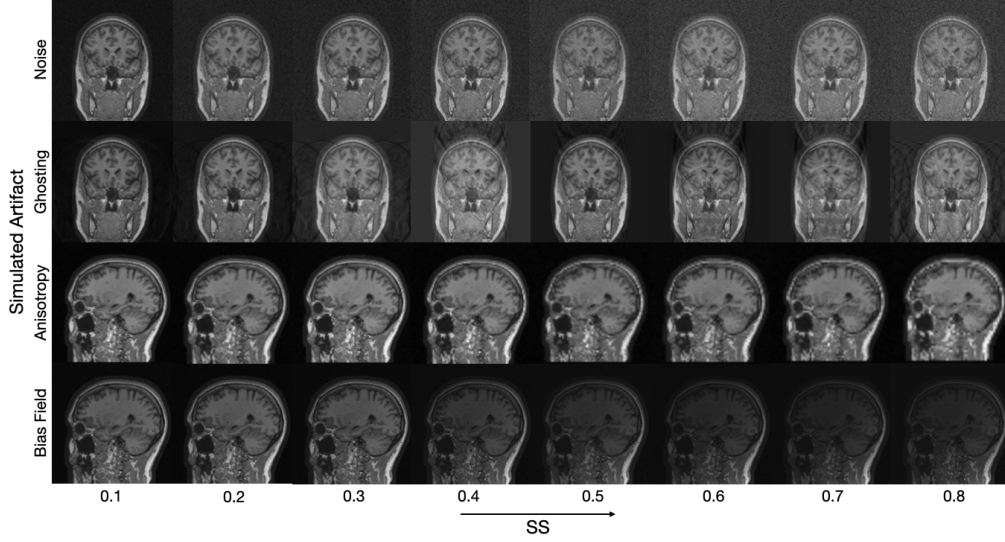


Figure 2: Increasing severity scores (SS) in the range  $[0, 1]$  (left to right) of the simulated artifacts (from top to bottom: noise, ghosting, anisotropy, bias) seen during training.

#### 2.4. Training and Artifact Simulation

To effectively train the model, we developed a data augmentation module that simulates common MR image artifacts in a controlled manner. These artifact transformations were implemented using the TorchIO library (Pérez-García et al., 2021) and include random noise, random ghosting, random bias field, and random anisotropy. Random noise introduces varying levels of noise in the images. Random ghosting simulates motion-induced ghosting with varying intensities and repetition. Random bias field introduces intensity inhomogeneities mimicking scanner-specific artifacts. Random anisotropy reduces spatial resolution to simulate anisotropic acquisitions. Each transformation is parameterized and assigned a calculated severity score (SS) in the range  $[0, 1]$ , as detailed in Table 1. The parameters are randomly sampled to ensure continuity in the artifact space exposing the model to diverse image degradations. Figure 2 illustrates example images with increasing severity scores. During training, the severity scores are used as the margin in the triplet loss to emphasize the relative differences between clean and artifact-degraded images. The model itself is only constrained to produce non-negative outputs. It is not constrained to produce outputs within a specific range. This design allows the model to flexibly assign scores based on the learned features.

#### 2.5. Loss Functions

During training, three images are passed through the network. The anchor is a clean image without any artifact simulation. The positive sample is another randomly selected clean image distinct from the anchor. The negative sample is the anchor with a randomly applied

Table 1: Each artifact and its parameters used in the severity score (SS). The parameters were uniformly sampled in the corresponding range to ensure continuity of the artifact space.

Artifact	Input	Parameters	Severity Score (SS)
Noise	std	$\mathcal{U}[0.005, 0.2]$	$\frac{\text{std} - 0.005}{0.2 - 0.005}$
Ghosting	num_ghosts	$\mathcal{U}\{2, \dots, 10\}$	$\frac{(\text{intensity} - 0.2) + \frac{\text{num\_ghosts}}{10}}{(1.5 - 0.2) + 1}$
	intensity	$\mathcal{U}[0.2, 1.5]$	
Bias Field	coefficients	$\mathcal{U}[0.01, 0.3]$	$\frac{\text{coefficients} - 0.01}{0.3 - 0.01}$
Anisotropy	scale	$\mathcal{U}[1, 4]$	$\frac{\text{scale} - 1}{4 - 1}$

artifact simulation. Two loss functions are employed to guide the model’s learning. First, we calculated an L2 loss. This measures the difference between the scores of the anchor and the positive sample. By minimizing this loss, the model ensures that clean images, which lack artifacts, are assigned consistently low scores, indicating high image quality. The second loss is the triplet loss (sometimes referred to as the contrastive loss). The triplet loss uses the scores from the anchor, the positive sample, and the negative sample. This loss encourages the score of the anchor to be similar to that of the positive sample, while the score from the negative sample is further apart. The triplet loss incorporates a margin that varies dynamically based on the artifact severity score of the negative sample. For high-severity artifacts, the margin is set higher, emphasizing a larger separation between clean and degraded images. For low-severity artifacts, the margin is smaller, reflecting subtler differences in quality. This adaptive margin ensures that the model’s scores are interpretable and directly correlated with artifact severity.

### 3. Experiments and Results

#### 3.1. Public Dataset

We first evaluated our model on a sample from the OASIS dataset (Marcus et al., 2007) ( $N = 20$ ). Based on our model scores during inference and visual interpretation, a value of 1 is a reasonable threshold for the model scoring. Anything with a score under 1 is considered to be of sufficient quality for subsequent processing. Anything with a higher score than 1 should be excluded from processing or undergo correction steps dependent on the artifact type. To assess the model’s ability to rank artifact severity, we simulated artifacts on the OASIS images resulting in evaluation of 120 images. We compared our model’s scores with MRIQC (v25.0.0) IQMs (Table 2). Our model consistently ranked images by artifact severity, whereas MRIQC’s IQMs showed inconsistent trends across different artifacts. Notably, MRIQC processing failed to report IQMs on images of T<sub>2</sub>-w contrast and high levels of simulated artifacts. This restricted our experiments to one modality and one artifact type and SS rather than testing a variety of severities per artifact.



Table 2: The specified artifact type is added with the specified severity score (SS) as outlined in Table 1. We compare the MRIQC statistics against the score from our model ( $N = 20$  for each artifact type). The arrows should indicate improving image quality. Our scores increase as SS increases. MRIQC IQM’s do not have a direct relationship with SS.

Artifact Type	SS	MRIQC				Ours ↓
		CJV ↓	CNR ↑	EFC ↓	FBER ↑	
None	0.0	$0.77 \pm 0.13$	$1.14 \pm 0.25$	$0.49 \pm 0.05$	$6962 \pm 2097$	$0.17 \pm 0.47$
Bias	0.1	$0.80 \pm 0.19$	$1.08 \pm 0.27$	$0.51 \pm 0.60$	$2367 \pm 1572$	$0.01 \pm 0.03$
Motion	0.3	$0.87 \pm 0.19$	$1.05 \pm 0.27$	$0.51 \pm 0.05$	$5501 \pm 2010$	$1.74 \pm 0.49$
Anisotropy	0.6	$0.78 \pm 0.05$	$1.32 \pm 0.23$	$0.52 \pm 0.05$	$9876 \pm 3326$	$2.32 \pm 0.48$
Ghosting	0.8	$1.03 \pm 0.18$	$0.82 \pm 0.12$	$0.53 \pm 0.06$	$4161 \pm 1852$	$2.34 \pm 0.60$
Noise	0.9	—	—	—	—	$3.54 \pm 0.06$

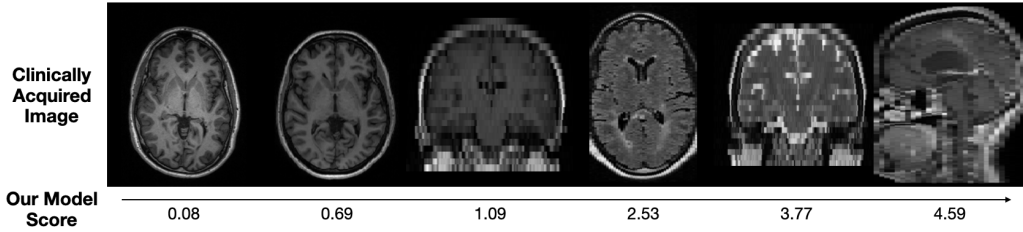


Figure 3: Clinically acquired images from the TREAT-MS dataset. Although images are acquired at various sites following a standardized protocol, several low resolution 2D acquisitions are observed.

### 3.2. Private Clinical Dataset

We further validated our model using 124 structural MR volumes from the TREAT-MS pragmatic, clinical trial (NCT03500328). These images were acquired following a standardized protocol but still exhibited substantial variation in image quality, particularly in resolution. Many images are 2D acquired at varying levels of resolutions. Figure 3 shows clinically acquired images ranked by their scores. Our model assigned scores close to 0 for high-quality images and higher scores for low-resolution 2D acquisitions, indicating strong detection of anisotropic resolution artifacts. Based on our findings, we recommend images with scores  $\leq 1$  are of sufficient quality for downstream processing. Images with scores  $> 1$  should be flagged for review or undergo artifact correction (e.g., super-resolution, background removal). This evaluation demonstrates that our model effectively identifies artifact severity in both public and clinical datasets, providing interpretable scores that facilitate automated MR image quality assurance.

#### 4. Discussion and Conclusion

In this study, we introduced an unsupervised, interpretable framework for QA in multi-contrast MR imaging. By training on diverse data with simulated artifact transformations, our model produces scores that correlate with image quality, offering an efficient and scalable solution for QA workflows. Unlike existing methods, our approach eliminates the need for manual labels or preprocessing. The framework is compatible with T<sub>1</sub>-w, T<sub>2</sub>-w, T<sub>2</sub>-w FLAIR, and PD images.

Although MRIQC can be directly run using their provided docker, we experienced many issues and limitations running this program. MRIQC requires data to be in BIDS format, this implies some level of preprocessing which we wish to avoid. For structural images, it can only retrieve IQMs for T<sub>1</sub>-w and T<sub>2</sub>-w images, while our approach can handle T<sub>1</sub>-w, T<sub>2</sub>-w, FLAIR, and PD images. However, MRIQC can also handle diffusion data which we cannot currently. There is also a large variation in computational time between the two methods. MRIQC does several steps like bias field correction, registration, and segmentation resulting in a process that takes several minutes per volume. Our model only requires about 1 second per volume. For many of the T<sub>2</sub>-w images and images with a simulated artifact, MRIQC failed to complete—which we found was a common experience in the MRIQC discussion forums with no known solution. This is the primary reason for our limited experiment and missing metrics for the noise artifact type in Table 2.

A key feature of the framework is the threshold for artifact scores. Based on empirical results, a threshold of 1 was determined as a reasonable cutoff for distinguishing between high- and low-quality images. During training, the margin in the triplet loss is derived from simulated artifact severity scores bounded between 0 and 1. This margin emphasizes the distinction between clean and artifact-ridden images while allowing flexibility for the model’s outputs during inference. Consequently, the model’s unrestricted scoring capability results in a broader range of scores (0 to 6) during real-world application. The simulated artifacts used in training currently focus on one artifact type at a time. In practice, MR images can exhibit multiple overlapping artifacts. We hypothesize that the model detecting multiple artifacts might be a reasoning for the broader range of scores during application. Future work would incorporate combinations of simulated artifacts during training, making the model more robust and reflective of real-world data. Additionally, incorporating artifact-specific weighting into the scoring process could improve interpretability by assigning higher penalties to artifacts that are particularly detrimental to downstream analysis.

The proposed framework has significant implications for large-scale studies, where it can streamline data preprocessing by automating QA and identifying low-quality images prior to analysis. Its unsupervised nature ensures adaptability across different datasets without reliance on labeled training data. This adaptability makes it particularly valuable for multi-site studies with heterogeneous imaging protocols and scanner hardware. Overall, this work lays a robust foundation for advancing automated MR image QA and emphasizes the importance of interpretable, unsupervised learning solutions in medical imaging. By providing artifact severity scores that are both flexible and actionable, this framework enables more efficient and reliable analyses, paving the way for enhanced QA workflows in both clinical and research settings.



## Acknowledgments

This material is partially supported by the Johns Hopkins University Percy Pierre Fellowship (Hays) and the National Science Foundation Graduate Research Fellowship under Grant No. DGE-2139757 (Hays) and Grant No. DGE-1746891 (Remedios). Development is partially supported by FG-2008-36966 (Dewey), CDMRP W81XWH2010912 (Prince), NIH R01 CA253923 (Landman), NIH R01 CA275015 (Landman), the National MS Society grant RG-1507-05243 (Pham) and Patient-Centered Outcomes Research Institute (PCORI) grant MS-1610-37115 (Newsome and Mowry). The statements in this publication are solely the responsibility of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee.

## References

- Fidel Alfaro-Almagro, Mark Jenkinson, Neal K Bangerter, Jesper L R Andersson, Ludovica Griffanti, Gwenaelle Douaud, Stamatis N Sotiropoulos, Saad Jbabdi, Miguel Hernandez-Fernandez, Emilie Vallee, et al. Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. *NeuroImage*, 166:400–424, 2018.
- David Atkinson et al. Automatic Correction of Motion Artifacts in Magnetic Resonance Images Using an Entropy Focus Criterion. *IEEE Trans. Med. Imag.*, 16:903–910, 12 1997. doi: 10.1109/42.650886.
- Matt A. Bernstein et al. *Handbook of MRI Pulse Sequences*. Elsevier Inc., September 2004. ISBN 9780120928613. doi: 10.1016/B978-0-12-092861-3.X5000-6.
- Oscar Esteban et al. MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLOS ONE*, 12(9):1–21, 09 2017.
- Marco Ganzetti et al. Intensity Inhomogeneity Correction of Structural MR Images: A Data-Driven Approach to Define Input Algorithm Parameters. *Frontiers in Neuroinformatics*, 10, 2016. ISSN 1662-5196. doi: 10.3389/fninf.2016.00010.
- Vincent Magnotta and Lee Friedman. Measurement of Signal-to-Noise and Contrast-to-Noise in the fBIRN Multicenter Imaging Study. *Journal of Digital Imaging*, 19:140–7, 07 2006. doi: 10.1007/s10278-006-0264-x.
- Daniel S. Marcus, T. R. Wang, Jonathan Parker, John G. Csernansky, John C. Morris, and Randy L. Buckner. Open access series of imaging studies (oasis): Cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19(9):1498–1507, 2007.
- Fernando Pérez-García et al. TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer Methods and Programs in Biomedicine*, 208:106236, 2021. doi: 10.1016/j.cmpb.2021.106236.

- Samuel W. Remedios et al. Self-supervised super-resolution for anisotropic MR images with and without slice gap. In *Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI) held in conjunction with the 26<sup>th</sup> International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2023)*, volume 14288, pages 118–128, 2023.
- Zarrar Shehzad et al. The Preprocessed Connectomes Project Quality Assessment Protocol - a resource for measuring the quality of MRI data. *Frontiers in Neuroscience*, 9, 01 2015. doi: 10.3389/conf.fnins.2015.91.00047.
- Nicholas J Tustison et al. N4ITK: Improved N3 bias correction. *IEEE Trans. Med. Imag.*, 29(6):1310–1320, 2010.
- Maxim Zaitsev et al. Motion artifacts in MRI: A complex problem with many partial solutions. *Jrnl. of Magnetic Resonance Imaging*, 42(4):887–901, 2015.
- Lianrui Zuo et al. HACA3: A unified approach for multi-site MR image harmonization. *Computerized Medical Imaging and Graphics*, 109(102285), 2023a.
- Lianrui Zuo et al. A latent space for unsupervised MR image quality control via artifact assessment. In *Medical Imaging 2023: Image Processing*, volume 12464, pages 278–283. SPIE, 2023b.