BioMedical Engineering OnLine

REVIEW

Open Access

Deep learning for the harmonization of structural MRI scans: a survey

Soolmaz Abbasi[1], Haoyu Lan[2], Jeiran Choupan[2], Nasim Sheikh-Bahaei[3], Gaurav Pandey[4] and Bino Varghese[3*]

*Correspondence:
bino.varghese@med.usc.edu

[1] Department of Computer Engineering, Yazd University, Yazd, Iran
[2] Department of Neurology, University of Southern California, Los Angeles, CA, USA
[3] Department of Radiology, University of Southern California, Los Angeles, CA, USA
[4] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Abstract

Medical imaging datasets for research are frequently collected from multiple imaging centers using different scanners, protocols, and settings. These variations affect data consistency and compatibility across different sources. Image harmonization is a critical step to mitigate the effects of factors like inherent differences between various vendors, hardware upgrades, protocol changes, and scanner calibration drift, as well as to ensure consistent data for medical image processing techniques. Given the critical importance and widespread relevance of this issue, a vast array of image harmonization methodologies have emerged, with deep learning-based approaches driving substantial advancements in recent times. The goal of this review paper is to examine the latest deep learning techniques employed for image harmonization by analyzing cutting-edge architectural approaches in the field of medical image harmonization, evaluating both their strengths and limitations. This paper begins by providing a comprehensive fundamental overview of image harmonization strategies, covering three critical aspects: established imaging datasets, commonly used evaluation metrics, and characteristics of different scanners. Subsequently, this paper analyzes recent structural MRI (Magnetic Resonance Imaging) harmonization techniques based on network architecture, network learning algorithm, network supervision strategy, and network output. The underlying architectures include U-Net, Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), flow-based generative models, transformer-based approaches, as well as custom-designed network architectures. This paper investigates the effectiveness of Disentangled Representation Learning (DRL) as a pivotal learning algorithm in harmonization. Lastly, the review highlights the primary limitations in harmonization techniques, specifically the lack of comprehensive quantitative comparisons across different methods. The overall aim of this review is to serve as a guide for researchers and practitioners to select appropriate architectures based on their specific conditions and requirements. It also aims to foster discussions around ongoing challenges in the field and shed light on promising future research directions with the potential for significant advancements.

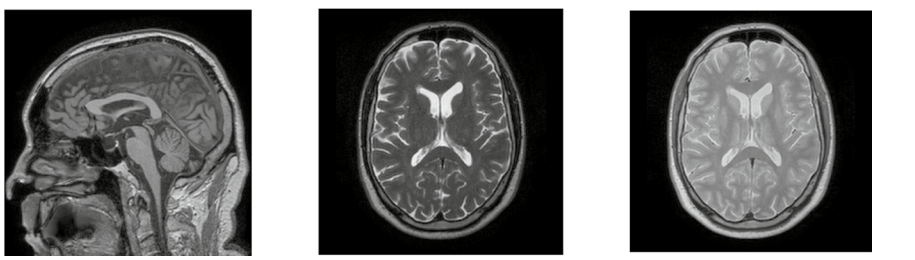
Keywords: Harmonization, Structural MRI, Generative adversarial networks, Variational autoencoders, Disentangled representation learning

Introduction

Neuroimaging techniques like magnetic resonance imaging (MRI), positron emission tomography (PET) and computerized tomography (CT) play a vital role in studying the brain's structure and function [1–3]. These medical imaging modalities provide invaluable insight for diagnosing neurological disorders, understanding brain development, and investigating neurodegenerative processes. In comparison to other neuroimaging modalities, MRI is known for its ability to generate excellent soft tissue contrast, making it instrumental for studying subtle tissue details including the brain or spinal cord [4]. Additionally, it is vital for many differential diagnoses of neurological disorders, including tumors [5], inflammatory conditions [6], and degenerative disorders [7]. MRI utilizes various imaging options and pulse sequences according to clinical needs. These sequences generate images with different contrasts, such as T_1-weighted, T_2-weighted, and PD-weighted (Fig. 1).

Variations in scanner hardware, imaging parameters, and acquisition protocols can lead to systematic differences in the appearance and quantitative measures derived from neuroimages. These divergences, if unaddressed, can reduce the statistical power of neuroimaging studies, limit the generalizability of findings across sites, and impede efforts to pool and analyze multi-site datasets. Consequently, there is an increasing demand for harmonizing neuroimaging data to mitigate unwanted inter-site and inter-scanner effects.

In recent years, data-driven harmonization techniques leveraging machine learning have gained significant traction. Deep learning models have demonstrated remarkable ability to capture complex data representations and transformations, making them well-suited for tackling neuroimaging harmonization challenges. By learning from data acquired across multiple sites and scanners, these models can disentangle biologically relevant signals from technical artifacts, enabling the generation of harmonized neuroimages or derived measures. However, compared to other neuroimaging harmonization efforts, one of the factors that make MRI harmonization more complex is that grayscale-based signal intensity in MRI lacks a standardized measure, unlike semi-quantitative measures such as standardized uptake value (SUV) of PET or quantitative measures such as Hounsfield units (HU) of CT [3]. This accentuates the variability in MRI, resulting in



T_1-weighted
Repetition time = 9.813
Echo time = 4.603
Number of Phase Encoding Steps = 192
Echo Train Length = 0
Reconstruction Diameter = 240
Flip Angle = 8

T_2-weighted
Repetition time = 8178.34
Echo time = 100
Number of Phase Encoding Steps = 187
Echo Train Length = 16
Reconstruction Diameter = 240
Flip Angle = 90

PD-weighted
Repetition time = 8178.34
Echo time = 8
Number of Phase Encoding Steps = 187
Echo Train Length = 16
Reconstruction Diameter = 240
Flip Angle = 90

Fig. 1 Example images acquired from Guy's Hospital using a Philips 1.5T system, sourced from the IXI dataset [8]

differences in contrast-to-noise ratio, temporal resolution, and spatial resolution. These variations have been found to affect the reliability of radiomic analysis [9].

With larger and more varied datasets available in open-source databases, researchers can conduct more in-depth analyses and leverage subtle details within the data. By involving more participants and tracking them over an extended period, researchers gain deeper insights into how the study impacts the subjects, such as the effects of a disease or the aging process [10]. This longitudinal approach also enables a better understanding of the underlying causes of certain diseases, ultimately helping identify optimal treatments, diagnostic methods, and care plans. While larger datasets lead to more accurate results, they also introduce increased variability in how the data are acquired. This acquisition variability issue is almost always present for various reasons. Even at a single site using one scanner, patients may require repeat scans on different MRI scanners if they need additional medical care. Furthermore, the imaging environment itself is prone to changes due to scanner upgrades or replacements occurring over the course of a study [11]. When dealing with images from multiple sites, the inconsistency becomes more pronounced and can lead to domain shift problems.

In the realm of medical imaging, addressing domain shift is crucial for ensuring the reliability and consistency of AI models across different imaging sites and scanners. Domain adaptation (DA) involves fine-tuning a model to perform well on data from a specific target domain, thereby enhancing its accuracy and applicability within that domain. On the other hand, domain generalization (DG) aims to develop models that generalize effectively to unseen domains, without specific access to target domain data during training [12]. An emerging technique, image harmonization, focuses on reducing inter-site variation to facilitate meaningful comparisons and analyses of images across diverse imaging environments.

Domain adaptation enhances model accuracy by injecting domain-specific knowledge into a general AI model. This process typically requires access to target domain data during training, allowing the model to better capture domain-specific features and nuances. However, challenges include the need for additional training data [13], inter-modality heterogeneity [14], and rigorous model evaluation. Despite these challenges, domain adaptation significantly improves model performance within specific domains.

In contrast, domain generalization tackles the broader challenge of developing models that can generalize well across unseen domains. This approach is particularly beneficial in scenarios where acquiring labeled data from every possible domain is impractical. By learning from multiple related domains during training without direct exposure to the target domain, domain generalization reduces the need for extensive labeling efforts and enhances the model's adaptability to new tasks. However, the lack of direct access to target domain data makes domain generalization more challenging than domain adaptation in practical applications [15].

Image harmonization focuses on minimizing inter-site variation in medical imaging, enabling consistent analysis and comparison across different imaging sites and scanners [16]. This technique has shown promising results in standardizing imaging data, thereby improving the reliability of downstream analysis and clinical decision-making. Unlike domain adaptation and generalization, which primarily focus on model training strategies, harmonization directly addresses data variability at the preprocessing stage. This

approach simplifies model deployment across diverse clinical settings but may require specialized algorithms tailored to specific imaging modalities.

In the context of MRI harmonization, a "traveling subject" refers to a person or phantom (an object designed to mimic certain properties of human tissues) that is scanned on multiple MRI scanners at different sites. The traveling subject provides a common reference point when being scanned on MRI scanners located at different sites or institutions, which may have different scanner models, field strengths, or acquisition protocols. The data from scanning the traveling subject across sites are used to characterize and correct for scanner-specific variations in the MRI data. The use of traveling subjects is a crucial step in many MRI harmonization pipelines, especially for large multi-site neuroimaging studies, as it provides a way to quantify and mitigate scanner-related effects that could otherwise confound the analysis and interpretation of the pooled dataset.

On a related note, in some cases paired MRI data are used for image harmonization. While the two concepts are related, they are fundamentally different in their application. The purpose of the traveling subject is to directly measure and characterize the scanner-specific variations that need to be harmonized. Paired MRI data, however, refer to having two sets of MRI scans acquired from the same subject, which can be done either within a single scanner (intra-site) or across different scanners (inter-site). Intra-site paired data typically include different imaging modalities (e.g., T1-weighted and T2-weighted images) acquired from the same subject on the same scanner. This intra-site pairing facilitates training by providing complementary information from different modalities. Inter-site paired data, which involve scanning the same subject on both a "source" and a "target" scanner, help in harmonizing data between different scanners by offering corresponding data points under different scanner conditions.

While a traveling subject provides a common reference scanned across all scanners, paired data have separate source and target scans. Also, while traveling subjects are fewer in number but scanned widely across different sites, paired data can involve the full set of study subjects scanned at two sites. Lastly, while traveling subjects directly measure scanner effects, paired data rely on the corresponding subject scans to estimate the scanner-specific transformations required for harmonization. Therefore, while a traveling subject provides a direct measurement of scanner effects, paired data provide a way to estimate and apply those effects to each subject's data during the harmonization process. To provide an example of image harmonization, Fig. 2 demonstrates the traveling subject from six different scanners of the SRPBS Multi-disorder MRI Dataset [17]. As observed, the image contrast varies across scanners, which is known to affect downstream tasks such as tissue segmentation and disease classification.

In 2022, a comprehensive overview was presented on the various approaches for radiomics harmonization [18]. This review study classified these methods into two categories: image-based harmonization and feature-based harmonization. Image-based harmonization techniques are applied directly to the images before extracting radiomics features, while feature-based harmonization aims to reduce the differences between the extracted features themselves. The choice of these techniques can be constrained by the number of samples available for analysis. The review concluded that, up to that point, none of these harmonization methods had been definitively established as the most effective approach within the analysis process.

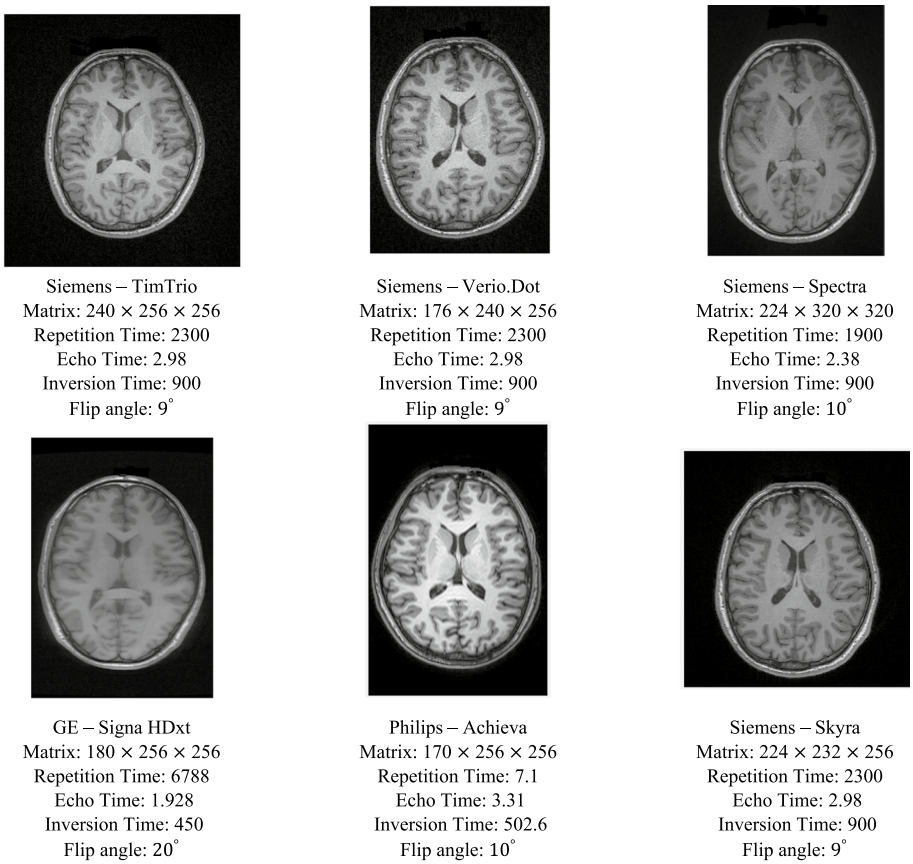


Fig. 2 Traveling subject from different scanners of the SRPBS Multi-disorder MRI Dataset [12] with detailed acquisition parameters. The contrast has been changed across scanners

Statistical methods have been used for feature-based harmonization [19]. These methods offer fine-grained adjustments, enabling researchers to target specific features for precise correction while maintaining interpretability [20]. The primary advantages of statistical methods include the following:

- Interpretability: Researchers can understand which specific features are being adjusted.
- Targeted Adjustments: Precise correction of predefined features is possible, making these methods ideal for datasets with well-understood relevant features.

However, these methods also have notable limitations:

- Dependence on Predefined Features: They rely on predefined features for correction, potentially limiting their effectiveness in complex datasets where relevant features are poorly defined [21].
- Potential for Overfitting: In scenarios with high variability and noise, statistical methods might overfit to the training data.

Image-based approaches particularly leverage deep learning models for accurate mapping between source and target MRI images [21]. Deep learning has revolutionized the field of medical imaging by enabling automated analysis, diagnosis, and prognosis through the extraction of meaningful features from medical images. Image classification, segmentation [22], image reconstruction [23], and image registration [24] are just a few successful applications of deep learning in medicine. Along these lines, recent development of various neural network architectures that have been proposed specifically for the task of medical image harmonization serve as the primary focus of this review article. The main advantages of image-based approaches are as follows:

- Automation and Scalability: They can automatically learn complex mappings without requiring predefined features.
- Handling Complex Variability: They excel in scenarios with complex and high-dimensional data, effectively capturing intricate patterns and relationships.

However, these methods also face challenges:

- Black-Box Nature: Deep learning models are often criticized for their lack of interpretability, making it difficult to understand how corrections are being made.
- Data Requirements: They typically require large amounts of training data and computational resources, which may not be available in all settings.

These methods can be selected according to the specific characteristics of their datasets and research goals. Statistical methods are preferable when interpretability and targeted feature correction are paramount, especially in well-understood datasets. In contrast, image-based approaches are better suited for complex datasets with poorly defined features and where automated, scalable solutions are needed.

Combining statistical and image-based approaches could potentially yield better performance by leveraging the strengths of both methods. For instance, a hybrid model could use statistical methods to correct well-defined features while employing deep learning to handle more complex, undefined variability. This combination could enhance both the accuracy and interpretability of harmonization efforts. Future research could explore integrated frameworks that synergistically use both approaches, potentially leading to more robust and generalizable harmonization techniques.

Wen et al. provided an examination of image harmonization methods for brain MRI data, with a particular focus on machine learning (ML)-based approaches used in both explicit and implicit ways [25]. They reported that a uniform imaging dataset can be achieved by implementing explicit methods that harmonize intensity values and image-derived metrics, or by using implicit methods to improve the performance of a downstream task. They also noted that traveling subject datasets are crucial for the effective implementation of explicit harmonization, as these enable the machine learning models to avoid learning biased information from the population. However, contemporary traveling subject datasets have limitations in terms of size and issues

related to scan–rescan reliability, which can hinder the performance of the ML models. In contrast, implicit methods do not require a traveling subject dataset. Researchers only need to determine the source and reference domains to develop the machine learning algorithm for harmonizing the MRI scans.

Different harmonization solutions, including the image domain and feature domain, have been discussed in another survey [26]. The image domain harmonization encompasses acquisition protocols and data augmentation, while the feature domain category includes statistical normalization, Combat [27], and deep learning. GAN, Neural Style Transfer (NST), and their combination were discussed as deep learning-based harmonization techniques.

Hu et al. investigated both statistical and deep learning methods for harmonization [20]. The study noted that statistical techniques could provide robustness and effectiveness, especially in scenarios with smaller sample sizes or when dealing with confounding factors. Conversely, deep learning models may be better suited to handle the complex nature of image-level data, which poses significant challenges for conventional statistical approaches. A recent book chapter by Zuo et al. [28] reviewed disentangled representation learning methods for MR image harmonization, demonstrating how disentangled representations can be learned through both supervised and unsupervised image-to-image translation techniques.

- Literature search

To narrow down our primary focus, we conducted a comprehensive literature search that met specific inclusion criteria including applying harmonization technique on structural MRI, employing deep learning methods, and in some papers harmonizing images for downstream tasks. The search was organized across the PubMed database using the terms "MRI harmonization" AND "deep learning," "MRI harmonization" AND "structural MRI," and "MRI harmonization" AND "deep learning" AND "structural MRI" through February 2024. This search returned 285 papers and the duplicate papers were excluded in the first phase. The same keywords "disentanglement representation learning for MRI" and "MRI harmonization using transformers" were used for searches on Google Scholar. Studies that were excluded encompass those that (1) were applied on the PET, fMRI, dMRI, or anything other than structural MRI, (2) did not employ a deep learning architecture, and (3) did not belong to the aim and scope of this review according to titles and abstracts. After applying criteria to the literature search and conducting screening, we incorporated a total of 38 papers into our study.

All the papers included in this study were published between 2019 and 2024, with approximately 26% from 2022 and 34% from 2023. The identified studies were analyzed in terms of their network architecture, learning algorithm, network framework, and network output. Based on the deep network architecture, the approaches can be classified into U-Net, GANs, VAEs, flow-based generative models, transformers, and custom Networks. Some networks learn via disentangled representation learning, which involves extracting and separating meaningful features or factors from the MRI data associated with different imaging characteristics. Among the studies, 29% utilized custom networks with or without disentanglement representation, 24% employed GANs, 13% used U-Net, and 13% were based on VAEs. Additionally, two of

them relied on transformers, and three utilized flow-based generative models, while the rest employed a combination of two networks.

The majority of articles, about 74%, relied on publicly available datasets; however, 26% used local datasets. In addition to that about 53% of the articles proposed a 3D model or worked with 3D images. It is noteworthy that 66% of the papers employed harmonization solely on T1-weighted images, while the rest also utilized other contrasts. Furthermore, 63% of the studies conducted tests on the harmonized images for downstream tasks, whereas the remainder focused merely on image harmonization. Additionally, 15 of the articles have publicly available source code.

- Motivations

Based on the survey of prior efforts, the motivation behind this review is outlined as follows:

1) A thorough, wide-ranging evaluation of deep learning-based methods for harmonizing structural MRI data across various benchmarks is lacking. Such a comprehensive investigation could shed light on the advantages and limitations of existing approaches in this domain.
2) While previous review articles have examined various harmonization techniques, they have not covered harmonization strategies that leverage transformers, flow-based generative models, or custom-designed neural network architectures.
3) A comprehensive comparison of large-scale brain imaging datasets for training and evaluating harmonization methods has not been conducted.
4) There has been a lack of detailed discussion surrounding the evaluation metrics used for assessing MRI harmonization methods, including considerations for scenarios with or without the presence of traveling subject data.

Driven by the aforementioned motivations, the primary goals of this study are to address existing gaps, elucidate the current limitations for a comprehensive comparison of harmonization techniques, and analyze the pros and cons of deep learning architectures for MRI harmonization. As such, our focus is solely on deep learning-based harmonization techniques. The design of networks for MRI harmonization can draw inspiration from networks used for domain adaptation [29] and image-to-image translation [30] in various computer vision applications. However, due to the complex structure of the brain and the intensity levels of MRI images, MRI harmonization networks have fundamental differences from networks in other applications. MRI harmonization aims to preserve the anatomical (content) information of the source image while transforming the contrast (style) to the target domain.

- Contributions

SA performed literature search and wrote first draft of the manuscript. BV reviewed the first draft and conceived the project. SA, HA, JC, NSB, GP, and BV reviewed all subsequent drafts and approved the final manuscript.

1. This study provides a categorization of state-of-the-art deep learning-based MRI harmonization techniques based on their architectural design (e.g., GAN, U-Net,

VAE), learning algorithms, network frameworks, and network outputs. This systematic classification offers valuable insights into the appropriate design considerations for developing effective harmonization networks.

2. This study conducts a comprehensive comparison of well-established large-scale datasets based on their fundamental characteristics, such as the number of participants, age range, target challenges, scanner types, and image modalities. Additionally, it discusses the lack of a dedicated harmonization dataset that addresses specific challenges faced by the medical imaging community.
3. This study examines the commonly used evaluation metrics in explicit image harmonization techniques, such as the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). It highlights the limitations of these evaluation criteria, pointing out their potential shortcomings in enabling accurate comparisons across different harmonization methods.

The rest of the paper is structured as follows. Sect. "The Overview of Harmonization" introduces harmonization. Sect. "Deep Learning-based Harmonization Taxonomy" categorizes different harmonization techniques. While acknowledging that certain papers may fit into more than one category, they are classified based on their predominant focus. Sect. "Applicability and Limitations of Harmonization" discusses the applicability and limitations of harmonization in MRI neuroimaging in a general context. Sect. "Discussion" analyzes the strengths and weaknesses of the various image harmonization methods. Sect. "Conclusion and Future Direction" provides concluding remarks.

The overview of harmonization

The MRI harmonization is formulated as a transformation that maps images from a source domain X_s to a target domain X_t. Let x_s denote an MRI image within the source domain, and x_t signify its counterpart within the target domain. The objective of MRI harmonization is to learn a mapping function f :

$$f : X_s \rightarrow X_t. \tag{1}$$

The function f should preserve the anatomical content of the source image while aligning its contrast to match that of the target domain. This can be expressed mathematically as follows:

$$x_t = f(x_s) + \in \tag{2}$$

where ϵ represents the residual error introduced during the harmonization process. To learn the mapping function f, a deep learning model using adversarial and auxiliary loss functions can be employed to minimize the discrepancy between the distribution of synthesized images x_t and real images from the target domain X_t.

There are several key factors to consider when evaluating image harmonization techniques. This section will discuss these factors, including the types of datasets used for training, the scanners employed for image acquisition (if relevant), and the metrics used to assess the success of harmonization approaches. Each of these aspects can influence the effectiveness and suitability of a particular method and can present unique challenges (Fig. 3).

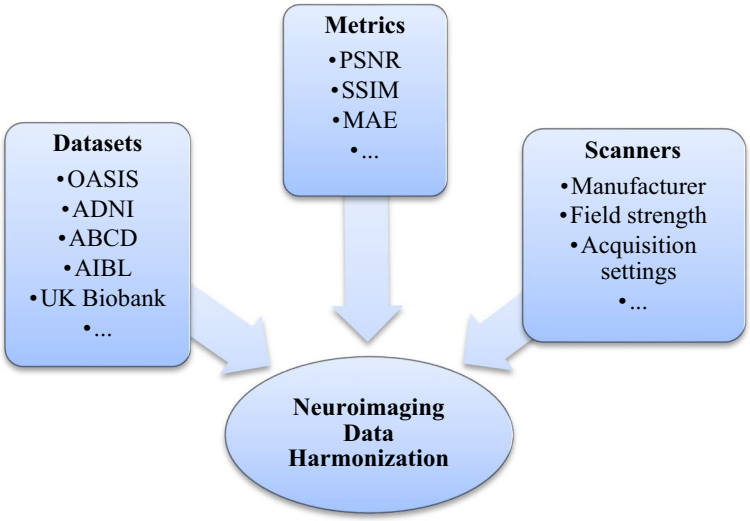


Fig. 3 Overview of Harmonization

Datasets

Pooling MRI data from diverse sources is crucial for high-powered and large-scale brain imaging studies. These sources include different sites, scanners, and acquisition protocols. Fortunately, several large neuroimaging datasets already exist, such as the UK Biobank [31], ABIDE [32], and ADNI [33]. Apart from heterogeneities in data due to scanners and modalities, the following complexities further complicate the harmonization process:

- Disease complexity: While some harmonization methods use healthy brain scans, brain diseases pose additional challenges. Subtle changes in diseased brains are crucial for studies like Alzheimer's disease progression, and harmonization techniques must not obscure this information.
- Segmentation and classification complexity: When diverse images from different scanners are harmonized, it must allow for precise and accurate segmentation (identification of specific brain structures) and classification (grouping images based on disease state).

Biological covariate balance complexity: Ensuring a balanced distribution of biological covariates such as gender and age in training datasets remains important. This is critical if the sample size of the training data is limited. A balanced representation of biological covariates and accounting for their difference helps achieve more generalizable and realistic results.

While Table 1 provides an overview of common datasets used in harmonization research, these datasets present challenges due to their inherent diversity. These large-scale datasets are often collected from multiple sites across different countries, and harmonization studies frequently utilize only a subset of the data or even employ entirely different datasets. This variation in data origin and usage hinders standardization and makes direct statistical comparisons between studies difficult.

Table 1 Large-scale neuroimaging datasets

Dataset name	No. participants	Scanners	Modalities	Gender % female	Age range	Target challenge
UK Biobank [31]	85,000	Standard Siemens Skyra 3T Siemens 32-channel RF	T1-weighted, rfMRI, T2_FLAIR, MRI, SWI, tfMRI	50	40–69	Cardiovascular diseases, cancer, diabetes, neurological disorders
ABIDE I [32] (Autism Brain Imaging Data Exchange)	1112	17 International sites	R-fMRI	–	7–64	Autism spectrum disorder (ASD)
ADNI [33]	The number of participants has evolved over time	–	PET and MRI	–	a broad spectrum, including older adults	Alzheimer's disease (AD)
IXI Dataset [8]	600	Philips 3T Philips 1.5T GE 1.5T	T1, T2, and PD-weighted images MRA images Diffusion-weighted images [15 directions]	50	20–90	healthy subjects
ABCD (Adolescent Brain Cognitive Development) [34]	11,875 children and adolescents	21 research sites	MRI-fMRI	–	9–10 and following participants for 10 years	evaluating behavioral and brain development
AIBL (Australian Imaging Biomarkers and Lifestyle Study of Ageing) [35]	3045	–	PET and MR imaging	–	older adults and individuals in late adulthood	Alzheimer's disease (AD)
OASIS-1[36]: Cross-sectional MRI Data	416	1.5T Vision scanner (Siemens, Erlangen, Germany)	T1-weighted	–	18 to 96	Alzheimer's disease and dementia
OASIS-2[37]: Longitudinal MRI Data	150	1.5-T Vision scanner (Siemens, Erlangen, Germany)	T1-weighted	–	60 to 96	
OASIS-3 [38]: Longitudinal Multimodal Neuroimaging	1379	Siemens scanner models 1.5T, TIM Trio 3T, and Biograph mMR PET-MR 3T	T1-weighted, T2-weighted, FLAIR, and TSE, PET	–	42 to 95	

'-' indicates that certain information is not provided in the dataset. "-" indicates that the information is not provided

Scanners

The global estimation for the number of MRI machines was around 36,000 with 2500 machines being manufactured annually [39]. While MRI is a vital tool in medical diagnosis, their inherent limitations can affect image quality and accuracy. These limitations arise from two main sources:

Natural effects

The MRI process itself and the equipment involved are susceptible to natural phenomena that can cause issues. Examples include magnetic field inhomogeneity, gradient nonlinearities, and variations in radiofrequency (RF) coil sensitivity. These effects manifest as variations in image intensity, distortions, and artifacts, ultimately impacting image quality and potentially leading to misdiagnosis. Fortunately, manufacturers are constantly working on mitigating these natural effects to improve the reliability and quality of MRI systems.

Acquisition settings

The specific settings used during an MRI scan significantly impact the resulting images. Factors like pulse sequence type, repetition time (TR), echo time (TE), and flip angle contribute to the appearance and accuracy of the scan. Table 2 provides a detailed breakdown of these factors. Selecting and optimizing these settings is crucial to minimize inherent limitations like noise, artifacts, and inconsistencies in image quality. By tailoring these settings to specific diagnostic needs, healthcare professionals can ensure high-quality imaging outcomes.

Metrics

Accurately assessing the effectiveness of harmonization algorithms is critical. This subsection delves into several essential metrics used for this purpose:

Peak signal-to-noise ratio (PSNR)

This metric measures the ratio between the maximum possible signal (image intensity) and the corrupting noise that affects image quality. Higher PSNR values generally indicate better harmonization and is essential for accurate diagnosis and assessment in clinical settings [46]. Improved PSNR aims to preserve important anatomical details, aiding radiologists in making precise evaluations. Thus, the PSNR ratio is a highly effective quality indicator in evaluating the effectiveness of harmonization architecture.

This ratio is derived from the difference between the original image and the harmonized version. The PSNR is calculated using the following equation:

$$PSNR = 10 \log_{10} \left(\frac{R^2}{MSE} \right), \tag{3}$$

where R demonstrates the maximum fluctuation in the input image data type.

Mean absolute error (MAE)

MAE calculates the average of the absolute differences in pixel intensity values between the original and harmonized images. Lower MAE values indicate that the harmonized image

Table 2 Technical factors that can affect the image quality of MRI scan

Factors	Descriptions
Field strengths	Magnetic field strength, measured in Tesla (T), is a key characteristic of MRI scanners. Higher field strengths translate to stronger magnetic fields, offering several advantages: • **Improved signal-to-noise ratio:** This translates to clearer images with less interference • **MR spectroscopy capability:** Enables the study of chemical composition within tissues • **Faster scan times:** Reducing patient discomfort and improving workflow • **High-resolution images:** Providing greater detail for accurate diagnosis However, despite these benefits, the high cost of high-field MRI scanners remains a significant barrier to their widespread adoption[40]
Field of view	The field of view (FOV) in MRI refers to the specific region of the patient's body captured during the scan. It essentially determines the size of the resulting image and the anatomical structures included. FOV is typically measured in centimeters (cm) [41] **Choosing the right FOV:** • **Larger FOV:** Provides a broader view of the anatomy, which can be helpful for initial examinations or when looking at larger structures. However, this may come at the expense of: o Lower spatial resolution: This means the image may appear less detailed, potentially making it harder to see fine structures • **Smaller FOV:** Offers higher spatial resolution, resulting in a more detailed image. This is preferable when examining smaller structures or needing a magnified view of a specific area. However, it may not capture the entire region of interest Finding the optimal FOV involves balancing the desired level of detail with the need to image the entire area of interest
Slice thickness	Slice thickness refers to the depth of each individual image layer acquired during an MRI scan. It directly impacts the image resolution along the z-axis, which represents depth within the body [42] **The impact of slice thickness:** • **Thinner slices:** o Advantages: Offer greater detail and improved visualization of small structures o Disadvantages: May require longer scan times due to the increased number of slices needed to cover the same region. They can also lead to a larger overall data volume, which might require additional storage space • **Thicker slices:** o Advantages: Enable faster scan times by capturing a larger area with fewer slices. They also result in a smaller data volume o Disadvantages: May lead to lower image resolution, potentially obscuring fine details
Echo time	Echo time (TE) is a critical parameter in MRI, measured in milliseconds (ms). It represents the time interval between the application of the initial radiofrequency (RF) pulse and the peak of the echo signal received by the scanner [43]. TE plays a significant role in determining the image contrast, which highlights differences in tissue properties **The impact of TE:** • **Short TE values:** o Result in images with a predominantly T1-weighted contrast. This means the contrast between tissues is primarily influenced by their longitudinal relaxation time (T1). Tissues with shorter T1 values tend to appear brighter in these images o Short TE values are often used to highlight anatomical details with high T1 values, such as fat and blood vessels
Repetition time	Repetition time (TR) is another crucial parameter in MRI, measured in milliseconds (ms). It represents the time interval between the application of consecutive radiofrequency (RF) pulses in a single pulse sequence. TR plays a significant role in determining the signal strength and image contrast
Inversion Time	Inversion time (TI) is a parameter specific to certain MRI pulse sequences, particularly those utilizing inversion recovery. It is measured in milliseconds (ms) and refers to the time interval between the application of a special radiofrequency (RF) pulse called an inversion pulse and the subsequent start of data acquisition.[44]
Gradient strength	Magnetic field gradients are controlled variations in the strength of the main magnetic field used in MRI scanners. These gradients are not uniform and are applied along specific directions within the scanner
Gradient orientation	Magnetic field gradients in MRI modulate the strength of the main magnetic field, but they do not have a single direction. Instead, these gradients create controlled variations in intensity along specific axes (e.g., X, Y, Z) within the scanner. This allows for spatial encoding of the signal, which essentially translates to determining the origin of the signal within the patient

Table 2 (continued)

Factors	Descriptions
Flip angle	The flip angle is a crucial parameter in MRI that governs the signal intensity in the resulting image. It is measured in degrees and reflects the angle at which the main magnetic field of the scanner nudges protons (tiny spinning particles within tissues) away from their equilibrium position. This nudge is achieved by applying brief radiofrequency (RF) pulses **Impact of flip angle:** **· Larger flip angle:** o Increases the signal intensity: This translates to brighter images with a higher signal-to-noise ratio, potentially offering greater detail **· Smaller flip angle:** o Decreases the signal intensity: This results in dimmer images but can improve contrast between tissues with different relaxation times Choosing the optimal flip angle involves balancing the desired signal intensity with the need for specific types of contrast in the image [45]

closely resembles the original, preserving essential diagnostic information. This accuracy in intensity values is crucial for tasks in image harmonization, where the goal is to closely approximate the original data. This fidelity is important for various applications, including diagnostic assessments and other clinical evaluations where accurate representation is paramount.

$$MAE = \frac{\sum_{M,N} |I_1(m,n) - I_2(m,n)|}{M \times N}, \tag{4}$$

where M and N represent the row and column of the input image, respectively.

PSNR and MAE both quantify the difference between the original and harmonized images, but they do so in different ways. PSNR is a logarithmic measure that emphasizes larger differences, making it useful for detecting significant deviations in image quality. MAE, on the other hand, provides a linear measure of average differences, offering a straightforward assessment of overall image fidelity. Together, these metrics provide a comprehensive view of image quality by highlighting both large and small discrepancies.

Structural similarity index measure (SSIM)

SSIM goes beyond just measuring noise levels. It compares the overall structural similarity between two images, considering luminance, contrast, and structure. Clinically, higher SSIM values suggest that the harmonized image retains the structural integrity of the original, which is critical for identifying subtle anatomical changes [47]. This metric helps ensure that the harmonization process does not distort important clinical features, thereby supporting accurate diagnosis and treatment planning.

The SSIM is based on illumination, contrast, and structural terms (Eqs. 6–8). In these equations, μ and σ are local mean and standard deviation, respectively.

$$SSIM(x,y) = \left[l(x,y)\right]^{\alpha}.\left[c(x,y)\right]^{\beta}.\left[s(x,y)\right]^{\gamma}, \tag{5}$$

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \tag{6}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \tag{7}$$

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}. \tag{8}$$

SSIM complements PSNR and MAE by focusing on structural information rather than just pixel-wise differences. While PSNR and MAE quantify numerical discrepancies, SSIM evaluates how well the harmonized image preserves the structural integrity of the original, considering luminance, contrast, and structural similarity. This combination ensures that both numerical accuracy and structural fidelity are assessed.

In addition to these core metrics, evaluation techniques specific to generative models might also be employed for certain harmonization approaches.

Recent research [48] has highlighted a crucial disconnect between how well harmonization algorithms perform according to traditional metrics and their actual impact on downstream applications. This suggests that common image similarity metrics like PSNR and SSIM might not fully capture the effectiveness of harmonization in improving compatibility across different datasets (cross-domain consistency). As a result, there is a growing need to re-evaluate current metrics to ensure they accurately assess the success of harmonization techniques.

To address this limitation, new assessment methods have been proposed [49]. These methods focus on two key aspects of harmonization:

- Intensity Harmonization: The Wasserstein Distance (WD) is used to measure how well intensity levels are harmonized. It achieves this by calculating the movement of histograms between images, essentially quantifying how similar the intensity distributions become. Clinically, ensuring consistent intensity levels across images from different scanners or protocols can improve the reliability of quantitative measurements, such as volumetric analysis.
- Anatomy Preservation: To evaluate how well anatomical structures are preserved during harmonization, segmentation is performed on both the original and harmonized images. The relative Absolute Volume Difference (rAVD) is then calculated to compare the segmentation results. This provides a measure of how closely the harmonized image retains the anatomical information from the original image. Accurate anatomical preservation ensures that clinical assessments, such as tracking disease progression or planning interventions, are based on reliable data.

Summarizing, WD measures the similarity of intensity distributions, ensuring that the overall intensity levels are harmonized across images. rAVD, on the other hand, evaluates how well anatomical structures are preserved during harmonization. By combining these metrics, we can assess both the consistency of intensity harmonization and the preservation of anatomical details, offering a dual perspective on harmonization effectiveness.

By incorporating these more specific metrics alongside traditional ones, researchers can gain a more comprehensive understanding of how different harmonization algorithms perform and their suitability for real-world applications.

In many real-world applications, "ground truth" data, which represents the perfect or ideal outcome, may not be available. This is especially common in situations involving unpaired datasets, where there is no direct correspondence between elements. Image generation tasks frequently encounter this scenario.

To address this challenge, researchers have developed metrics that assess image quality without relying on ground truth. These metrics offer valuable insights for objectively evaluating model performance:

- Inception Score (IS): A popular metric for assessing the quality and diversity of images generated by models like Generative Adversarial Networks (GANs) [50]. IS considers both the distinctiveness of individual images and the overall variety within the generated set.
- Fréchet Inception Distance (FID): This metric compares the similarity between the distribution of real images and the distribution of generated images in a feature space extracted by a deep learning model [51] (often the Inception network). Lower FID values indicate better alignment between the real and generated data distributions.
- Kernel Inception Distance (KID): Similar to FID, KID measures the discrepancy between the feature representations of real and generated images. However, KID utilizes kernel methods for the comparison. It is primarily used to assess the overall quality of generated images [52].
- Learned Perceptual Image Patch Similarity (LPIPS): This metric goes beyond basic image statistics and leverages a pre-trained deep learning model to assess the perceptual similarity between images. LPIPS considers factors like human visual perception and aims to quantify how similar two images appear to the human eye [53].

By employing these metrics alongside traditional methods, researchers gain a more comprehensive understanding of how models perform in scenarios lacking ground truth data.

While Inception Score (IS), Fréchet Inception Distance (FID), Kernel Inception Distance (KID), and Learned Perceptual Image Patch Similarity (LPIPS) are valuable tools for assessing the quality and diversity of generated images, they may not directly translate to evaluating image harmonization techniques because of the following:

- Differing Goals: Image generation aims to create entirely new, realistic images, whereas harmonization focuses on aligning existing images while preserving their anatomical content. Metrics like IS and FID prioritize diversity and novelty, which might not be desirable in harmonization.
- Focus on Anatomy: Preserving anatomical accuracy is paramount in harmonization. These metrics, however, do not explicitly assess how well anatomical structures are maintained during the process.

However, there might be situations where these metrics could be incorporated into a broader evaluation framework for harmonization, for example, target domain matching. In this scenario, if the harmonization process involves generating synthetic images to match a specific target domain (e.g., MRI scans from a particular scanner model), these metrics could be used to measure the discrepancy between real images from the target domain and the synthetic images produced during harmonization. This could provide insights into how well the harmonized images capture the characteristics of the target domain but would not necessarily address anatomical preservation.

In conclusion, while the mentioned metrics offer valuable insights for image generation, alternative methods are needed to comprehensively assess the success of image harmonization techniques, particularly regarding anatomical fidelity.

Lastly, LPIPS measures the distance between feature representations extracted from pre-trained deep neural networks, like VGG or ResNet. These feature representations capture high-level perceptual qualities of the images, such as textures, shapes, and structures. Unlike metrics like IS or FID, LPIPS considers these learned features, potentially making it a valuable tool for image harmonization. While LPIPS was not specifically designed for harmonization, it offers a unique advantage: assessing the perceptual quality and fidelity of harmonized images. Specifically, it reports on the similarity between the harmonized images and the images from the target domain in terms of human perception. Further research is needed to determine its full effectiveness for harmonization evaluation in various contexts.

These metrics offer insights into the quality and diversity of harmonized images. Clinically, they can be useful in scenarios where harmonization involves generating synthetic images to match a target domain. For instance, when harmonizing images to a specific scanner's characteristics, these metrics help ensure that the generated images align well with the clinical standards of the target domain, thus facilitating consistency in diagnostic practices. IS measures the quality and diversity of generated images, while FID and KID compare the distribution of real and generated images in a feature space. LPIPS evaluates perceptual similarity based on high-level features. When used in harmonization tasks, these metrics can help assess how well the harmonized images match the target domain characteristics, particularly in terms of perceptual and feature-level fidelity.

In addition to the mentioned metrics, it is crucial to evaluate not only the technical aspects of the harmonization process but also its practical utility in real-world applications. It is important to consider how well the harmonized MRI data perform in tasks such as disease classification, age estimation, and ROI segmentation. For instance, ImUnity model [54] assessed the classification ability to identify individuals with ASD (Autism Spectrum Disorder) within the ABIDE database, both before and after the harmonization process. In [55] the improvement of Alzheimer's disease classification was reported after applying a harmonization strategy. In another study [56], the segmentation of the thalamus from various MR image modalities was performed, and the impact of harmonization on the segmentation algorithm was investigated.

Common metrics used to evaluate MRI harmonization methods for downstream tasks like segmentation and classification include Dice Similarity Coefficient, Jaccard Index,

Accuracy, Precision, and Recall. Other metrics such as Hausdorff Distance, F1-score, AUC-ROC, and Sensitivity are also employed [57]. These metrics offer quantitative measures of the performance of harmonization methods in tasks like anatomical segmentation and disease classification.

Deep learning-based harmonization taxonomy

Designing a successful harmonization network hinges on four critical elements:

1. Data Availability: The type of data available for training, whether paired (corresponding images from source and target domains) or unpaired (images from each domain without direct matches), significantly impacts the design choices
2. Loss Functions: These functions mathematically quantify the errors made by the network during training. The specific loss function chosen guides the network toward achieving the desired harmonization goals.
3. Backbone Architecture: The underlying architecture of the neural network serves as the foundation for learning image representations. Different architectures offer varying capabilities for feature extraction and image transformation.
4. Learning Procedure: The optimization algorithm used to train the network plays a crucial role in its effectiveness. This includes techniques for adjusting network weights and parameters to minimize errors.

The following sections delve into a systematic categorization of deep learning-based harmonization methods. This categorization is based on these four key aspects:

- Network Architecture: The underlying structure of the neural network.
- Network Learning Algorithm: The specific optimization technique used for training.
- Network Supervision Strategy: The approach used to guide the training process of the neural network.
- Network Output: The form of the output generated by the network (e.g., harmonized image, segmentation map).

Figure 4 provides a visual representation of this proposed classification scheme, highlighting the different aspects considered.

Network architecture

The choice of network architecture is critical for MRI harmonization because it dictates the model's ability to learn and represent the complex relationships between images acquired from different sources or scanners. Different architectures offer varying degrees of complexity and flexibility, which ultimately influence their performance in harmonizing MRI data. Figure 5 provides a historical timeline illustrating the evolution of deep learning networks for MRI harmonization in recent years.

To better understand these methods, we can categorize them based on the underlying network architecture they employ. Some of the commonly used architectures (Fig. 6) in MRI harmonization include U-Net, GANs, VAEs, Flow-based Generative Models,

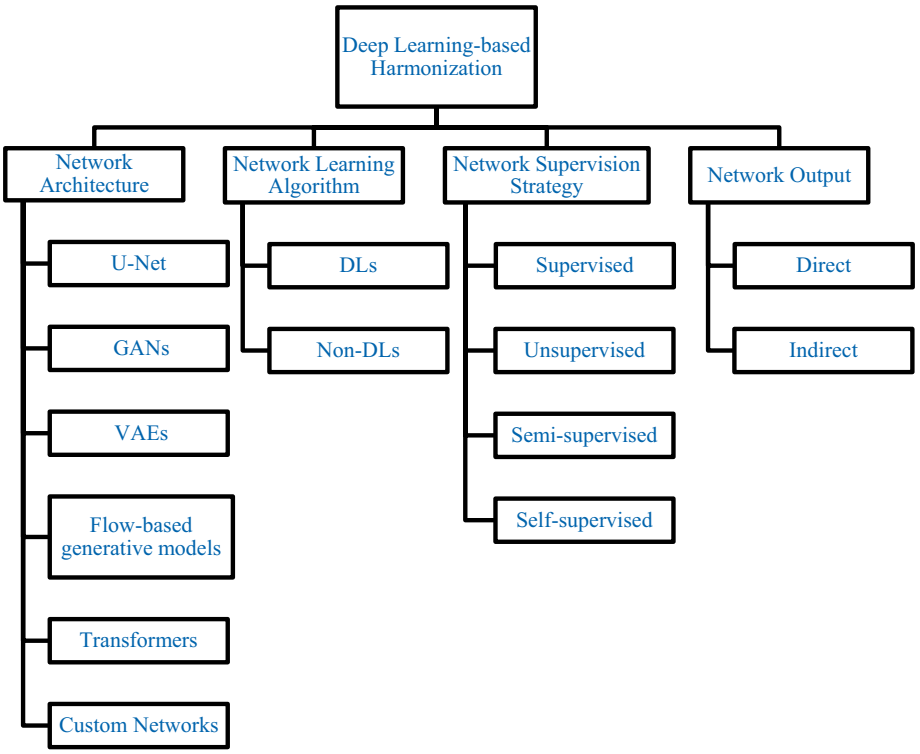


Fig. 4 Taxonomy of deep learning-based MRI harmonization approaches

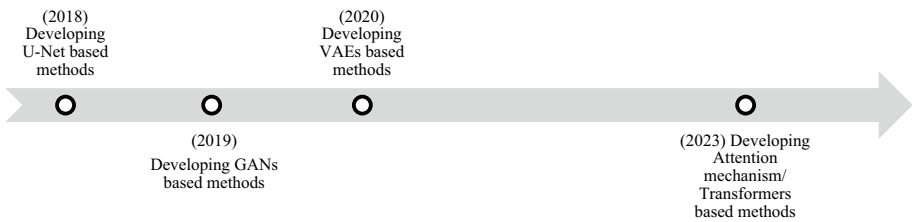


Fig. 5 Timeline of deep harmonization methods

Transformers, and Custom Networks (approaches are not exclusively in one category and are a combination of several networks).

U-Net

The U-Net architecture has been shown to be successful in segmenting [58] and synthesizing medical images [59]. The U-Net architecture has provided good results in dealing with large and diverse datasets in medical imaging. Due to the skip connections, it effectively retains the finer details from the initial images and has demonstrated strong performance in image-to-image translation [60] (Fig. 6A). One notable advantage of employing U-Net is its ability to enhance data with elastic deformation. It also can extract a large number of feature channels in upsampling. However, an inherent limitation is its comprehensive downsampling, which may result in the loss of spatial information [61].

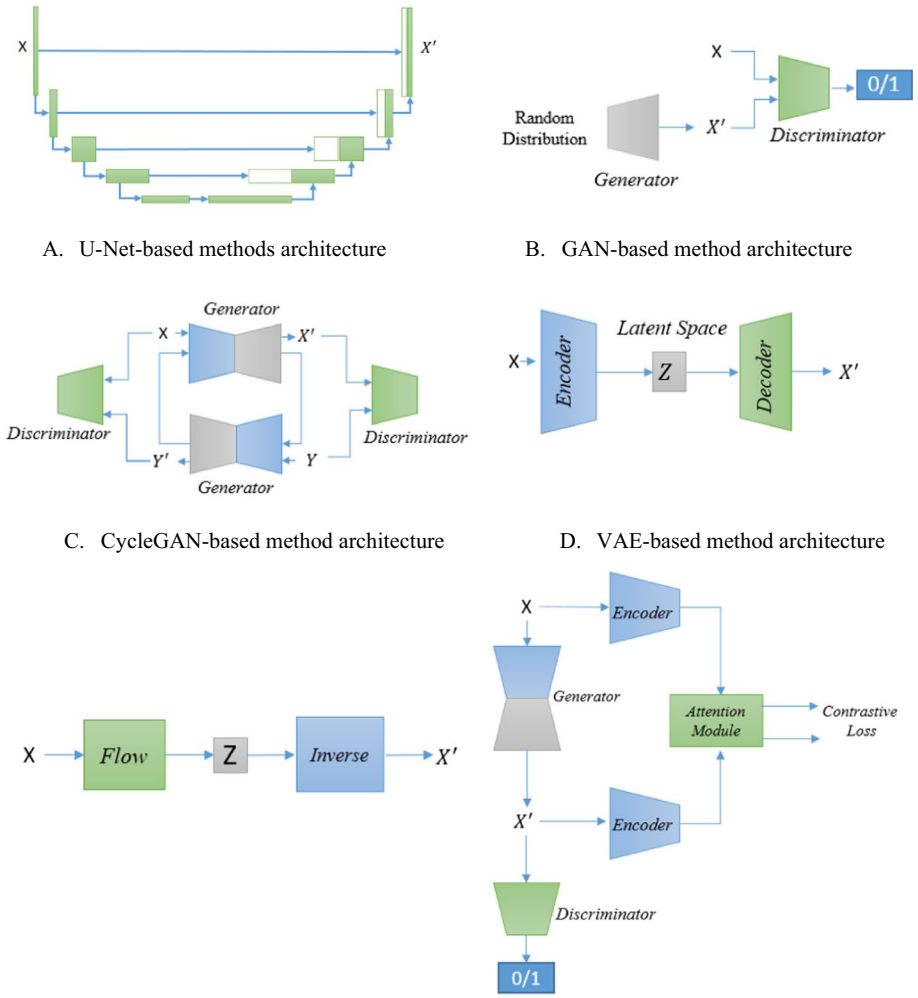


Fig. 6 General diagram for deep learning-based medical image harmonization. **A** U-Net-based methods architecture. **B** GAN-based method architecture. **C** CycleGAN-based method architecture. **D** VAE-based method architecture. E. Flow-based generative model. **F** Transformers-based architecture

On the basis of U-Net architecture, a supervised contrast harmonization has been introduced which is called DeepHarmony [62]. An overlap cohort was provided through two different protocols in order to get training data. This technique transforms images from protocols to generate harmonized ones that imitate the contrast of the target protocol. DeepHarmony is trained in two ways: one-to-one and many-to-one. Both of them require four separate networks. The former uses a single contrast from the source protocol and produces an image with the corresponding contrast from the target protocol. The latter employs four of the input contrasts including T1-weighted, FLAIR, PD-weighted, and T2-weighted from the source protocol to generate an output contrast from the target protocol. According to the result, this approach improves the result compared with one-to-one, but it needs further parameters. This approach needs training data of paired traveling subjects. In large-scale studies, it is hard to acquire.

Bottani et al. [63] utilized three architectures based on 3D U-Net to synthesize T1w non-contrast enhancement (T1w-nce) from T1w contrast enhancement (T1w-ce). These modifications included a version with added residual connections referred to as Res-U-Net, a version with incorporated attention mechanisms called Att-U-Net, and a version incorporating both transformer and convolutional layers known as Trans-U-Net. These models were employed as standalone generators and also incorporated into a conditional GAN setup, along with the addition of a patch-based discriminator. Although the models offered a degree of interpretability and provided promising results in brain image segmentation, there is a limitation to creating paired T1w-nce and T1w-ce due to the time and cost constraints. In another study [64], a U-Net model was developed to learn the non-linear transformation from the contrast of a source image to that of a target image across three MRI contrasts. The training and validation have been accomplished using 2D paired MR images.

The U-Net architecture synthesizes images at the pixel level using paired data, necessitating precise image coregistration for effective model training. Consequently, inadequate alignment of paired MR images could result in the loss of certain brain structure information in the generated images.

Generative adversarial network (GAN)

Generative adversarial networks are an approach to generative modeling using deep learning methods. Such architecture can be considered as image-to-image translation, which generates an image in an unsupervised learning task. GANs have attracted enormous interest in image translation, typically to generate new images from existing ones. The incorporation of a generator and a discriminator creates a GAN model (Fig. 6B).

A CycleGAN [65], which is widely used in MRI harmonization, stands as a robust deep learning framework facilitating image-to-image translation without the necessity of paired training data. It is made up of two GANs, including two discriminators and two generators (Fig. 6C). The objective of the model is to comprehend the attributes of the target domain and produce novel images from the source domain that exhibit these attributes. CycleGAN offers several key advantages over other image-to-image translation models. It excels in terms of accuracy by utilizing unpaired data, thus delivering superior results without the need for an extensive collection of paired training images. Its robustness to domain shifts in the data makes it versatile, allowing it to perform well with input images from various domains, enabling a wide range of translation tasks. Moreover, CycleGAN can generate high-quality images with smaller datasets, making it particularly valuable for tasks with limited training data, such as medical image-to-image harmonization [66, 67].

In [68], an unsupervised image-to-image canonical mapping based on CycleGAN was learned from a diverse dataset to a reference domain. This approach was evaluated on brain age prediction and schizophrenia classification to show how can mitigate confounding data variation while retaining semantic information. A Maximum Classifier Discrepancy Generative Adversarial Network (MCD-GAN) was introduced [69], leveraging the benefits of both generative models and maximum discrepancy theory. Komandur et al. [66] employed a 3D CycleGAN to harmonize brain MRI data from diverse

sources. They concluded that GAN-harmonized data yield higher accuracy compared to raw data for the age prediction task.

Training the CycleGAN can be time consuming, especially when dealing with a large number of training images. Overfitting is a concern with CycleGAN, possibly resulting in inferior outcomes on unseen data. Interpretability can be a hurdle with CycleGAN, making it challenging to comprehend the rationale behind the model's generated results. To address infant neuroimaging datasets harmonization, S2SGAN (Surface-to-Surface GAN) was introduced [70]. This method combines the spherical U-Net and the CycleGAN. The presented cycle-consistent adversarial networks are based on a spherical cortical surface for harmonizing cortical thickness maps between different scanners.

The majority of unsupervised approaches are unable to differentiate between variability caused by image acquisition and that originating from population differences across different sites. As a result, these methods necessitate datasets to include subjects or patient groups with comparable clinical or demographic profiles. Deep learning frameworks have shown success in dealing with image translation by breaking it down into basic content (e.g., line contours and orientation) and complex style (primarily color and texture) [71]. This framework is known by the "Disentangled Representation" term (more details are provided in part 3–2). All images with the same domain have the same content space but the style can vary. This allows the framework to make changes to the style while maintaining the original content. The aim is to maintain consistency in content but adjust the image style. This approach has been used to produce highly realistic translation results.

In [72], cross-site MRI image harmonization has been considered a style transfer problem instead of a domain transfer problem to overcome the need for the datasets to include patient groups or subjects with homogeneous demographic information. The proposed approach tries to overcome the limitation of the statistical method [73]. Some statistical methods need certain clinical or demographic characteristics of subjects within the dataset to control the acquisition-based variance. This method is capable of harmonizing MRI images without prior knowledge of their scan/site labels and harmonized by infusing style information derived from a single reference image.

The StarGAN is a version of GAN that enables Image-to-Image GANs to perform mapping across more than two domains using a single generator and acquire shared features applicable to all domains [74], whereas conventional models need multiple generators. However, it has limited ability to capture minor feature variations. The StarGAN v2 was utilized to process various datasets using canonical mapping from different sites to a reference domain [75]. By doing so, they reduced the impact of site-based variance while preserving the meaning provided by the input data.

A 3D model named Image Generation with Unified Adversarial Networks (IGUANe) was introduced [76] that benefits domain translation and style transfer methods to harmonize multicenter brain MR images. It extends the CycleGAN architecture by integrating multiple domains for training using a many-to-one approach.

Addressing hallucinations in GANs

Hallucinations refer to the generation of artificial structures in the output images that do not correspond to real anatomical features. This is particularly problematic in medical

applications, where the authenticity of every detail is crucial for accurate diagnosis and treatment planning.

The causes of hallucinations in GANs are multifaceted [77]. One primary cause is the imbalance between the generator and the discriminator during training, where the generator may learn to produce plausible but incorrect details to fool the discriminator. Another cause is the lack of sufficient and diverse training data, which can lead to overfitting and the generation of unrealistic artifacts. Additionally, the inherent randomness in GANs can introduce noise that manifests as hallucinations.

To mitigate hallucinations, several strategies have been suggested in the literature. One approach is to enhance the quality and quantity of training data [78]. Techniques such as data augmentation can also help in this regard. Another method is to improve the training process through techniques like progressive growing of GANs [79], where the model is trained on low-resolution images initially and progressively moves to higher resolutions, allowing for more stable training and better-quality outputs.

Regularization techniques, such as spectral normalization and gradient penalty [80], can also help by stabilizing the training dynamics and reducing the likelihood of hallucinations. Additionally, incorporating domain-specific knowledge through the use of hybrid models that combine GANs with traditional image processing techniques or other deep learning models can provide more reliable outputs. By integrating these strategies, future work can aim to reduce the incidence of hallucinations in GAN-generated images, thereby enhancing their clinical applicability and reliability.

Variational autoencoders (VAEs)

The GAN networks and VAE are two of the most popular AI image generators. The VAEs consist of two main architectures: encoders and decoders (Fig. 6D). The encoder learns and encodes the representation of input data and maps it to the latent space. The decoder converts the latent space to get back the original data [81]. According to the comparative study regarding anomaly segmentation on brain MRI images, GAN-based models are recognized for their ability to generate ultra-realistic and sharp images [82]. Meanwhile, AutoEncoders are known for their propensity to produce blurry reconstructions.

Torbati et al. proposed a multi-scanner harmonization framework [83]. This encoder–decoder architecture maps the MRIs from multi-scanners to the latent space and then maps the latent embedding to the harmonized image space. It considers two training steps to preserving the anatomical structure: (1) the harmonized images and input image should be as similar as possible and the variance of embeddings across the scanners should be minimized; (2) ensuring that the output images remain similar across scanners. This step helps maintain uniformity and consistency between various scans.

One-shot learning learns from limited data and has shown significant results in many tasks of medical imaging [84–86]. Based on VAEs, a one-shot learning method was proposed for harmonization across imaging locations [49]. During testing, the architecture utilizes an image from a clinical site to create an image that aligns with the intensity scale of the cooperating sites. In another study, a zero-shot learning framework using style-blind autoencoders was introduced [87]. The network was trained to recognize

and extract essential content information exclusively. Consequently, the trained network demonstrated the capability for zero-shot harmonization by discarding unknown scanner-dependent contrast information.

The architectures based on the combination of GAN networks and VAEs have been presented in multiple studies. Cackowski et al. introduced ImUnity [54]. To decrease the effect of the scanner or site identity on the training results, the generator (VAE) was equipped with a bias learning module connected to the bottleneck. Additionally, a biological preservation module was proposed to maintain pertinent biological information within the latent space representation.

Flow-based generative model

A flow-based generative model is a type of generative model that transforms a simple input distribution into a more complex data distribution using a series of invertible transformations called flows (Fig. 6E). These models offer exact likelihood evaluation, making them suitable for tasks like density estimation. They are flexible, scalable, and can handle high-dimensional data, making them applicable to various tasks such as image generation and denoising.

Recently, BlindHarmony [88] was introduced as a solution for blind harmonization, where a flow-based blind MR image harmonization framework was developed. BlindHarmony utilized only the target domain dataset during training. The objective is to discover a harmonized image that retains the anatomical structure and contrast of the input source domain image while ensuring a high likelihood in the flow model, thus facilitating harmonization for the target domain by leveraging the invertibility of flow models. Beizaee et al. proposed an unsupervised MR harmonization method based on normalizing flow [89]. Within this framework, a shallow harmonizer network was trained to restore images of the source domain from their augmented counterparts. Subsequently, a normalizing flow network was trained to understand the distribution of the source domain. Ultimately, during testing, modifications were made to the harmonizer network so that the resulting images aligned with the distribution learned by the normalizing flow model of the source domain. In another study, a causal flow-based approach was proposed to address the issue of varying feature distributions in multi-site data utilized for Parkinson's disease classification [90].

Flow-based models are inherently invertible, allowing for bidirectional mapping between domains without loss of information and the transformation process is interpretable so facilitates a better understanding of the harmonization process. However, compared to other generative models, flow-based methods are relatively newer in the field of MRI harmonization, leading to fewer established techniques and benchmarks.

Transformers

Recently, transformers with attention mechanisms (Fig. 6F) have gained promising performance in medical image processing [91] and image-to-image translation [92]. Yao et al. [93] employed two attention-based image-to-image translation frameworks, Morph-UGATIT and QS-Attn [94] for MRI harmonization. The effectiveness of these harmonization strategies was evaluated and compared to the conventional CycleGAN by performing a subcortical segmentation task on a heterogeneous dataset acquired at

1.5T and 3T. Among the frameworks assessed, QS-Attn stands out with the most optimal performance. Morph-UGATIT shows comparable performance to QS-Attn and exhibits enhancements in most subcortical regions compared to the CycleGAN model. They concluded that attention-based harmonization techniques demonstrate notable improvements over the baseline frameworks, especially when combined with diverse downstream tasks like segmentation. In [95] two transformer encoders were introduced to extract both style and content information from MR images, and two decoders were utilized to generate harmonized image patches. Additionally, the impact of changes in image resolution on position encoding was addressed. To capture semantic information in images of varying scales, a content-aware positional encoding scheme method was employed, effectively accommodating images of different sizes.

Custom networks

The custom-designed networks are characterized by architectures that, while sharing similarities with U-Net, GANs, VAEs, flow-based generative models, and transformer-based approaches, include distinct components and configurations tailored to address specific challenges in this research domain. These architectures utilize elements that diverge from typical implementations of the mentioned architectures. They incorporate unique configurations of convolutional layers, pooling layers, domain classifiers, and specially designed blocks that may not fit into these established categories.

Inspired by the adversarial framework and domain adaptation techniques, a harmonization approach was introduced that can be effective for classification, regression, and segmentation tasks while employing two diverse network architectures [96]. Image harmonization can be considered a multi-source joint domain adaptation problem. This approach tries to produce shared feature representations that are invariant to the acquisition scanner while still completing the main task of interest across scanners and acquisition protocols with minimum performance compromise.

An attention-guided domain adaptation was introduced for multi-site MRI harmonization and was applied to automated brain disorder identification [97]. In this framework, the attention discovery and domain transfer modules were defined to automatically pinpoint discriminative dementia-related regions in each whole-brain MRI scan and facilitate knowledge transfer between the source and target domains, respectively. Wolleb et al. [98] introduced a constraint in the latent space of an encoder–classifier network to ignore scanner-related characteristics.

Network learning algorithm

The learning strategies for MRI harmonization has been categorized into two main groups: disentangled learning (DL) methods and non-disentangled learning methods (non-DLs).

1. Disentangled Learning (DL) Methods: Disentangled learning methods refer to approaches where the neural network or algorithm is explicitly designed to learn

separate and interpretable factors or features from the input data. In the context of MRI harmonization,

- o DL methods aim to disentangle latent factors such as imaging artifacts, variations in acquisition protocols, tissue types, and other confounding factors that contribute to variability in MRI scans.
- o These methods typically employ architectures such as VAEs, adversarial training techniques, or other models with explicit mechanisms to learn invariant representations across different datasets.
- o The goal of DL methods is to improve the robustness and generalization of MRI harmonization by separating out and modeling the underlying factors of variability.

2. Non-Disentangled Learning Methods (non-DLs): Non-disentangled learning methods, in contrast, do not prioritize the disentanglement of underlying factors in the input data:

- o These methods may include traditional neural networks, regression-based models, or simpler machine learning algorithms.
- o They focus on direct mapping from input (MRI scans with variability) to output (harmonized MRI scans) without explicitly modeling or separating out the distinct factors contributing to variability.
- o While effective in certain scenarios, non-DL methods may be less robust to dataset variations and might not generalize as well across different MRI datasets with varying acquisition conditions.

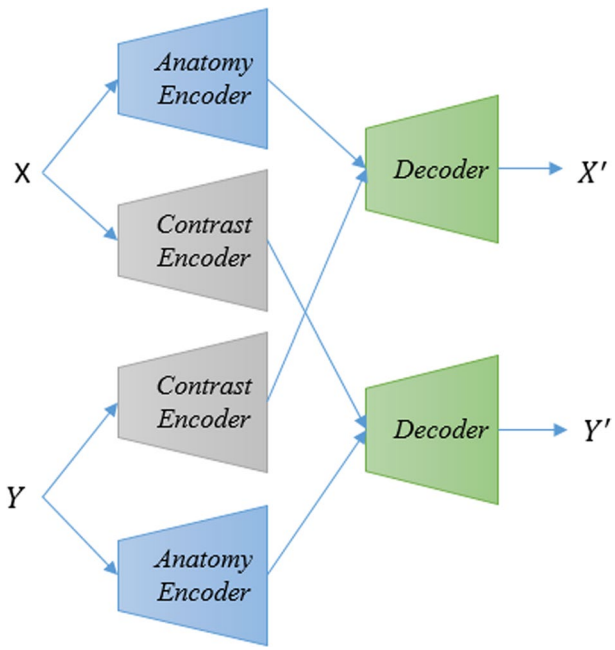


Fig. 7 Disentanglement representation learning diagram

Studies have shown that inverting the magnetic resonance imaging signal equation produces an encoded image with disentangled contrast data and contrast-invariant anatomical data. This means the image can be separated into two distinct parts that are denoted by β and θ for anatomical map and contrast, respectively. Figure 7 demonstrates the disentanglement representation learning diagram which uses encoder–decoder architecture.

Dewey et al. proposed an approach that includes two sub-networks [99]. The first one is the encoder and the second one is the decoder. These two sub-networks are connected by the latent space which contains disentangled contrast data and contrast-invariant anatomical data. This encoder and decoder are based on U-Net architecture. For training the network, the T1-weighted and T2-weighted images of the same anatomy using different scanning protocols. It is expected that the network will produce the same β value for the same anatomies and equal θ for images generated from scanners with the same protocol. This approach utilizes the multiple contrast magnetic resonance (MR) images acquired within each site. These intra-site paired data can be found in the same session. But, relying on this technique alone will not provide a globally disentangled latent space. Zuo et al. introduced a similar architecture which is called CALAMITI [100] based on information bottleneck theory. The algorithm learns a global latent space of anatomical and contrast information and it can be adapted to a new testing site using only the data collected at the new site. This architecture requires paired images from the same site during training which might limit certain applications, especially in cases where obtaining multi-contrast images is not feasible. Zuo et al. introduced a method that uses a single MR modality [101]. The inputs of the encoder–decoder are two slices from different orientations of the same 3D volume instead of paired images. Additionally, they defined a new information-based metric for evaluating disentanglement.

According to the inspiration of advancements in multi-domain image translation, Multiple-site Unsupervised Representation Disentanglement (MURD) was introduced [102]. The harmonized images were produced using combining the content of the original image with styles from a specific site or a generator. The style generator enables the generating of multiple appearances concerning natural style variations associated with each site. In [103], another work based on disentangled representation was introduced that disengaged the image into content and scanner-specific space. This method was evaluated on healthy controls and multiple sclerosis (MS) cohorts. Zhao et al. developed a deep learning model to harmonize the multi-site cortical data using a surface-based autoencoder [104]. The encoded cortical features were subsequently decomposed into components related to site-specific characteristics and those unrelated to site effects. An adversarial strategy was employed to promote the disentanglement of these components. Subsequently, the decoding of the site-unrelated features, combined with other site-related features, facilitates the generation of mappings across different sites.

A disentangled latent energy-based style translation (DLEST) framework was introduced [105] in order to harmonize image-level structural MRI. The proposed model disentangles site-specific style translation and site-invariant image generation through the utilization of an energy-based model and a latent autoencoder.

Harmonization with Attention-based Contrast, Anatomy, and Artifact Awareness (HACA3) was suggested to overcome some drawbacks of synthetic-based disentanglement [106]. The previous methods to harmonize MR images are limited by their reliance on assumptions that contrast images from the same subject share the same anatomy. These assumptions are doubtful as different contrasts are aimed at highlighting distinct anatomical features. Moreover, these methods require a fixed set of images for training, which is often limited to T1-weighted and T2-weighted data. Finally, the existing methods are sensitive to artifact images and other image artifacts, making them less useful in practical applications.

Utilizing DRL's advantages, it becomes possible to learn and align independent factors within generation objectives with the latent representation through disentanglement. Consequently, this enables effective control over the generation process [107].

In contrast, non-disentangled representations encapsulate multiple factors of variation in a more intertwined manner, making it challenging to isolate individual factors and understand their influence on the harmonization process. While non-disentangled approaches may offer simplicity and computational efficiency, they often lack the interpretability and robustness necessary for reliable MRI harmonization across diverse datasets.

Network supervision strategy

Broadly, harmonization techniques can be classified into major categories, including supervised, unsupervised, semi-supervised, and self-supervised approaches. Supervised harmonization methods [62, 108] are employed to harmonize images from different scanners/sites using a cross-domain dataset. These methods require a group of subjects to be scanned in both domains. This arrangement provides the training and validation data that the model requires. Due to the logistics and costs of acquiring data, gathering cross-domain data is uncommon in practice. Additionally, cross-domain data are limited. Typically, data from multiple domains are available without cross-domain data. This necessitates an unsupervised harmonization method [66, 89, 109], which requires a training method without data from the same subjects from multiple domains.

Semi-supervised approaches [110], on the other hand, are trained using a dataset containing under-sampled acquisitions of both source and target contrasts from MRI scans. In contrast, self-supervised approaches in the realm of harmonization enable models to learn from the inherent structure of the data itself, eliminating the need for external labels. However, it is important to note that many methods introduced in the field of harmonization are categorized under unsupervised techniques.

Network output

In MRI harmonization, methods can be categorized into two categories based on their network output: direct and indirect. In the direct network output category, methods focus on predicting the target image directly from the reference image. Deep learning models within this category are specifically trained for harmonizing data, allowing for straightforward evaluation by a radiologist. Conversely, methods in the indirect network output category involve training models on a downstream task, such as classification, registration, segmentation, or age prediction. In this category,

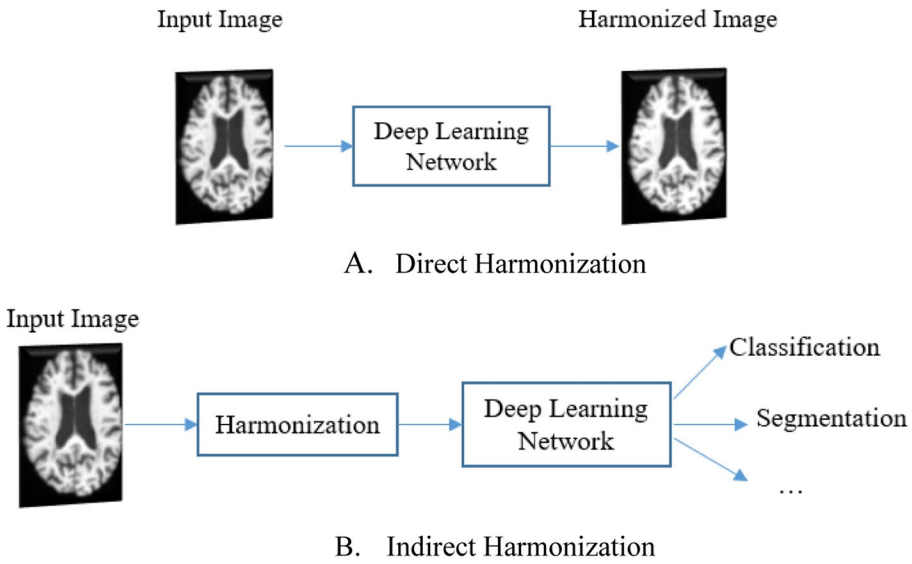


Fig. 8 Classifying Harmonization according to Network Output

the harmonization process occurs implicitly through optimization during training, resulting in harmonized data that remain concealed from direct observation. Figure 8 illustrates the general diagram of direct and indirect harmonization.

There are some papers in the literature that consider harmonization for downstream tasks. Grigorescu et al. [111] explored two unsupervised domain adaptation techniques, seeking the optimal solution for tissue segmentation maps using T2-weighted magnetic resonance imaging data from an unseen neonatal population born preterm. In [56], a 3D U-Net architecture was presented to segment the thalamus from multiple MR image modalities, and the effect of harmonization on the segmentation algorithm was investigated. Tor-Diez et al. [112] used an unpaired image-to-image translation strategy based on adversarial networks and supervised segmentation for the anterior visual pathway. They concluded that harmonization can improve the segmentation results significantly. In another study, with the aim of boosting Alzheimer's disease classification, the Attention-Guided Generative Adversarial Network (AG-GAN) was used for data harmonization [55]. Komandur et al. [66] proposed a CycleGAN-based harmonization for improving the results of age estimation. According to the assumption that neglecting downstream applications during harmonization can hinder overall performance, the goal-specific harmonization framework was proposed [113]. This VAE-based architecture utilizes downstream application performance to regulate the harmonization procedure. They concluded that while this approach enhances downstream performance, it may also limit generalization to new downstream applications, potentially necessitating repetition of the training procedure for each one.

In indirect approaches, according to downstream tasks, different objective functions for clustering, classification, or regression can be defined, so they need a diverse range of learning procedures, parameters, and optimization algorithms.

Table 3 Details of harmonization techniques that were discussed in this review article. DRL indicates Disentanglement Representation Learning

Authors	Year	Network	Data	Number of patients	2D/3D	Open-Source	Image modality	Downstream Task
Dewey et al. [62]	2019	U-Net	local dataset	12	2D & 3D	–	T1-weighted FLAIR PD-/ T2-weighted	No
Zhao et al. [70]	2019	U-Net and CycleGAN	BCP dataset	183	3D	–	T1-weighted T2-weighted	Yes
Bashyam et al. [68]	2020	CycleGAN	local dataset	9701	2D	–	T1-weighted	Yes
Tor-Diez et al. [112]	2020	VAE and GAN	Clinical Center (CNH, CHOP, CHC)	18	3D	–	T1-weighted	Yes
Dewey et al. [99]	2020	Custom Network with DRL	IXI Two private datasets	50	2D	–	T1-weighted T2-weighted	No
Mengting Liu et al. [72]	2021	CycleGAN	UKBB, PPMI, ADNI, ICBM	518	2D	–	T1-weighted	No
Eshaghzadeh et al. [83]	2021	VAE	local dataset	18	2D	Code	T1-weighted	No
Dinsdale et al. [96]	2021	Custom Network	UK Biobank, OASIS, Whitehall II	8418	2D	Code	T1-weighted	Yes
Guan et al. [97]	2021	Custom Network	ADNI	2572	3D	–	T1-weighted	Yes
Zuo et al. [100]	2021	VAE	OASIS3 IXI	120	2D	Code	T1-weighted T2-weighted	No
Li et al. [103]	2021	Custom Network with DRL	local dataset	150	2D	Code	FLAIR	Yes
Grigorescu et al. [111]	2021	U-Net	dHCP ePrime	403	3D	Code	T_2-weighted	Yes
Sinha et al. [55]	2021	GAN	ADNI AIBL OASIS	854	3D	–	T1-weighted	Yes
Bottani et al. [63]	2022	U-Net	local dataset	307	3D	–	T1-weighted	No
Osman et al. [64]	2022	U-Net	BRATS'2018	477	3D	code	T1-weighted T2-weighted FLAIR	No
Yan et al. [69]	2022	GAN	Simulated data (Double moon) ABCD	8087	3D	–	T1-weighted	No
Fatania et al. [87]	2022	VAE	CC359 public dataset	250	2D	–	T1-weighted	No
Wolleb et al. [98]	2022	Custom Network	local dataset ADNI	Not provided	3D	Code	T1-weighted	Yes
Bashyam et al. [75]	2022	StarGAN	local dataset	8876	2D	–	T1-weighted	Yes
Zuo et al. [101]	2022	Custom Network with DRL	IXI	150	2D	–	T1-weighted or T2-weighted	No
Chang et al. [109]	2022	Custom Network	local dataset	116	2D	–	T2-weighted	Yes
Yurt et al. [110]	2022	GAN	IXI Vivo Brain Dataset	104	3D	Code	T1- / T2-weighted T2- / PD-weighted	No
Shao et al. [56]	2022	U-Net	local dataset	22	3D	–	T1-weighted	Yes
An et al. [113]	2022	VAE	ADNI AIBL MACC	2787	2D	Code	T1-weighted	Yes

Table 3 (continued)

Authors	Year	Network	Data	Number of patients	2D/3D	Open-Source	Image modality	Downstream Task
Komandur et al. [66]	2023	CycleGAN	UK Biobank ADNI AIBL OASIS-1 WHIMS	4941	3D	–	T1-weighted	No
Liu et al. [73]	2023	GAN	ADNI3 ICBM PPMI UK Biobank ABCD	718	2D	Code	T1-weighted	Yes
Jeong et al. [88]	2023	Flow-based generative model	OASIS3	20	2D	Code	T1-weighted	Yes
Beizaee et al. [89]	2023	Flow-based generative model	ABIDE	79	2D	Code	T1-weighted	Yes
Yao et al. [84]	2023	Transformers	SegData HarmoData	Not provided	3D	–	T1-weighted	Yes
Han et al. [95]	2023	Transformers	ADNI	391	2D	–	T1-weighted	Yes
Parida et al. [49]	2023	VAE	local dataset	180	3D	–	T1-weighted	No
Cackowski et al. [54]	2023	GAN-VAE	ABIDE OASIS SRPBS	1698	3D	–	T1-weighted	No
Zhao et al. [104]	2023	Custom Network with DRL	public dataset & local dataset	2342	3D	–	T1-weighted T2-weighted	Yes
Wu et al. [105]	2023	Custom Network with DRL	OpenBHB SRPBS	4092	2D	–	T1-weighted	Yes
Zuo et al. [106]	2023	Custom Network with DRL	public dataset & local dataset	210	3D	Code	T1-weighted T2-weighted FLAIR PD-weighted	Yes
Roca et al. [76]	2024	CycleGAN	ADNI MCIC PPMI COBRE ABIDE	676	3D	Code	T1-weighted	Yes
Vigneshwaran et al. [90]	2024	Flow-based generative model	6 sites	415	2D	–	T1-weighted	Yes
Liu et al. [102]	2024	Custom Networks with DRL	ABCD ADNI AIBL	> 6000	2.5D	Code	T1-weighted T2-weighted	Yes

Therefore, when addressing a different task, the harmonization technique needs to be initiated anew.

The details regarding the papers' information are presented in Table 3. Since the datasets and evaluation metrics are different the comprehensive comparison is limited. The variation in scanners and number of participants, healthy and patient cases, and the investigated contrasts can affect the result and applicability.

Applicability and limitations of harmonization

Harmonization techniques in MRI neuroimaging are essential for mitigating scanner-related variability and site effects, allowing for more reliable and comparable results across different studies and cohorts. These methods enhance the statistical power of studies by increasing the effective sample size and facilitating meta-analyses and multi-site collaborations. Harmonization is particularly beneficial in large-scale studies involving data collected from different scanners, protocols, and populations.

Despite its advantages, harmonization is not without limitations. One major concern is that harmonization algorithms may introduce bias if not properly validated across different datasets. The effectiveness of harmonization can vary depending on the specific characteristics of the datasets, including differences in scanner types, imaging protocols, and the populations being studied. Additionally, harmonization processes might inadvertently remove or obscure biologically relevant variations that are not related to scanner differences.

Another limitation is the complexity and computational cost associated with advanced harmonization techniques. Implementing these methods often requires significant expertise and resources, which may not be readily available in all research settings. Furthermore, the choice of harmonization method can impact the results, necessitating careful consideration and validation of the chosen approach.

Neuroimaging analysis should not always use harmonization, especially in scenarios where the primary goal is to investigate scanner-specific effects or when studying the inherent variability between different imaging systems. In such cases, harmonization could mask the very differences that are of interest. Additionally, in single-site studies with consistent imaging protocols, the need for harmonization may be minimal or unnecessary.

Harmonization might negatively impact results if applied inappropriately. For instance, over-harmonization can lead to the loss of important biological signals, resulting in reduced sensitivity to detect true effects. It is crucial to balance the removal of unwanted scanner-related variance with the preservation of genuine biological variability. Researchers should perform extensive validation to ensure that harmonization does not distort the data in a way that affects the study outcomes.

Discussion

Inconsistent contrast across MRI scans presents a significant hurdle for modern medical image analysis techniques. This becomes particularly evident when using deep learning models trained on specific image types. For example, a segmentation model designed for CT scans might perform poorly on MR images due to the fundamental differences in how these imaging techniques capture the body. While research efforts like cross-domain synthesis [114–117] have aimed to address these challenges, inconsistencies in contrast remain a persistent issue, even within the realm of MRI scans themselves [28].

It is important to distinguish between image harmonization and cross-domain synthesis, although both techniques address challenges with image variability. While image harmonization aims to align images from different sources (e.g., scanners) while preserving anatomical details and spatial relationships, cross-domain synthesis focuses on

generating images in a target domain based on images from a different source domain. While the former is particularly useful when combining datasets from different sources for analysis and it helps ensure consistency and reduces variability, enabling more reliable comparisons, the goal of the latter is to preserve key features like structures or textures, while also creating visually realistic images in the target domain. This technique can be used for data augmentation, domain adaptation, and image enhancement. In essence, harmonization aims to make existing images from different sources more compatible, while cross-domain synthesis aims to create entirely new images within a specific domain. Given this distinction, and our focus on harmonization techniques, we have not explored cross-domain synthesis techniques within this paper.

While most harmonization techniques reviewed here leverage 2D images, there is growing recognition that 3D models offer significant advantages. 3D models hold greater potential for capturing the full complexity of medical image features, potentially leading to improved learning performance. However, 3D models come with significant computational limitations such as increased memory requirements to store and process 3D data and longer training times for deep learning models due to the larger amount of data. To address these challenges, some harmonization approaches employ a hybrid strategy for example, 2D Slices with Multi-Orientation. In this approach, models are trained using 2D axial, coronal, and sagittal slices extracted from each 3D MR volume. Subsequently, these multi-directional 2D slices are then combined into a harmonized 3D volume [106].

A majority of current harmonization methods reported in the literature evaluate their performance primarily on T1-weighted MRI scans. While this is a common starting point, it is important to acknowledge the limitation of limited generalizability, i.e., these methods might not achieve the same level of success with other MRI contrasts, such as T2-weighted, PD-weighted, and T2-FLAIR images.

The preprocessing steps before harmonization approach can affect the harmonization outcome. In many cases, the harmonizing native MRI is an essential step. Subsequent investigation should clarify the influence of harmonization on native MRI and explore how the quality of harmonization can be conditional on the preprocessing procedure employed [73]. By understanding this interplay between preprocessing and harmonization, researchers can develop more robust and effective pipelines for MRI data analysis.

Evaluating and comparing different harmonization techniques presents several obstacles. Firstly, there is a current lack of standardized and comprehensive datasets encompassing a wide variety of MRI contrasts, scanner types, and patient demographics (healthy vs. patient groups). Many studies rely on subsets of data, focusing on scanners with minimal differences or specific patient groups. Secondly, given the importance of harmonization in medical image processing, establishing a reference dataset would be highly beneficial. This dataset should ideally include multiple MRI contrasts, involve data from various patient cohorts, and encompass a diversity of challenges commonly encountered in real-world scenarios. Such a benchmark would facilitate more comprehensive evaluation and comparison of harmonization techniques for downstream tasks. Thirdly, conventional image similarity metrics might not fully capture the effectiveness of harmonization. They may prioritize overall similarity without adequately considering factors like cross-domain consistency (compatibility between data from different sources) and preservation of crucial anatomical

details. Re-evaluating harmonization success requires metrics that comprehensively assess these key aspects.

While supervised learning approaches using U-Net convolutional neural networks have shown promise in MRI harmonization, they have limitations. Firstly, supervised methods require paired data, meaning the same patients need to be scanned on multiple scanners. This can be expensive and time consuming to acquire, limiting the applicability of these approaches. Secondly, supervised methods often work best for brain imaging due to the relative homogeneity of the MR signal and the feasibility of performing rigid image registration (aligning images based on anatomical landmarks). Addressing these challenges will be crucial to advance the development and evaluation of effective harmonization techniques for broader applications in medical image analysis.

Accurately assessing the effectiveness of different image harmonization techniques remains a challenge due to two key limitations. Firstly, there is an absence of standardized benchmark datasets encompassing a wide variety of factors hinders comprehensive evaluation and comparison. The current studies often rely on the following:

- Subsets of data: Some studies use a limited portion of a larger dataset, potentially missing valuable information.
- Homogeneous data: Some studies focus on data acquired from scanners with minimal differences, limiting the generalizability of findings.
- Data from specific patient groups: Some datasets might be restricted to healthy or diseased individuals, neglecting the real-world scenario where datasets may include both.

Given the crucial role of harmonization in medical image processing, a robust reference dataset is urgently needed. This dataset should ideally include multiple MRI contrasts (T1-weighted, T2-weighted, etc.), data from diverse patient cohorts (healthy and diseased) and a variety of challenges commonly encountered in real-world settings (e.g., scanner variations, acquisition protocols). Such a comprehensive benchmark would enable researchers to thoroughly evaluate and compare harmonization techniques, ultimately improving their performance in downstream tasks.

Conventional image similarity metrics (like PSNR or SSIM) primarily focus on overall image similarity. While important, they may not fully capture the cross-domain consistency, i.e., how well does the harmonized image align with data from a different source (e.g., another scanner)? They may not also fully provide anatomical preservation, i.e., does the harmonized image retain the crucial anatomical details present in the original image? To address these limitations, a re-evaluation of success metrics is necessary. New metrics should be developed, or existing ones adapted to comprehensively assess these essential aspects of harmonization.

Addressing these limitations in datasets and evaluation methods represents a crucial step toward the development and implementation of next-generation harmonization techniques for broader use in medical image analysis.

While U-Net convolutional neural networks have shown promise in supervised learning approaches for MRI harmonization, they face some limitations such as paired data

dependency, applicability constraints such as MR signal homogeneity and rigid image registration. These limitations restrict the broader applicability of supervised U-Net-based approaches for harmonization in medical image analysis.

Generative Adversarial Networks (GANs) have been employed to address harmonization by synthesizing images with a specified contrast, where the "content" from the input image is retained while adjusting the contrast to match that of a target scanner. The CycleGAN utilizes unpaired training data and unsupervised learning, showing promise in harmonization tasks and leading to developments in the prediction of brain age and classification. However, an inherent limitation of GANs is their inability to inherently distinguish content from contrast, potentially resulting in alterations to anatomical details to align more with the target scanner dataset, causing "geometry shifts." Preserving patient anatomy is crucial for precise diagnosis and treatment. In the absence of structural uniformity, the generated images might lack clinically significant specifics. Additionally, GANs are well known for producing artificial structures that are not present in the initial training data, a phenomenon commonly referred to as "hallucination."

Variational Autoencoders (VAEs) offer an alternative approach to MRI harmonization that addresses a key limitation of supervised learning: the need for paired data. VAEs can potentially harmonize data across multiple sites without requiring scans from the same subjects at each location. This is achieved by learning a latent representation of the data, which essentially captures the underlying characteristics of the images in a compressed form. The VAE then transforms data from one site into another using this latent space. However, some of the limitations of VAEs blurry reconstructions and challenging latent space interpretability. While VAEs hold promise for multi-site harmonization without paired data, further research is needed to address these limitations and improve the accuracy and detail preservation in the harmonized images.

Disentangled representation learning aims to separate an image's style (contrast) and content (anatomy) into distinct representations. This allows for modifications to the style while preserving the underlying anatomical details. However, in complex MRI data, factors like contrast and anatomical details can be intertwined and challenging to perfectly separate. This can lead to ambiguities in the disentanglement process, resulting in overlapping or mixed representations of the intended factors. Additionally, extending disentangled representation learning to 3D or higher dimensions presents additional challenges. The increased complexity of higher-dimensional spaces makes it more difficult to disentangle the features within them.

Vision Transformers (ViTs) have emerged as a powerful tool in computer vision, demonstrating effectiveness across diverse tasks like segmentation, classification, and image-to-image translation. This versatility stems from their core mechanism, self-attention. Unlike traditional convolutional neural networks, ViTs can directly analyze relationships between any two parts of an image, allowing them to capture long-range dependencies and gain a deeper understanding of the global context. However, ViTs also face some limitations due to need for high-resolution input images for optimal performance and substantial computational memory and processing power for training and inference. These factors can limit the applicability of ViTs in scenarios with limited computational resources or where processing speed is critical.

In reviewing the advancements in deep learning models for MRI harmonization, it is evident that even marginal improvements in image quality metrics can be of significant clinical value. However, these improvements often appear minor when comparing new network architectures. The statistical analysis of these improvements is crucial to determine their true significance. For instance [118], highlights the importance of using rigorous statistical methods such as Analysis of Variance (ANOVA) and Mixed Effects Models (MEM).

Similarly [119], investigated three U-Nets (dense, robust, and anisotropic) for upscaling low-quality MRI images. Despite non-statistically significant differences in basic evaluation metrics, mixed effects statistics illustrated significant differences. This suggests that while the detailed architecture of these U-Nets may not drastically alter the outcomes, the use of robust statistical techniques can reveal critical differences and interactions. These findings underscore the importance of employing comprehensive statistical methods to fully understand and validate the performance of different network configurations.

Furthermore, the application of robust statistical techniques, including cross-validation, paired t-tests [120], Wilcoxon signed-rank tests [121], and bootstrap methods, can enhance the reliability, generalizability, and rigor of findings in deep learning model evaluations. These approaches collectively provide a comprehensive framework for assessing model performance beyond subjective evaluation metrics alone. Thus, future research should prioritize not only advancing novel architectures but also ensuring meticulous statistical validation of performance improvements to substantiate their clinical efficacy.

In parallel with these advancements, the integration of foundation models into the harmonization process holds the potential to further refine image quality and consistency across diverse datasets. Foundation models, which are large-scale, pre-trained deep learning models, have recently attracted significant attention across various deep learning challenges. These models are trained on extensive datasets to enhance generalization, contextual reasoning, and adaptability across different modalities. They can be fine-tuned for new tasks using task-specific prompts without the need for extensive retraining or labeled data. The field of medical imaging is increasingly exploring these models to leverage their advanced capabilities and improve outcomes [122].

While the application of foundation models in image harmonization is still an emerging field, these models offer substantial potential for improving consistency and compatibility across diverse medical imaging datasets. Future research should focus on exploring and optimizing the use of these models, conducting comprehensive quantitative comparisons, and addressing the specific challenges associated with harmonization in medical imaging.

Conclusion and future direction

This review provides a comprehensive overview of state-of-the-art deep learning-based methods for harmonizing Magnetic Resonance Imaging (MRI) scans. We categorized harmonization approaches based on their underlying network architecture, including U-Net, Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), flow-based models, and transformers. We surveyed current literature on MRI harmonization and report significant progress in harmonization techniques, with improvements observed in downstream tasks that rely on harmonized images.

Despite these advancements, several challenges remain:

- Data Standardization: The wide variety of acquisition parameters across scanners, disease states, and patient demographics (e.g., gender) can affect brain size and pose challenges for harmonization robustness. Developing standardized datasets encompassing these diverse scenarios and incorporating comprehensive evaluations for each challenge would be valuable.
- Evaluation Metrics: Current metrics primarily focus on contrast similarity. Novel metrics are needed to assess how well harmonization techniques address the "shift problem" (differences in image intensity distributions) while preserving crucial anatomical information.
- Multi-Site Harmonization: Current methods often focus on harmonization between two specific sites. Exploring techniques that can handle data from multiple sites would be beneficial.
- Architectural Innovation: Combining the strengths of different network architectures (e.g., U-Net for segmentation and GANs for image generation) could lead to more robust harmonization solutions. Additionally, computational efficiency should be considered, as faster models are more practical for real-world applications.
- Generalizability: Extending harmonization frameworks beyond specific MRI contrasts (T1-weighted, PD-weighted, T2-FLAIR) and even exploring other modalities like PET or CT could be a promising research direction.

By addressing these challenges and exploring new avenues, deep learning has the potential to further revolutionize MRI harmonization, ultimately leading to improved medical diagnosis and treatment planning.

Acknowledgements
None

Author contributions
SA performed literature search and wrote first draft of the manuscript. BV reviewed the first draft and conceived the project. SA, HA, JC, NSB, GP, and BV reviewed all subsequent drafts and approved the final manuscript.

Funding
This work was funded by the National Institutes of Health (5R01NS128486-03).

Availability of data and materials
No datasets were generated or analyzed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 10 May 2024 Accepted: 6 August 2024
Published online: 31 August 2024

References

1. Bangerter NK, Morrell G, Grech-Sollars M. Magnetic resonance imaging. In: bioengineering innovative solutions for cancer. Cambridge: Academic Press; 2019. p. 163–94.
2. Khanduri S. Magnetic resonance imaging physics. In: textbook of radiology for CT and MRI technicians with MCQs. 2018;109–109. https://doi.org/10.5005/jp/books/14192.

3. Carré A, Battistella E, Niyoteka S, Sun R, Deutsch E, Robert C. AutoComBat: a generic method for harmonizing MRI-based radiomic features. Sci Reports. 2022;12(1):12762. https://doi.org/10.1038/s41598-022-16609-1.
4. Fratini M, Abdollahzadeh A, DiNuzzo M, Salo RA, Maugeri L, Cedola A, et al. Multiscale imaging approach for studying the central nervous system: methodology and perspective. Front Neurosci. 2020;14:72. https://doi.org/10.3389/fnins.2020.00072.
5. Lakshmi MJ, Nagaraja RS. Brain tumor magnetic resonance image classification: a deep learning approach. Soft Comput. 2022;26(13):6245–53. https://doi.org/10.3390/cancers15164172.
6. Han KM, Ham BJ. How inflammation affects the brain in depression: a review of functional and structural MRI studies. J Clin Neurol. 2021;17(4):503. https://doi.org/10.3988/jcn.2021.17.4.503.
7. Noor MB, Zenia NZ, Kaiser MS, Mamun SA, Mahmud M. Application of deep learning in detecting neurological disorders from magnetic resonance images: a survey on the detection of Alzheimer's disease, Parkinson's disease and schizophrenia. Brain Inform. 2020;7:1–21.
8. IXI Brain Development Dataset. https://brain-development.org/ixi-dataset/.
9. Tixier F, Jaouen V, Hognon C, Gallinato O, Colin T, Visvikis D. Evaluation of conventional and deep learning based image harmonization methods in radiomics studies. Phys Med Biol. 2021;66(24):245009. https://doi.org/10.1088/1361-6560/ac39e5.
10. Lawrence E, Vegvari C, Ower A, Hadjichrysanthou C, De Wolf F, Anderson RM. A systematic review of longitudinal studies which measure Alzheimer's disease biomarkers. J Alzheimer's Dis. 2017. https://doi.org/10.3233/JAD-170261.
11. Pomponio R, Erus G, Habes M, Doshi J, Srinivasan D, Mamourian E, et al. Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. Neuroimage. 2020;208:116450. https://doi.org/10.1016/j.neuroimage.2019.116450.
12. Wang J, Lan C, Liu C, Ouyang Y, Qin T, Lu W, Chen Y, Zeng W, Philip SY. Generalizing to unseen domains: a survey on domain generalization. IEEE Trans Knowl Data Eng. 2022;35(8):8052–72.
13. Guan H, Liu M. Domain adaptation for medical image analysis: a survey. IEEE Trans Biomed Eng. 2021;69(3):1173–85.
14. Chen C, Dou Q, Chen H, Qin J, Heng PA. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. IEEE Trans Med Imaging. 2020;39(7):2494–505.
15. Zhou K, Liu Z, Qiao Y, Xiang T, Loy CC. Domain generalization: a survey. IEEE Trans Pattern Anal Mach Intell. 2022;45(4):4396–415.
16. Zhang J, Zuo L, Dewey BE, Remedios SW, Hays SP, Pham DL, Prince JL, Carass A. Harmonization-enriched domain adaptation with light fine-tuning for multiple sclerosis lesion segmentation. InMed Imaging 2024 Clin Biomed Imaging. 2024;12930:635–41.
17. Tanaka SC, Yamashita A, Yahata N, Itahashi T, Lisi G, Yamada T, et al. A multi-site, multi-disorder resting-state magnetic resonance image database. Sci Data. 2021;8(1):227. https://doi.org/10.6084/m9.figshare.14716329.
18. Stamoulou E, Spanakis C, Manikis GC, Karanasiou G, Grigoriadis G, Foukakis T, et al. Harmonization strategies in multicenter MRI-based radiomics. J Imaging. 2022;8(11):303. https://doi.org/10.3390/jimaging8110303.
19. Orlhac F, Eertink JJ, Cottereau AS, Zijlstra JM, Thieblemont C, Meignan M, et al. A guide to ComBat harmonization of imaging biomarkers in multicenter studies. J Nuclear Med. 2022;63(2):172–9.
20. Hu F, Chen AA, Horng H, Bashyam V, Davatzikos C, Alexander-Bloch A, et al. Image harmonization: a review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization. Neuroimage. 2023;20:120125.
21. Roca V, Kuchcinski G, Pruvo JP, Manouvriez D, Leclerc X, Lopes R. A three-dimensional deep learning model for inter-site harmonization of structural MR images of the brain: extensive validation with a multicenter dataset. Heliyon. 2023. https://doi.org/10.1016/j.heliyon.2023.e22647.
22. Ayaz A, Al KY, Amirrajab S, Lorenz C, Weese J, Pluim J, et al. Brain MR image simulation for deep learning based medical image analysis networks. Comput Methods Programs Biomed. 2024;248:108115.
23. Klemenz AC, Albrecht L, Manzke M, Dalmer A, Böttcher B, Surov A, et al. Improved image quality in CT pulmonary angiography using deep learning-based image reconstruction. Sci Reports. 2024;14(1):2494.
24. Deng L, Lan Q, Zhi Q, Huang S, Wang J, Yang X. Deep learning-based 3D brain multimodal medical image registration. Med Biol Eng Comput. 2024;62(2):505–19.
25. Wen G, Shim V, Holdsworth SJ, Fernandez J, Qiao M, Kasabov N, et al. Machine learning for brain MRI data harmonisation: a systematic review. Bioengineering. 2023;10(4):397. https://doi.org/10.3390/bioengineering10040397.
26. Mali SA, Ibrahim A, Woodruff HC, Andrearczyk V, Müller H, Primakov S, et al. Making radiomics more reproducible across scanner and imaging protocol variations: a review of harmonization methods. J Personal Med. 2021;11(9):842. https://doi.org/10.3390/jpm11090842.
27. Bayer JMM, Thompson PM, Ching CRK, Liu M, Chen A, Panzenhagen AC, et al. Site effects how-to and when: an overview of retrospective techniques to accommodate site effects in multi-site neuroimaging analyses. Front Neurol. 2022;13:923988.
28. Zuo L, Liu Y, Prince JL, Carass A. An overview of disentangled representation learning for MR image harmonization. Deep Learn Med Image Anal. 2024;1:135–52.
29. Yang J, Liu J, Xu N, Huang J. TVT: transferable vision transformer for unsupervised domain adaptation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023; 520–530.
30. Ko K, Yeom T, Lee M. SuperstarGAN: generative adversarial networks for image-to-image translation in large-scale domains. Neural Netw. 2023;162:330–9. https://doi.org/10.1016/j.neunet.2023.02.042.
31. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12(3):e1001779. https://doi.org/10.1371/journal.pmed.1001779.
32. Di Martino A, Yan CG, Li Q, Denio E, Castellanos FX, Alaerts K, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol Psychiatry. 2014;19(6):659–67. https://doi.org/10.1038/mp.2013.78.

33. Weiner MW, Veitch DP, Aisen PS, Beckett LA, Cairns NJ, Green RC, et al. The Alzheimer's disease neuroimaging initiative 3: continued innovation for clinical trial improvement. Alzheimer's Dementia. 2017;13(5):561–71. https://doi.org/10.1016/j.jalz.2016.10.006.
34. Casey BJ, Cannonier T, Conley MI, Cohen AO, Barch DM, Heitzeg MM, et al. The Adolescent brain cognitive development (ABCD) study: imaging acquisition across 21 sites. Dev Cogn Neurosci. 2018;32:43–54. https://doi.org/10.1016/j.dcn.2018.03.001.
35. Ellis KA, Bush AI, Darby D, De Fazio D, Foster J, Hudson P, et al. The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. Int Psychogeriatrics. 2009;21(4):672–87. https://doi.org/10.1017/S1041610209009405.
36. Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J Cognit Neurosci. 2007;19(9):1498–507. https://doi.org/10.1162/jocn.2007.19.9.1498.
37. Marcus DS, Fotenos AF, Csernansky JG, Morris JC, Buckner RL. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. J Cognit Neurosci. 2010. https://doi.org/10.1162/jocn.2009.21407.
38. LaMontagne PJ, Keefe S, Lauren W, Xiong C, Grant EA, Moulder KL, et al. OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. MedRxiv. 2019. https://doi.org/10.1101/2019.12.13.19014902.
39. Ogbole GI, Adeyomoye AO, Badu-Peprah A, Mensah Y, Nzeh DA. Survey of magnetic resonance imaging availability in West Africa. Pan Afr Med J. 2018. https://doi.org/10.11604/pamj.2018.30.240.14000.
40. Rutt BK, Lee DH. The impact of field strength on image quality in MRI. J Magnet Reson Imaging. 1996;6(1):57–62. https://doi.org/10.1002/jmri.1880060111.
41. Ahmed SY, Hassan FF. Optimizing imaging resolution in brain MRI: understanding the impact of technical factors. J Med Life. 2023;16(6):920. https://doi.org/10.25122/jml-2022-0212.
42. Thrower SL, Al Feghali KA, Luo D, Paddick I, Hou P, Briere T, Li J, McAleer MF, McGovern SL, Woodhouse KD, Yeboa DN. The effect of slice thickness on contours of brain metastases for stereotactic radiosurgery. Adv Radiat Oncol. 2021;6(4):100708. https://doi.org/10.1016/j.adro.2021.100708.
43. Ma YJ, Jang H, Chang EY, Hiniker A, Head BP, Lee RR, Corey-Bloom J, Bydder GM, Du J. Ultrashort echo time (UTE) magnetic resonance imaging of myelin: technical developments and challenges. Quant Imaging Med Surg. 2020;10(6):1186.
44. Nazarpoor M. The effect of inversion times on the minimum signal intensity of the contrast agent concentration using inversion recovery t1-weighted fast imaging sequence. Med J Islamic Republic Iran. 2014;28:128.
45. Morrell GR, Schabel MC. An analysis of the accuracy of magnetic resonance flip angle measurement methods. Phys Med Biol. 2010;55(20):6157.
46. Sijbers J, Scheunders P, Bonnet N, Van Dyck D, Raman E. Quantification and improvement of the signal-to-noise ratio in a magnetic resonance image acquisition procedure. Magnet Reson Imaging. 1996;14(10):1157–63.
47. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process. 2004;13(4):600–12. https://doi.org/10.1109/TIP.2003.819861.
48. Ravano V, Démonet JF, Damian D, Meuli R, Piredda GF, Huelnhagen T, Maréchal B, Thiran JP, Kober T, Richiardi J. Neuroimaging harmonization using cGANs: image similarity metrics poorly predict cross-protocol volumetric consistency. Int Workshop Mach Learn Clin Neuroimag. 2022;18:83–92.
49. Parida A, Jiang Z, Anwar SM, Foreman N, Stence N, Fisher MJ, Packer RJ, Avery RA, Linguraru MG. Harmonization across imaging locations (HAIL): one-shot learning for brain MRI. arXiv. 2023. https://doi.org/10.48550/arXiv.2308.11047.
50. Treder MS, Codrai R, Tsvetanov KA. Quality assessment of anatomical MRI images from generative adversarial networks: human assessment and image quality metrics. J Neurosci Methods. 2022;374:109579.
51. Tronchin L, Sicilia R, Cordelli E, Ramella S, Soda P. Evaluating GANs in medical imaging. In: Engelhardt S, Oksuz I, Zhu D, Yuan Y, Mukhopadhyay A, Heller N, Huang SX, Nguyen H, Sznitman R, Xue Y, editors. Deep generative models, and data augmentation, labelling, and imperfections. MICCAI. Cham: Springer International Publishing; 2021. p. 112–21.
52. Bińkowski M, Sutherland DJ, Arbel M, Gretton A. Demystifying mmd gans. arXiv. 2018. https://doi.org/10.48550/arXiv.1801.01401.
53. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2018; 586–595.
54. Cackowski S, Barbier EL, Dojat M, Christen T. Imunity: a generalizable VAE-GAN solution for multicenter MR image harmonization. Med Image Anal. 2023;88:102799. https://doi.org/10.1016/j.media.2023.102799.
55. Sinha S, Thomopoulos SI, Lam P, Muir A, Thompson PM. Alzheimer's disease classification accuracy is improved by MRI harmonization based on attention-guided generative adversarial networks. Int Sympos Med Inform Proc Anal. 2021;12088:180–9. https://doi.org/10.1117/12.2606155.
56. Shao M, Zuo L, Carass A, Zhuo J, Gullapalli RP, Prince JL. Evaluating the impact of MR image harmonization on thalamus deep network segmentation. Med Imaging 2022 Image Proc. 2022;12032:115–21. https://doi.org/10.1117/12.2613159.
57. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging. 2015;15:1–28.
58. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical Image computing and computer-assisted intervention–MICCAI. Springer International Publishing: Cham; 2015. p. 234–41.
59. Dewey BE, Zhao C, Carass A, Oh J, Calabresi PA, van Zijl PCM, et al. Deep harmonization of inconsistent MR data for consistent volume segmentation. In: Gooya A, Goksel O, Oguz I, Burgos N, editors., et al., Simulation and synthesis in medical imaging: third international workshop, SASHIMI 2018, held in conjunction with MICCAI. Springer International Publishing: Cham; 2018. p. 20–30.

60. Shiri I, Ghafarian P, Geramifar P, Leung KHY, Ghelichoghli M, Oveisi M, et al. Direct attenuation correction of brain PET images using only emission data via a deep convolutional encoder-decoder (Deep-DAC). Eur Radiol. 2019;29:6867–79.
61. Zhang S, Niu Y. LcmUNet: a lightweight network combining CNN and MLP for real-time medical image segmentation. Bioengineering. 2023;10(6):712. https://doi.org/10.3390/bioengineering10060712.
62. Dewey BE, Zhao C, Reinhold JC, Carass A, Fitzgerald KC, Sotirchos ES, et al. DeepHarmony: a deep learning approach to contrast harmonization across scanner changes. Magn Reson Imaging. 2019;64:160–70. https://doi.org/10.1016/j.mri.2019.05.041.
63. Bottani S, Thibeau-Sutre E, Maire A, Stroër S, Dormont D, Colliot O, et al. Homogenization of brain MRI from a clinical data warehouse using contrast-enhanced to non-contrast-enhanced image translation with U-Net derived models. Med Imaging. 2022;12032:576–82. https://doi.org/10.1117/12.2608565.
64. Osman AF, Tamam NM. Deep learning-based convolutional neural network for intramodality brain MRI synthesis. J Appl Clin Med Phys. 2022;23(4):e13530. https://doi.org/10.1002/acm2.13530.
65. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proc IEEE Int Con Comput Vision. 2017. https://doi.org/10.1109/ICCV.2017.244.
66. Komandur D, Gupta U, Chattopadhyay T, Dhinagar NJ, Sophia I, California S, et al. Unsupervised harmonization of brain MRI using 3D CycleGANs and its effect on brain age prediction. Int Sympos Med Inform Proc Anal (SIPAIM). 2023. https://doi.org/10.1109/SIPAIM56729.2023.10373501.
67. Modanwal G, Vellal A, Buda M, Mazurowski MA. MRI image harmonization using cycle-consistent generative adversarial network. Med Imaging. 2020;11314:259–64. https://doi.org/10.1117/12.2551301.
68. Bashyam VM, Doshi J, Erus G, Srinivasan D, Abdulkadir A, Habes M, et al. Medical image harmonization using deep learning based canonical mapping: toward robust and generalizable learning in imaging. arXiv. 2020. https://doi.org/10.48550/arXiv.2010.05355.
69. Yan W, Fu Z, Sui J, Calhoun VD. 'Harmless' adversarial network harmonization approach for removing site effects and improving reproducibility in neuroimaging studies. Ann Int Conf IEEE Eng Med Biol Soc (EMBC). 2022. https://doi.org/10.1109/EMBC48229.2022.9871061.
70. Zhao F, Wu Z, Wang L, Lin W, Xia S, Shen D, Li G. UNC/UMN baby connectome project consortium. Harmonization of infant cortical thickness using surface-to-surface cycle-consistent adversarial networks. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, Yap P-T, Khan A, editors. International conference on medical image computing and computer-assisted intervention. Cham: Springer International Publishing; 2019. p. 475–83.
71. Huang X, Liu MY, Belongie S, Kautz J. Multimodal unsupervised image-to-image translation. In: Proceedings of the European conference on computer vision (ECCV). 2018. p. 172–89.
72. Liu M, Maiti P, Thomopoulos S, Zhu A, Chai Y, Kim H, et al. Style transfer using generative adversarial networks for multi-site MRI harmonization. Med Image Comput Comput Assist Intervent MICCAI. 2021. https://doi.org/10.1002/hbm.26422.
73. Liu M, Zhu AH, Maiti P, Thomopoulos SI, Gadewar S, Chai Y, et al. Style transfer generative adversarial networks to harmonize multisite MRI to a single reference image to avoid overcorrection. Hum Brain Mapp. 2023;44(14):4875–92. https://doi.org/10.1002/hbm.26422.
74. Saxena S, Teli MN. Comparison and analysis of image-to-image generative adversarial networks: a survey. arXiv. 2021. https://doi.org/10.48550/arXiv.2112.12625.
75. Bashyam VM, Doshi J, Erus G, Srinivasan D, Abdulkadir A, Singh A, et al. Deep generative medical image harmonization for improving cross-site generalization in deep learning predictors. J Magnet Reson Imaging. 2022;55(3):908–16. https://doi.org/10.1002/jmri.27908.
76. Roca V, Kuchcinski G, Pruvo JP, Manouvriez D, Lopes R. IGUANe: a 3D generalizable CycleGAN for multicenter harmonization of brain MR images. arXiv. 2024. https://doi.org/10.48550/arXiv.2402.03227.
77. Bai Z, Wang P, Xiao T, He T, Han Z, Zhang Z, Shou MZ. Hallucination of multimodal large language models: a survey. arXiv. 2024. https://doi.org/10.48550/arXiv.2404.18930.
78. Sarkar P, Ebrahimi S, Etemad A, Beirami A, Arık SÖ, Pfister T. Mitigating object hallucination via data augmented contrastive tuning. arXiv. 2024. https://doi.org/10.48550/arXiv.2405.18654.
79. Liang J, Yang X, Huang Y, Li H, He S, Hu X, Chen Z, Xue W, Cheng J, Ni D. Sketch guided and progressive growing GAN for realistic and editable ultrasound image synthesis. Med Image Anal. 2022;79:102461.
80. Miyato T, Kataoka T, Koyama M, Yoshida Y. Spectral normalization for generative adversarial networks. ArXiv. 2018. https://doi.org/10.48550/arXiv.1802.05957.
81. Kingma DP, Welling M. Auto-encoding variational bayes. ArXiv. 2013. https://doi.org/10.48550/arXiv.1312.6114.
82. Baur C, Denner S, Wiestler B, Navab N, Albarqouni S. Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study. Med Image Anal. 2021;69:101952. https://doi.org/10.1016/j.media.2020.101952.
83. Torbati ME, Tudorascu DL, Minhas DS, Maillard P, Decarli CS, Jae Hwang S. Multi-scanner harmonization of paired neuroimaging data via structure preserving embedding learning. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021; 3284–3293.
84. Jadon S. An overview of deep learning architectures in few-shot learning domain. arXiv. 2020. https://doi.org/10.48550/arXiv.2008.06365.
85. Pachetti E, Colantonio S. A systematic review of few-shot learning in medical imaging. ArXiv. 2023. https://doi.org/10.48550/arXiv.2309.11433.
86. Kotia J, Kotwal A, Bharti R, Mangrulkar R. Few shot learning for medical imaging. Mach Learn Algorithms Indust Appl. 2021.
87. Fatania K, Clark A, Frood R, Scarsbrook A, Al-Qaisieh B, Currie S, et al. Harmonisation of scanner-dependent contrast variations in magnetic resonance imaging for radiation oncology, using style-blind auto-encoders. Phys Imaging Radiat Oncol. 2022;22:115–22. https://doi.org/10.1016/j.phro.2022.05.005.

88. Jeong H, Byun H, Kang DU, Lee J. BlindHarmony." Blind" Harmonization for MR Images via Flow model. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023; 21129–21139.
89. Beizaee F, Desrosiers C, Lodygensky GA, Dolz J. Harmonizing flows: unsupervised MR harmonization based on normalizing flows. International Conference on Information Processing in Medical Imaging. 2023; 347–359.
90. Vigneshwaran V, Wilms M, Camacho MI, Souza R, Forkert N. Improved multi-site Parkinson's disease classification using neuroimaging data with counterfactual inference. In: Medical Imaging with Deep Learning. PMLR; 2024. p. 1304–17.
91. He K, Gan C, Li Z, Rekik I, Yin Z, Ji W, et al. Transformers in medical image analysis. Intell Med. 2023;3(1):59–78. https://doi.org/10.1016/j.imed.2022.07.002.
92. Torbunov D, Huang Y, Yu H, Huang J, Yoo S, Lin M, et al. Uvcgan: UNET vision transformer cycle-consistent GAN for unpaired image-to-image translation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023; 702–712.
93. Yao X, Lou A, Li H, Hu D, Lu D, Liu H, Wang J, Stoebner Z, Johnson H, Long JD, Paulsen JS. Novel application of the attention mechanism on medical image harmonization. Med Imaging. 2023;12464:184–94. https://doi.org/10.1117/12.2654392.
94. Hu X, Zhou X, Huang Q, Shi Z, Sun L, Li Q. Qs-attn: Query-selected attention for contrastive learning in i2i translation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022; 18291–18300.
95. Han D, Yu R, Li S, Wang J, Yang Y, Zhao Z, Wei Y, Cong S. MR Image harmonization with transformer. In 2023 IEEE International Conference on Mechatronics and Automation (ICMA). 2023; 2448–2453.
96. Dinsdale NK, Jenkinson M, Namburete AIL. Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. Neuroimage. 2021. https://doi.org/10.1016/j.neuroimage.2020.117689.
97. Guan H, Liu Y, Yang E, Yap PT, Shen D, Liu M. Multi-site MRI harmonization via attention-guided deep domain adaptation for brain disorder identification. Med Image Anal. 2021;71:102076. https://doi.org/10.1016/j.media.2021.102076.
98. Wolleb J, Sandkühler R, Bieder F, Barakovic M, Hadjikhani N, Papadopoulou A, et al. Learn to ignore: domain adaptation for multi-site MRI analysis. International Conference on Medical Image Computing and Computer-Assisted Intervention. 2022; 725–735.
99. Dewey BE, Zuo L, Carass A, He Y, Liu Y, Mowry EM, et al. A disentangled latent space for cross-site MRI harmonization. International Conference on Medical Image Computing and Computer-Assisted Intervention. 2020; 720–729.
100. Zuo L, Dewey BE, Carass A, Liu Y, He Y, Calabresi PA, et al. Information-based disentangled representation learning for unsupervised MR harmonization. International Conference on Information Processing in Medical Imaging. 2021; 346–359.
101. Zuo L, Liu Y, Xue Y, Han S, Bilgel M, Resnick SM, et al. Disentangling a single MR modality. MICCAI Workshop on Data Augmentation, Labelling, and Imperfections. Cham: Springer Nature Switzerland; 2022. p. 54–63.
102. Liu S, Yap P-T. Learning multi-site harmonization of magnetic resonance images without traveling human phantoms. Commun Eng. 2024;3(1):6.
103. Li H, Gopal S, Sekuboyina A, Zhang J, Niu C, Pirkl C, et al. Unpaired MR image homogenisation by disentangled representations and its uncertainty. Uncertainty for safe utilization of machine learning in medical imaging, and perinatal imaging, placental and preterm image analysis: 3rd international workshop. Cham: Springer International Publishing; 2021. p. 44–53.
104. Zhao F, Wu Z, Zhu D, Liu T, Gilmore J, Lin W, Wang L, Li G. Disentangling Site Effects with Cycle-Consistent Adversarial Autoencoder for Multi-site Cortical Data Harmonization. International Conference on Medical Image Computing and Computer-Assisted Intervention. 2023; 369–379
105. Wu M, Zhang L, Yap PT, Lin W, Zhu H, Liu M. Structural MRI harmonization via disentangled latent energy-based style translation. In: Cao X, Xuanang X, Rekik I, Cui Z, Ouyang X, editors. International Workshop on machine learning in medical imaging. Cham: Springer Nature Switzerland; 2023. p. 1–11.
106. Zuo L, Liu Y, Xue Y, Dewey BE, Bilgel M, Mowry EM, et al. HACA3: a unified approach for multi-site MR image harmonization. Comput Med Imaging Graphics. 2023;109:102285. https://doi.org/10.1016/j.compmedimag.2023.102285.
107. Wang X, Chen H, Tang SA, Wu Z, Zhu W. Disentangled representation learning. ArXiv. 2022. https://doi.org/10.48550/arXiv.2211.11695.
108. Jog A, Carass A, Roy S, Pham DL, Prince JL. Random forest regression for magnetic resonance image synthesis. Med Image Anal. 2017;35:475–88. https://doi.org/10.1016/j.media.2016.08.009.
109. Chang X, Cai X, Dan Y, Song Y, Lu Q, Yang G, et al. Self-supervised learning for multi-center magnetic resonance imaging harmonization without traveling phantoms. Phys Med Biol. 2022;67(14):145004. https://doi.org/10.1088/1361-6560/ac7b66.
110. Yurt M, Dalmaz O, Dar S, Ozbey M, Tınaz B, Oguz K, Çukur T. Semi-supervised learning of MRI synthesis without fully-sampled ground truths. IEEE Trans Med Imaging. 2022;41(12):3895–906.
111. Grigorescu I, Vanes L, Uus A, Batalle D, Cordero-Grande L, Nosarti C, et al. Harmonized segmentation of neonatal brain MRI. Front Neurosci. 2021;15:662005. https://doi.org/10.3389/fnins.2021.662005.
112. Tor-Diez C, Porras AR, Packer RJ, Avery RA, Linguraru MG. Unsupervised MRI homogenization: application to pediatric anterior visual pathway segmentation. Machine learning in medical imaging: MICCAI. 2020. 180–188.
113. An L, Chen J, Chen P, Zhang C, He T, Chen C, Zhou JH, Yeo BT. Of Aging LS, Alzheimer's disease neuroimaging initiative. Goal-specific brain MRI harmonization. Neuroimage. 2022;263:119570. https://doi.org/10.1016/j.neuroimage.2022.119570.
114. Jin C-B, Kim H, Liu M, Jung W, Joo S, Park E, et al. Deep CT to MR synthesis using paired and unpaired data. Sensors. 2019;19(10):2361. https://doi.org/10.3390/s19102361.
115. Wang T, Lei Y, Fu Y, Wynne JF, Curran WJ, Liu T, et al. A review on medical imaging synthesis using deep learning and its clinical applications. J Appl Clin Med Phys. 2021;22(1):11–36. https://doi.org/10.1002/acm2.13121.

116. Pan Y, Liu M, Lian C, Zhou T, Xia Y, Shen D. Synthesizing missing PET from MRI with cycle-consistent generative adversarial networks for Alzheimer's disease diagnosis. Med Image Comput Comput Assist Intervent MICCAI. 2018. https://doi.org/10.1007/978-3-030-00931-1_52.
117. Zhang J, Cui Z, Jiang C, Zhang J, Gao F, Shen D. Mapping in cycles: dual-domain PET-CT synthesis framework with cycle-consistent constraints. International Conference on Medical Image Computing and Computer-Assisted Intervention. 2022; 758–767.
118. Sharma R, Tsiamyrtzis P, Webb AG, Leiss EL, Tsekos NV. Learning to deep learning: statistics and a paradigm test in selecting a UNet architecture to enhance MRI. Magn Reson Mater Phys Biol Med. 2023;21:1–22.
119. Sharma R, Tsiamyrtzis P, Webb AG, Seimenis I, Loukas C, Leiss E, Tsekos NV. A deep learning approach to upscaling "low-quality" MR Images: an in silico comparison study based on the UNet framework. Appl Sci. 2022;12(22):11758.
120. Atanda OG, Ismaila W, Afolabi AO, Awodoye OA, Falohun AS, Oguntoye JP. Statistical Analysis of a deep learning based trimodal biometric system using paired sampling T-Test. International Conference on Science, Engineering and Business for Sustainable Development Goals (SEB-SDG). IEEE. 2023;1:1–10.
121. Akter S, Shamrat FJ, Chakraborty S, Karim A, Azam S. COVID-19 detection using deep learning algorithm on chest X-ray images. Biology. 2021;10(11):1174.
122. Azad B, Azad R, Eskandari S, Bozorgpour A, Kazerouni A, Rekik I, Merhof D. Foundational models in medical imaging: a comprehensive survey and future vision. ArXiv. 2023. https://doi.org/10.48550/arXiv.2310.18689.

Publisher's Note