nature biomedical engineering

Article

https://doi.org/10.1038/s41551-024-01283-7

A foundation model for enhancing magnetic resonance images and downstream segmentation, registration and diagnostic tasks

Received: 2 August 2023

Accepted: 17 October 2024

Published online: 05 December 2024

Check for updates

Yue Sun [1,2], Limei Wang [1,2], Gang Li [3], Weili Lin [3] & Li Wang [1]

In structural magnetic resonance (MR) imaging, motion artefacts, low resolution, imaging noise and variability in acquisition protocols frequently degrade image quality and confound downstream analyses. Here we report a foundation model for the motion correction, resolution enhancement, denoising and harmonization of MR images. Specifically, we trained a tissue-classification neural network to predict tissue labels, which are then leveraged by a 'tissue-aware' enhancement network to generate high-quality MR images. We validated the model's effectiveness on a large and diverse dataset comprising 2,448 deliberately corrupted images and 10,963 images spanning a wide age range (from foetuses to elderly individuals) acquired using a variety of clinical scanners across 19 public datasets. The model consistently outperformed state-of-the-art algorithms in improving the quality of MR images, handling pathological brains with multiple sclerosis or gliomas, generating 7-T-like images from 3 T scans and harmonizing images acquired from different scanners. The high-quality, high-resolution and harmonized images generated by the model can be used to enhance the performance of models for tissue segmentation, registration, diagnosis and other downstream tasks.

Structural magnetic resonance imaging (MRI) is an established, non-invasive and safe technique to characterize the human brain, owing to excellent soft tissue contrast and lack of radiation[1]. However, despite its advantages, any participant motion from head movement or even heart beats, breathing or blinking during acquisition can cause serious blurring and ghost artefacts[2], resulting in confounding effects and in challenges for subsequent neuroimaging analyses, including tissue segmentation, registration, atlas construction, parcellation and group comparisons. Moreover, motion artefacts are commonly present in MRIs of individuals across all age groups, from foetuses to elderly adults. In particular, collecting high-quality MRIs from young children (2–4 years of age) is extremely challenging[3], as they are typically active and have difficulty remaining still throughout the entire scan, with only 33–60% success rate as reported in refs. 4,5.

To mitigate motion artefacts, various techniques have been proposed and can be broadly classified into two categories: prospective[6,7] and retrospective correction methods[8–12]. Prospective methods aim to prevent artefacts during data acquisition. The most common

[1]Developing Brain Computing Lab, Department of Radiology and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [2]Joint Department of Biomedical Engineering, University of North Carolina at Chapel Hill and North Carolina State University, Chapel Hill, NC, USA. [3]Department of Radiology and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. e-mail: li_wang@med.unc.edu

https://doi.org/10.1038/s41551-024-01283-7

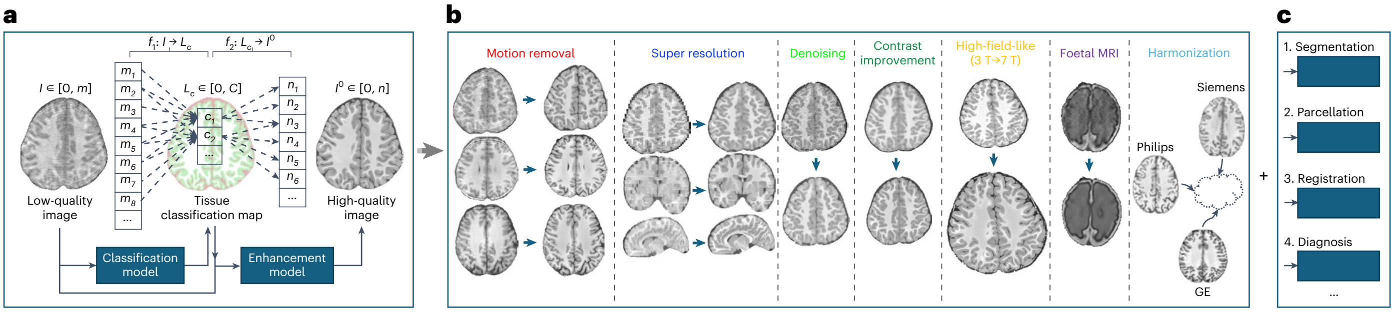

Fig. 1 | Overview of the BME-X model. a, A flowchart of the tissue-aware foundation model to improve image quality. **b**, Seven applications of the foundation model, including motion removal, super-resolution reconstruction, denoising, contrast improvement, high-field-like reconstruction, enhancement for foetal MRIs and harmonization. **c**, Extensions for downstream tasks, including tissue segmentation, parcellation, registration and diagnosis.

prospective strategy is to maintain a relatively constant relation between scanning coordinates and the object of interest, such as a head marker. However, this strategy has yet to be fully validated for routine use[13] and requires either expensive additional hardware or sequence modifications that can increase scan duration[14,15]. In contrast, retrospective methods remove or reduce artefacts after acquisition using navigators[16] or trackers[17], or iterative algorithms. Examples of retrospective correction methods include iterative estimation of phase correction[18], autofocusing methods[19,20], parallel-imaging reconstruction[21] and specific *k*-space sampling techniques (for example, periodically rotated overlapping parallel lines with enhanced reconstruction (PROPELLER)[16] or distributed and incoherent sample orders for reconstruction deblurring using encoding redundancy (DISORDER)[11]). These methods are in general computationally expensive and require supplemental data alongside the reconstructed images, such as the raw frequency domain (*k*-space) data, which is not always available for large-scale open datasets[13].

Recent deep learning-based methods have emerged as a potential solution for motion correction[9,10,22–27]. These methods eliminate the need for expensive additional hardware, sequences modification or supplemental *k*-space data. For instance, an artefact-correction method based on a multi-scale fully convolutional neural network was proposed[10] that extracts both high- and low-level features from input images across three different scales. In addition, an unsupervised, multi-contrast, brain image-based in-plane motion correction framework was developed[26]. A disentangled unsupervised cycle-consistent adversarial network (DUNCAN)[9], inspired by image translation techniques such as cycle-consistent generative adversarial networks (CycleGAN)[28] and Pix2Pix[29], can mitigate artefacts without requiring paired artefact-free and artefact-corrupted data during training. However, while these methods have shown success, they still produce images with residual artefacts and anatomically incorrect tissues. In particular, they suffer from two key limitations. First, most existing methods target adult brain MRIs, with few focusing on the entire lifespan, particularly the challenging young children (2–4 years of age) phase that is often affected by severe motion artefacts. Second, most existing methods ignore the brain anatomy, resulting in fake anatomical structures in the corrected results.

Besides motion artefacts, brain MRI is often acquired with large through-plane thickness or low resolution, due to limitations in imaging hardware, signal-to-noise ratio (SNR), time constraints and participant comfort. To mitigate these limitations, many super-resolution techniques[30–32] have been proposed to effectively increase image resolution. For instance, a non-local MRI upsampling (NLUP) method[30] was proposed that uses data-adaptive patch-based regularization combined with a subsampling coherence constraint to enhance voxel resolution. SynthSR[31] is an artificial intelligence technique that is publicly available for image synthesis and super resolution. In addition, brain MRI is further degraded by inevitable imaging noise, mainly from thermal noise susceptibility, stochastic variation, physiological process and other sources[33]. Denoising is therefore a critical task in image processing systems to improve image quality. However, most existing methods treat motion correction, super resolution and denoising as separate tasks, or propose sequential pipelines to address them one by one, resulting in cumulative errors and suboptimal solutions.

Another notable challenge in using MRI data is the large intercentre data heterogeneity, stemming from variations in scanner manufacturers, models and imaging parameters. This heterogeneity hinders comparability across sources and adversely affects both clinical and research outcomes. Although recent advancements in MRI technology have enhanced image quality, harmonization remains crucial, particularly in multi-centre studies. State-of-the-art methods, including machine-learning approaches and generative adversarial networks[28,34], have shown promise in mitigating these discrepancies. In addition, techniques such as combining batches (ComBat)[35], which employs empirical Bayes frameworks to adjust for batch effects, are widely adopted in neuroimaging. However, these methods often depend on specific training data or batch configurations, limiting their ability to generalize across diverse imaging protocols and scanners, which can lead to inconsistent harmonization results.

In this work, we propose a flexible and easy-to-implement brain MRI enhancement foundation (BME-X) model, as shown in Fig. 1, for notably improving the quality of brain MR images through motion correction, super resolution, denoising, harmonization and contrast enhancement. Our approach is rooted in the fundamental assumption that increasing image quality yields clearer and sharper image appearances, accompanied by a convergence of intensity ranges within each tissue type to a single value. Therefore, the complicated task of image reconstruction can be simplified to a tissue classification task. Then, high-quality images can be straightforwardly estimated with the guidance of classification results. Specifically, the BME-X model consists of a tissue classification network and a tissue-aware enhancement network. The tissue classification network is trained to predict the tissue labels, with the advantage of deep learning-based methods being robust to classify motion-corrupted MR images[36]. Then, the tissue-aware enhancement network leverages these tissue labels to generate high-quality images. To thoroughly evaluate the performance of our approach, we conducted extensive validation on 2,448 synthesized corrupted images and 10,963 in vivo images from 19 public datasets with mixed Siemens, General Electric (GE) and Philips scanners, covering the lifespan from foetuses to elderly individuals. The results demonstrate that BME-X substantially outperforms state-of-the-art methods both qualitatively and quantitatively, and is remarkably effective in motion correction, super-resolution reconstruction, denoising, handling pathological brain MRIs (such as those with multiple sclerosis (MS) or gliomas) and harmonization across scanners or sites.

 https://doi.org/10.1038/s41551-024-01283-7

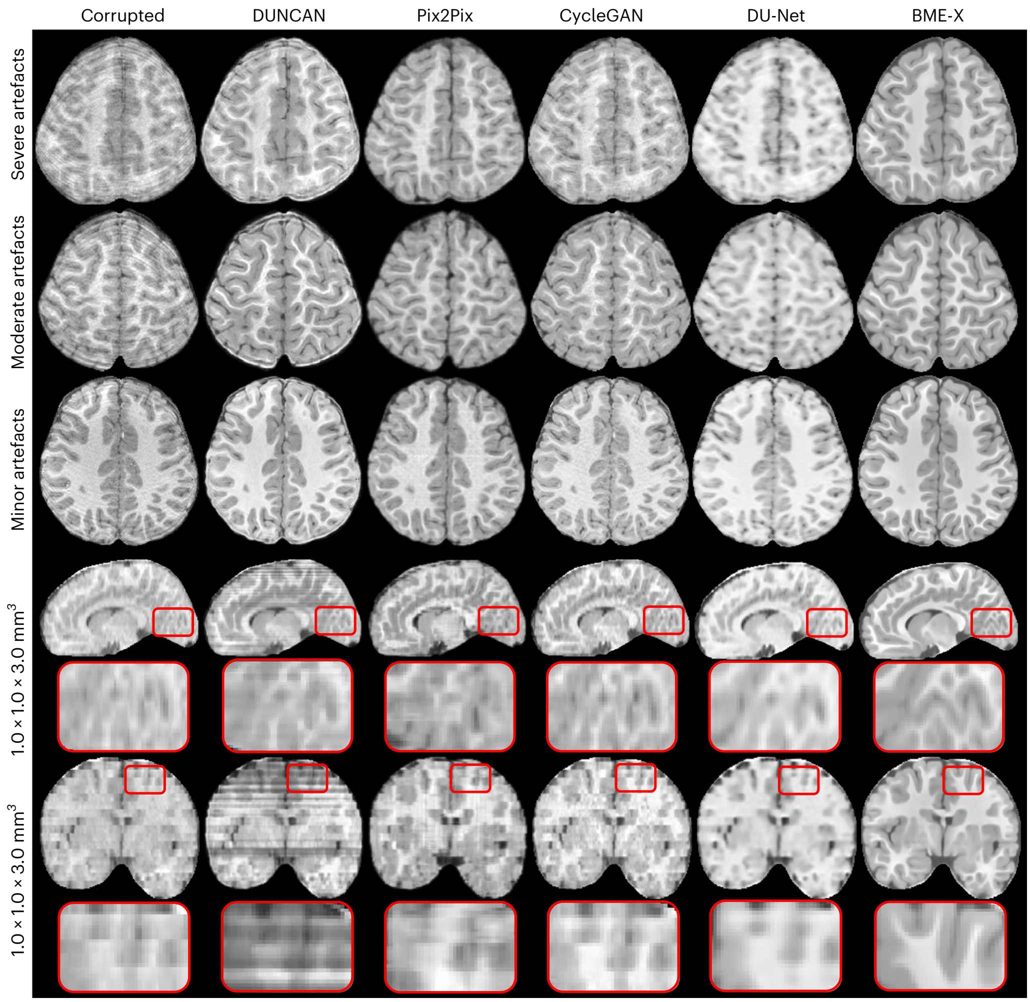

Fig. 2 | Visual comparison of the enhanced results for in vivo T1w images at 24 months old. In the first column, from top to bottom are T1w images with severe, moderate and minor artefacts, and two low-resolution images ($1.0 \times 1.0 \times 3.0$ mm^3). The corresponding enhanced images generated by different methods are shown from the second column to the last column. The resolution of the two-dimensional sagittal and coronal slices (the corrupted images in the last two rows) is 1.0×3.0 mm^2. The corresponding enhanced results are in a resolution of $0.8 \times 0.8 \times 0.8$ mm^3.

Moreover, our harmonized, high-quality, high-resolution images generated by BME-X facilitate various downstream tasks including tissue segmentation, parcellation, registration and diagnosis, as shown by our experimental results.

Results and discussion

We first show the advantages of the BME-X model by validating it on a diverse set of 2,088 synthesized corrupted images from 6 datasets (in 'Motion correction and super resolution on 24-month-old images' and 'Performance on 1,908 synthesized lifespan images' sections) and 10,963 in vivo images from 19 datasets (in 'Performance on 10,963 in vivo images, foetal to adulthood' section). Second, we conduct an ablation study to validate the effectiveness of the tissue classification module in enhancing image quality (in 'Ablation study' section). Third, we present a comprehensive evaluation of BME-X's robustness when handling corrupted images with varying levels of motion artefacts, downsampling, Gaussian noise, Rician noise and smoothing (in 'Robustness quantification' section). Fourth, we assess whether any potential bias was introduced during reconstruction (in 'Bias quantification during reconstruction' section). Fifth, we explore the capability of BME-X to estimate high-field-like (7-T-like) images from 3 T MRIs (in 'Application on reconstruction 7-T-like images from 3 T MRIs' section). Sixth, we demonstrate the effectiveness of BME-X in removing artefacts from abnormal brain MRIs with various conditions (in 'Application on pathological brain MRIs with lesions or gliomas' section). Seventh, we apply BME-X for the harmonization of brain MRIs across different scanners (in 'Application on harmonization across scanners' section). Eighth, we extend the BME-X model to several downstream tasks, including tissue segmentation, registration, parcellation and diagnosis (in 'Downstream tasks' section). Lastly, we discuss the limitations of this study and outline potential future work (in 'Outlook' section).

Competing methods

We compared the BME-X model with five state-of-the-art methods: (1) DUNCAN[9], (2) Pix2Pix[29], (3) CycleGAN[28], (4) densely connected U-Net (DU-Net)[37] and (5) NLUP[30]. DUNCAN (version 3.0) is a well-trained model that is publicly available (https://doi.org/10.5281/zenodo.3742351 (ref. 38)) for artefact removal using training data from University of North Carolina at Chapel Hill/University of Minnesota (UNC/UMN) Baby Connectome Project (BCP)[39]. To ensure a fair comparison, the competing artefact removal methods (Pix2Pix, CycleGAN and DU-Net) share the same training dataset with the BME-X model. Of note, to demonstrate the advantages of the tissue classification module in enhancing image quality, we conducted ablation studies by comparing the proposed framework with the DU-Net model that was trained for enhancement without the tissue classification module. NLUP (version 2.0, https://personales.upv.es/jmanjon/upsampling.htm) is a state-of-the-art super-resolution algorithm for brain images that uses a data-adaptive patch-based regularization in conjunction with a subsampling coherence constraint.

Evaluation metrics

To evaluate the quality of enhanced results, we initially regarded images that appear artefact-free as the ground truth reference. However, imaging noise is inevitable during acquisition, even for images that appear artefact-free. To ensure the quality of the ground truth, we applied a multi-resolution non-local means filter[40] to further denoise the original artefact-free images. We then used these denoised artefact-free images as the reference for six quality metrics, including mean squared error (MSE), peak SNR (PSNR), structural similarity index measure (SSIM)[41], multi-scale SSIM (MS-SSIM)[42], universal quality index (UQI)[43] and visual information fidelity (VIF)[44]. MSE measures the voxel-by-voxel difference between the reference and degraded images, while PSNR represents a measure of the peak error and is derived from MSE. SSIM quantifies the similarity in luminance, contrast and structural content of the reference and degraded images. MS-SSIM extends SSIM by computing it across multiple scales. UQI models any image distortion as a combination of three factors: loss of correlation, luminance distortion and contrast distortion. VIF evaluates the information shared between the reference and degraded images by exploring the connections between image information and visual quality. This metric demonstrates a higher correlation with radiologists' assessments of MR image quality compared with other metrics[45]. Except for MSE, where lower values indicate better performance, higher values in the remaining metrics reflect superior image quality. Default parameters were used to compute all six metrics in this study.

For in vivo images, evaluating image quality is challenging due to the absence of ground truth artefact-free references. To address this, a commonly used metric is the tissue contrast *t*-score (TCT)[46], which is defined as

$$\mathrm{TCT} = \frac{|\mu_{\mathrm{wm}} - \mu_{\mathrm{gm}}|}{\sqrt{\sigma_{\mathrm{wm}}^2 + \sigma_{\mathrm{gm}}^2}}, \qquad (1)$$

where (μ_{wm}, σ_{wm}^2) and (μ_{gm}, σ_{gm}^2) are the means and variances of white matter (WM) and grey matter (GM) intensities, respectively. TCT measures the contrast between WM and GM, as well as the intensity variation within each tissue. It is assumed that, after image enhancement, TCT should increase as the contrast between GM and WM increases and the intensity variation within each tissue decreases. By tracking TCT values, we can assess whether image quality has been improved after enhancement. In this work, we employed the Infant Brain Extraction and Analysis Toolbox (iBEAT V2.0)[47] to extract WM and GM for TCT calculation.

Motion correction and super resolution on 24-month-old images

In this experiment, we first performed a comprehensive comparison of our BME-X model against four competing methods on 24-month-old testing images. We have specifically chosen this age group because scans of young children often have substantial artefacts caused by their high activity level and difficulty in remaining still throughout the entire scan[3–5]. The testing data include both in vivo images with artefacts and low resolutions, and 180 synthesized corrupted images sourced from the BCP dataset[39]. Further information on the generation of these 180 images can be found in 'Testing dataset' section in Methods.

Performance comparison on in vivo images. Figure 2 shows a visual comparison of the enhanced results for in vivo corrupted images (acquired at 24 months old), with varying degrees of motion artefacts, and low-resolution images from our in-house collection. Our BME-X model was compared against four competing methods. The first three rows demonstrated that our method effectively removed all degrees of motion artefacts and enhanced tissue contrast, without introducing new artefacts. In contrast, the competing methods either failed to completely remove the artefacts or introduced new artefacts. The last two rows of Fig. 2 show an example of low-resolution images (1.0 × 1.0 × 3.0 mm^3), and the corresponding enhanced results (0.8 × 0.8 × 0.8 mm^3) obtained by different methods. The zoomed-in views reveal that the BME-X model can effectively restore tissue structure details and substantially improve image quality compared with the four competing methods.

Performance comparison on synthesized images. To evaluate the effectiveness of our BME-X model, we conducted a quantitative analysis on 180 synthesized corrupted images with simulated minor, moderate and severe motion artefacts, as well as varying resolutions. A representative set of corrupted images is shown in the first row of Fig. 3a. We used six metrics, MSE, PSNR, SSIM, MS-SSIM, UQI and VIF, to evaluate the performance. The quantitative results are reported in Fig. 3b, along with the statistical significance between our method and each competing method.

Our analysis revealed that none of the competing methods could completely remove the motion artefacts in the synthesized corrupted images, as shown in the first five rows of Fig. 3a. Moreover, we observed that the performance of the competing methods deteriorated considerably when dealing with severe motion artefacts and low resolutions. For instance, while Pix2Pix appeared to be more effective in removing artefacts than other competing methods based on qualitative inspection, motions and fuzziness were still present in its results, as seen in the third row. As an ablation study, we trained DU-Net with the same settings as our proposed framework but without the tissue classification module. The results showed that DU-Net could only remove minor motion artefacts at the expense of producing blurred images. On the other hand, BME-X demonstrated superior performance in removing motion artefacts and increasing image resolution, resulting in sharper and clearer enhanced results in the last row of Fig. 3a. The ablation comparison with DU-Net clearly highlighted the importance of the tissue classification module in enhancing image quality.

In addition to qualitative analysis, BME-X outperformed the competing methods in all six evaluation metrics for all cases, as shown in Fig. 3b. The differences between our method and each competing method were statistically significant with $P < 0.001$. These results demonstrated the effectiveness and robustness of our foundation model in enhancing image quality and removing motion artefacts.

Performance on 1,908 synthesized lifespan images

To demonstrate the robustness and effectiveness of the BME-X model in removing artefacts, we conducted extensive validation on a larger and more diverse dataset consisting of 1,908 images from five different datasets. The images were acquired with different scanners and age ranges, covering individuals from 6 months to 86 years of age, as detailed in Fig. 4a. Specifically, 883 images were scanned with Siemens scanners, 249 with GE scanners and 776 with Philips scanners. To evaluate the motion correction performance of each method, severe motion artefacts were intentionally added to all testing data. However, as the DUNCAN model did not include 6-month-old images as training data, we excluded it from the comparison on this age group.

Figure 4b shows a visual comparison of the corrected results on seven representative images. The results showed that BME-X effectively removed motion artefacts and enhanced image quality, while the competing methods still suffered from residual artefacts, with blurred boundaries and other issues. Furthermore, we calculated six metrics, MSE, PSNR, SSIM, MS-SSIM, UQI and VIF, to quantitatively evaluate the performance on 1,908 images. The results are shown in Fig. 4c and demonstrate that BME-X outperformed each of the competing methods, with lower MSE scores and higher PSNR, SSIM, MS-SSIM, UQI and VIF scores ($P < 0.001$).

https://doi.org/10.1038/s41551-024-01283-7

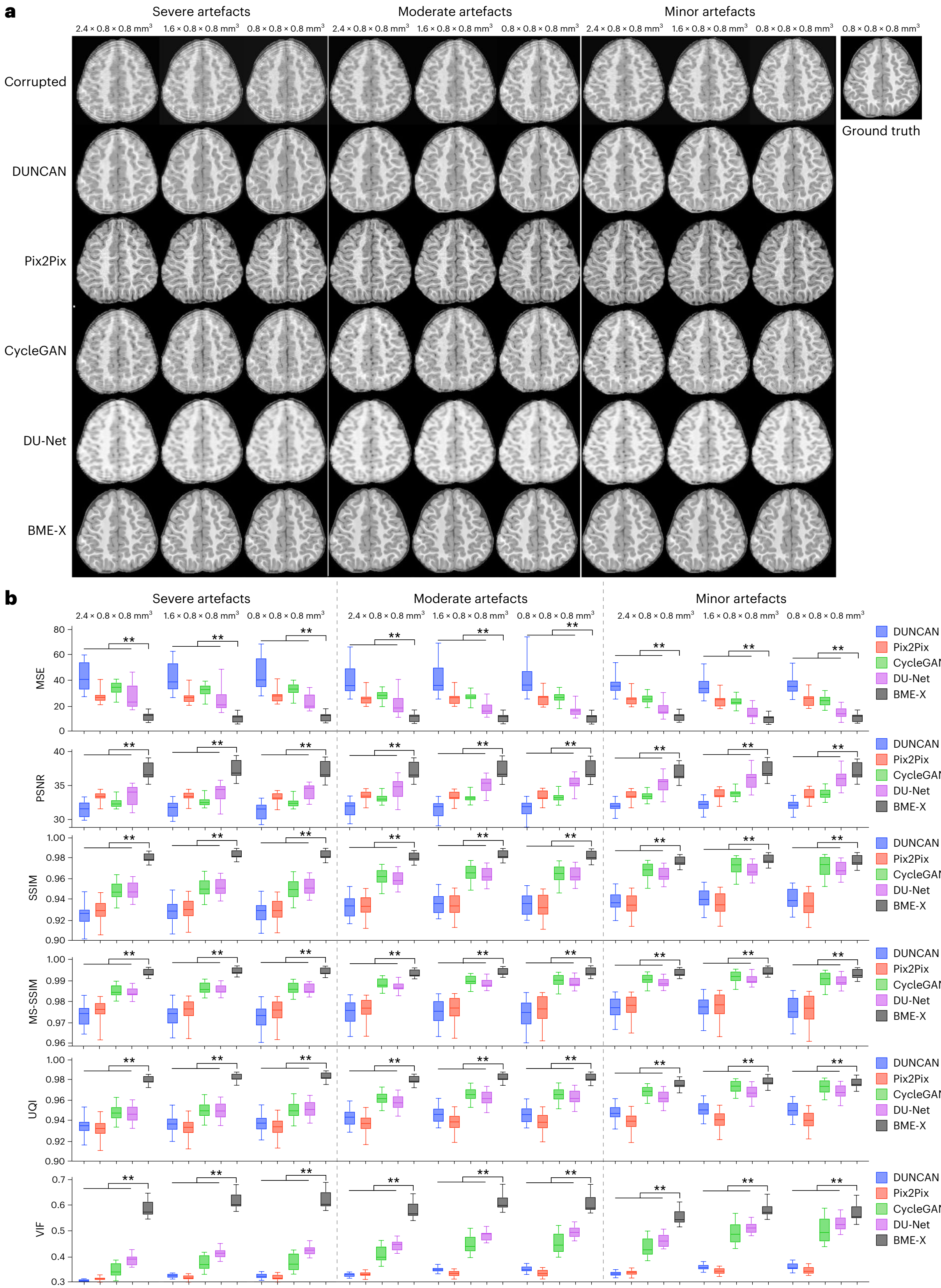

Fig. 3 | Enhanced results for 180 synthesized corrupted T1w images from BCP at 24 months old, generated by four competing methods and the foundation model. a, A visual comparison for enhanced results. The first row shows an exemplary T1w MRIs with different levels of simulated artefacts, at the original resolution (that is, 0.8 × 0.8 × 0.8 mm³) and lower resolutions. From the second row to the last row, the enhanced images are generated by four competing methods and the BME-X model. The corresponding ground truth is in the right top corner for reference. **b**, Quantitative comparison on 180 synthesized corrupted images using various image quality metrics (**$P < 0.001$, two-sided t-test, $n = 20$). The exact measurement values and corresponding P values are provided as Source data. In each box plot, the midline represents the median value, and its lower and upper edges represent the first and third quartiles. The whiskers go down to the smallest value and up to the largest.

https://doi.org/10.1038/s41551-024-01283-7

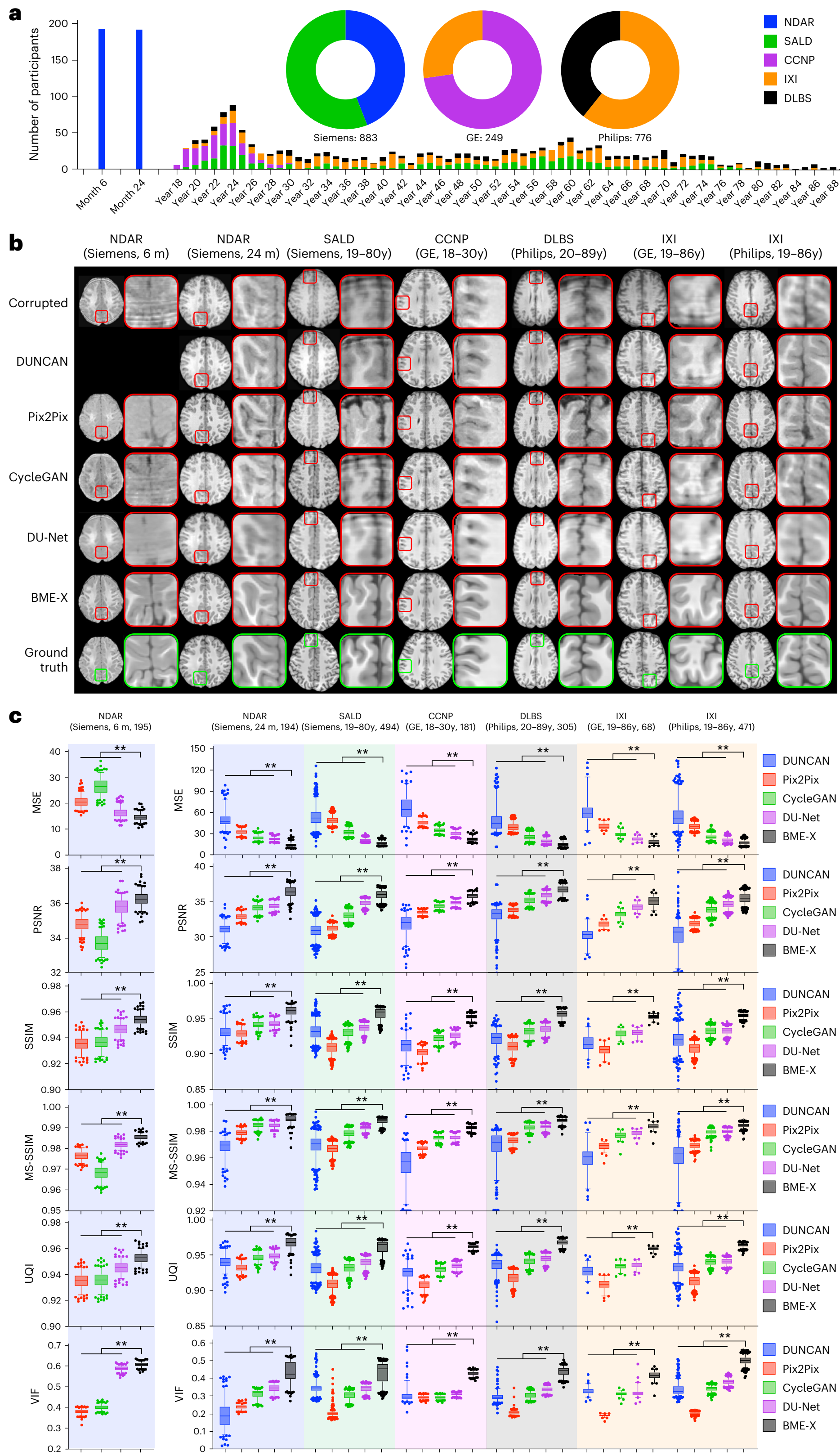

Fig. 4 | Enhanced results on 1,908 synthesized corrupted images from five datasets. a, Age distribution histograms and the corresponding distributions of three imaging scanners. **b**, A visual comparison of the enhanced results. Note that DUNCAN is not used to test the NDAR data at 6 months old since it did not include 6-month-old images as training data. **c**, A quantitative comparison of the enhanced results (**$P < 0.001$, two-sided t-test, n as indicated in the figures; m, month; y, year). The exact measurement values and corresponding P values are provided as Source data. In each box plot, the midline represents the median value, and its lower and upper edges represent the first and third quartiles. The whiskers are drawn down to the 5th percentile and up to the 95th. Points below and above the whiskers are drawn as individual points.

https://doi.org/10.1038/s41551-024-01283-7

To evaluate the performance of BME-X in terms of different scanners, we compared the quantitative evaluations of the enhanced results for each scanner type at approximately the same age range (15–90 years), as shown in Supplementary Fig. 1. The Siemens scans included 494 images from Southwest University Adult Lifespan Dataset (SALD)[48] dataset, the GE scans consisted of 249 images from Chinese Color Nest Project (CCNP)[49,50] and Information eXtraction from Images (IXI, http://brain-development.org/ixi-dataset/) datasets, and the Philips scans contained 776 images from Dallas Lifespan Brain Study (DLBS)[51] and IXI datasets. The comparison showed that BME-X achieved consistently better results than the competing methods on each of the three scanner types, with statistically significant differences ($P < 0.001$) for all six metrics. Moreover, BME-X achieved relatively consistent performance across scanners. These results highlighted the robustness and effectiveness of BME-X in removing motion artefacts and enhancing image quality, regardless of scanner type and participant age.

Performance on 10,963 in vivo images, foetal to adulthood

In addition to validating the BME-X model on synthesized data, we conducted further validation on 10,963 in vivo lifespan images from 19 datasets, as listed in Supplementary Table 1. Figure 5a shows the age distribution of the 10,963 testing images. These images were acquired using various scanners (Siemens, GE and Philips) and spanned ages from foetuses through to the elderly. Figure 5b–d shows results on representative lifespan images: (1) foetuses from 21 to 36 gestational weeks, (2) infants at 0, 3, 6, 9, 12 and 18 months old and (3) images from birth to late childhood, young adulthood, middle adulthood and late adulthood.

Mid-late foetuses. Figure 5b shows in vivo T2-weighted (T2w) foetal images from nine participants. These images are visibly affected by substantial amounts of motion, blurring and noise. Moreover, in some areas, severe artefacts render the thin GM layer nearly invisible. Nonetheless, BME-X effectively removed most of the artefacts, resulting in enhanced images that are much sharper visually. In addition, the thin GM layer was restored to a reasonable extent. This improvement is particularly noticeable in the second pair of corrupted and enhanced images in the sagittal view.

Infants from newborn to 18 months old. In Fig. 5c, we show in vivo low-resolution T1-weighted (T1w) images ($1.0 \times 1.0 \times 3.0$ mm^3) and their corresponding enhanced results ($0.8 \times 0.8 \times 0.8$ mm^3) at various ages (0, 3, 6, 9, 12 and 18 months old), from our in-house collection. These images exhibit visible artefacts from the sagittal and coronal views, probably due to their large through-plane thickness. However, by using our foundation model, we successfully restored detailed tissue structures with a high isotropic resolution, especially for sagittal and coronal views. For instance, a comparison between the corrupted and enhanced images at 18 months clearly highlighted that BME-X yielded sharper and clearer visual appearances, with cerebrospinal fluid (CSF), GM and WM tissue boundaries being distinctly visible.

Lifespan images from newborn to late adulthood. Figure 5d shows in vivo randomly selected lifespan images from various datasets, including Developing Human Connectome Project (dHCP)[52,53], National Database for Autism Research (NDAR)[54], Healthy Brain Network (HBN)[55], SALD, Human Connectome Projects (HCP)[56], Open Access Series of Imaging Studies 3 (OASIS3)[57], IXI, DLBS and Alzheimer's Disease Neuroimaging Initiative (ADNI)[58], acquired using mixed scanners. Although most images exhibit relatively good quality, motion artefacts and imaging noise remain evident, particularly in infants at 12 and 24 months and in elderly participants, due to the inherent challenges in acquiring data from these age groups. However, the BME-X model achieved substantial improvements in image quality, offering better tissue contrast and clearer visuals. These enhancements were particularly notable for infant participants at 12 and 24 months and elderly participants, providing improved visualization and interpretation of critical anatomical features.

As there are no ground truth reference images for the in vivo images, conventional reference-based metrics such as MSE, PSNR, SSIM, MS-SSIM, UQI and VIF cannot be used for quantitative evaluation. Alternatively, we used the TCT metric[46], which measures the contrast between WM and GM and the intensity variation within each tissue. Higher TCT metric values indicate an improvement in image quality. Figure 5e shows the TCT values before and after enhancement for each of the 10,963 in vivo testing images. The results demonstrated a substantial improvement in image quality (two-sided t-test, $P < 0.001$) after enhancement, with higher TCT values. It is important to note that the TCT values for testing images taken before 24 months tended to be lower than those of images taken after 24 months. This is due to the fact that, during this period (especially between 3 and 9 months old), the intensity distributions of voxels in GM and WM largely overlap due to ongoing myelination and maturation, leading to lower tissue contrast compared to adult brain images[59]. Overall, this extensive validation on a large and diverse in vivo dataset underscores the efficacy of the proposed work in improving the quality of lifespan imaging datasets acquired from mixed scanners.

Ablation study

Notably, the comparisons between the DU-Net model and the BME-X model in Figs. 2–4 and Supplementary Fig. 1 constituted ablation studies that investigated the importance of the tissue classification module in enhancing image quality. The ablation results clearly demonstrated that BME-X outperformed DU-Net both qualitatively and quantitatively.

In addition, we augmented the parameters of DU-Net to match the parameter count of BME-X (DU-Net-1) and increased it by 1.5 times (DU-Net-1.5), as detailed in Supplementary Note 1. Despite the increased parameters, the ablation comparison compellingly demonstrated that the proposed tissue-aware enhancement framework achieved superior performance, providing conclusive evidence for the substantial contribution of the embedded tissue classification module to enhancing image quality.

Robustness quantification

We examined BME-X's capability to handle different levels of artefact severity. To accomplish this, we applied our approach to 380 artefact-corrupted images that were artificially degraded from artefact-free images (the same images used in 'Motion correction and super resolution on 24-month-old images' section) with varying degrees of motion (Supplementary Fig. 2a), downsampling (Supplementary Fig. 2b), additive Gaussian noise and Rician noise (Supplementary Fig. 2c) and smoothing (Supplementary Fig. 2d). The visual comparisons of the enhanced results are shown in the left part of Supplementary Fig. 2, while the corresponding quantitative evaluations are presented in the right part. We used six metrics (that is, MSE, PSNR, SSIM, MS-SSIM, UQI and VIF) to measure performance.

Motion correction. Supplementary Fig. 2a shows three levels of motion-corrupted images (generated by an image-based motion-simulation approach using the software[60]): severe (random strength of motion amplitude amp = 3, frequency hz = $0.42 \times$ TR, where TR is the repeating time), severe$^+$ (amp = 4, hz = $0.48 \times$ TR) and severe^{++} (amp = 5, hz = $0.54 \times$ TR). From visual inspection, it was evident that BME-X effectively handled the severe case and achieved a satisfactory result for the severe$^+$ case. However, for the severe^{++} case, BME-X generated anatomically incorrect tissues, as indicated by the orange arrows. Quantitative evaluations also demonstrated a decline in the model's performance as motion artefact severity increased. These results suggested that, while BME-X was effective at mitigating motion artefacts in most scenarios, the severe^{++} artefacts remained challenging.

https://doi.org/10.1038/s41551-024-01283-7

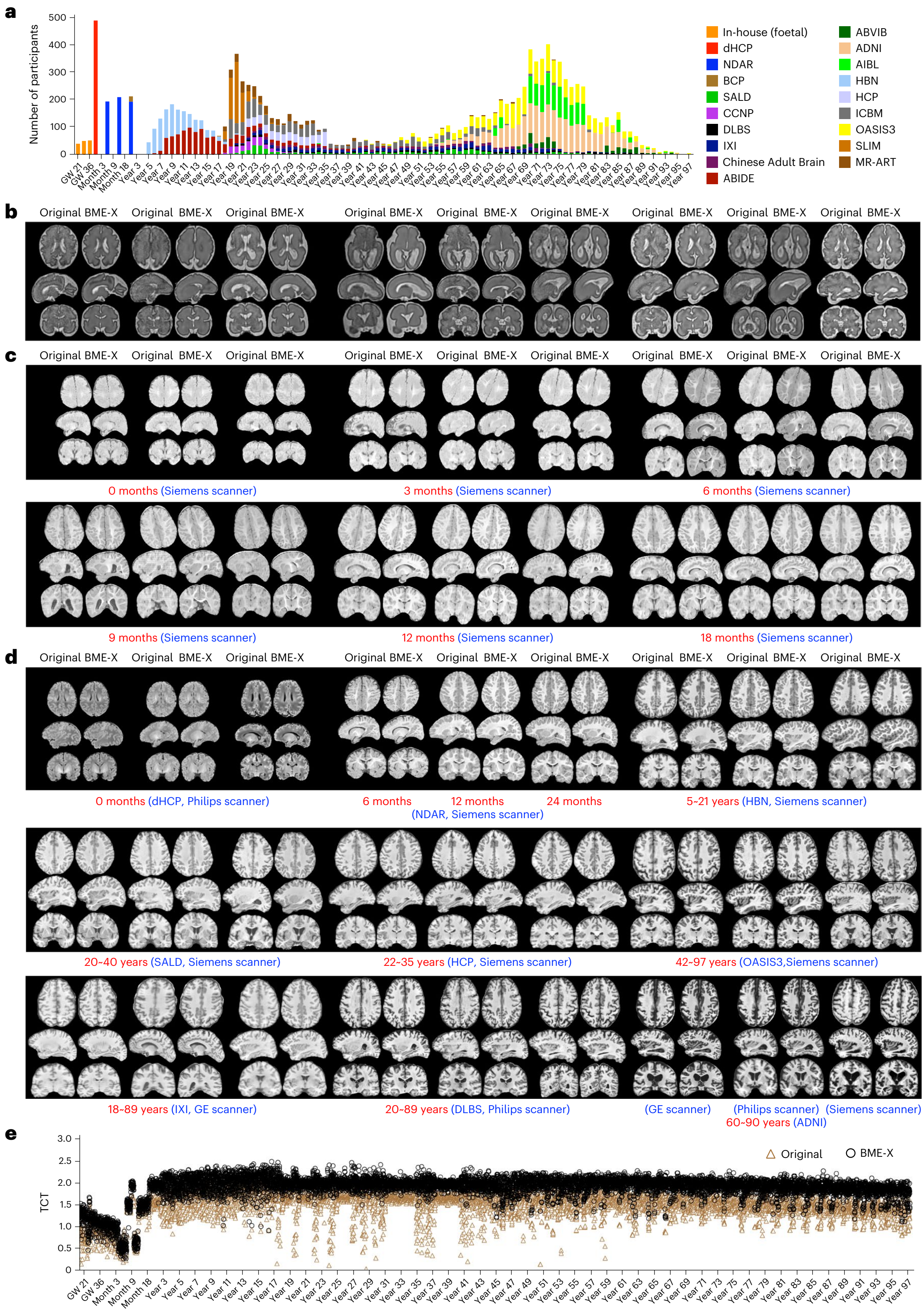

Fig. 5 | Enhanced results of the BME-X model for 10,963 in vivo low-quality images across the whole human lifespan, collected from 19 datasets. **a**, Age distribution. **b**, Mid-late foetal: original T2w images with severe motion and noise, and the corresponding enhanced results. **c**, From neonatal to early childhood: low-resolution T1w images (1.0 × 1.0 × 3.0 mm^3) and the corresponding enhanced results (0.8 × 0.8 × 0.8 mm^3). **d**, From neonatal to late adulthood: T1w images and the corresponding enhanced results. **e**, A comparison of TCT values for the 10,963 original images and the corresponding enhanced images by BME-X. The exact TCT values are provided as Source data.

https://doi.org/10.1038/s41551-024-01283-7

This indicated the need for further improvements, potentially through more advanced deep learning models or the incorporation of additional prior anatomical knowledge, to handle such extreme cases.

Super resolution. Although the BME-X model was not specifically designed for super-resolution tasks, we have found that it was capable of reconstructing low-resolution images to higher resolutions.

In this experiment, we downsampled 20 24-month-old images from $0.8 \times 0.8 \times 0.8$ mm^3 to lower resolutions (that is, $2.4 \times 0.8 \times 0.8$ mm^3, $3.2 \times 0.8 \times 0.8$ mm^3, $4.0 \times 0.8 \times 0.8$ mm^3 and $4.8 \times 0.8 \times 0.8$ mm^3), as shown in the first row of Supplementary Fig. 2b. We compared it with a state-of-the-art super-resolution algorithm, NLUP[30], which uses data-adaptive patch-based regularization in combination with a subsampling coherence constraint. Visual comparison in Supplementary Fig. 2b showed that the results of NLUP lacked consistent CSF structures compared with the ground truth, while BME-X generated more reasonable results. However, both NLUP and BME-X generated anatomically incorrect tissues when reconstructing ultralow-resolution images, such as $4.8 \times 0.8 \times 0.8$ mm^3, due to the limited information provided from the input. This is demonstrated in the last column of Supplementary Fig. 2b, where areas marked by orange arrows represent WM tissues, while the region indicated by blue arrows is a connected gyrus. We calculated quantitative evaluations using six metrics, and the results showed that both methods were sensitive to resolution. However, BME-X substantially outperformed NLUP.

Denoising. Noise is a common artefact during image acquisition, which is usually modelled by Rician distribution in magnitude MR images[61], making denoising a crucial task in image processing systems. Specifically, in the foreground regions where the SNR is high enough, the Rician noise can be approximated by a Gaussian distribution[62]. Nevertheless, to demonstrate the denoising performance of BME-X, we synthesized corrupted images with simulated Gaussian and Rician noise, respectively. As shown in Supplementary Fig. 2c, from left to right, the columns show corrupted images affected by increasing amounts of noise. The second (fifth) and third (sixth) rows depict the corresponding denoised images generated by DUNCAN[9] and BME-X, respectively. For images corrupted by Gaussian noises, DUNCAN can handle images with minor to moderate noise levels (standard deviation $\sigma = 0.01$, $\sigma = 0.03$ and $\sigma = 0.05$), while BME-X produced clearer and sharper results with better evaluation scores ($P < 0.001$). However, when denoising images with heavy Gaussian noise ($\sigma = 0.08$), neither method can produce satisfactory results. For images corrupted by Rician noises, compared with the competing method, BME-X consistently generated better results, as well as notably better quantitative results ($P < 0.001$).

Contrast improvement. In Supplementary Fig. 2d, we investigated the capability of BME-X in enhancing tissue contrast. To accomplish this, we first applied Gaussian filters with varying σ to smooth the images and then tested the smoothed images with both DUNCAN and BME-X. The results in Supplementary Fig. 2d demonstrate that BME-X outperformed DUNCAN in enhancing tissue contrast, even for images with ultralow contrast (that is, $\sigma = 1.2$). The enhanced images by BME-X are sharper and clearer than those produced by DUNCAN, which still suffer from blurring and incorrect anatomy, as indicated by the orange arrows. Furthermore, BME-X achieved notably better quantitative evaluations, with $P < 0.001$.

Bias quantification during reconstruction

In the above experiments, we demonstrated that BME-X exhibited a supreme ability to remove artefacts and improve image quality for both synthesized images and in vivo images, covering a broad age range from foetuses to elderly individuals. In this subsection, we further validated whether potential bias was introduced during the reconstruction, especially when we applied BME-X trained with data at 24+ months of age to adult images. Therefore, we compared brain characteristics, such as the tissue volumes and mean cortical thickness, from motion-free images and the enhanced images by BME-X and the four competing methods. Tissue volumes (WM, GM, CSF, ventricle and hippocampus) and mean cortical thickness were extracted by iBEAT V2.0[47]. Our hypothesis posits that, if the enhanced images closely resemble real images, characterizations such as volumes and thickness derived from enhanced images and original motion-free images should align.

To evaluate this, we used 420 in vivo images from the movement-related artefacts (MR-ART)[63] dataset, which includes both motion-free and motion-corrupted data acquired from the same participants, covering the adult lifespan (18–75 years old). We first validated the performance in artefact removal, as shown in Fig. 6a. The first column shows slight head motion (head motion 1, HM1) images, while the seventh column displays the paired images with more excessive head motion (head motion 2, HM2). Notably, for both HM1 and HM2, BME-X consistently outperformed competing methods in eliminating real artefacts while preserving anatomical details. To further scrutinize potential bias during reconstruction, we compared tissue volumes and mean cortical thickness for motion-free images (STAND) and enhanced images from motion-affected data (HM1 and HM2). Tissue volumes (WM, GM, CSF, ventricle and hippocampus) and mean cortical thickness are shown in Fig. 6b, along with the corresponding differences between the results on STAND and enhanced images. In addition, Cohen's d was employed to validate statistical differences between the results for STAND and enhanced images, as outlined in Supplementary Table 2.

As shown in Fig. 6b, the analysis revealed that enhanced images produced by DUNCAN, Pix2Pix and CycleGAN tended to undersegment WM and oversegment the remaining tissues (GM, CSF, ventricle and hippocampus). Supplementary Table 2 further confirms that most tissue volumes exhibited medium or large effect sizes compared with STAND (Cohen's $d > 0.4$), indicating that these three competing methods tended to introduce bias during the enhancement process. Conversely, the enhanced results obtained from DU-Net and BME-X consistently generated tissue volumes (WM, GM, CSF and ventricle tissues) in line with STAND, as indicated by volume differences close to zero (Fig. 6b). However, DU-Net tended to cause oversegmentation of the hippocampus, while BME-X consistently produced the most congruent results with STAND.

Furthermore, the detailed statistical analysis presented in Supplementary Table 2 underscores that BME-X demonstrated greater consistency with STAND results. Most Cohen's d values for the BME-X model were consistently smaller than those of all competing methods. Moreover, there was no significant effect size for tissue volumes of BME-X (Cohen's $d < 0.2$ for WM, GM, CSF and ventricle on HM1 and HM2 images, and hippocampus on HM1 images). Regarding mean cortical thickness, BME-X achieved consistency with STAND images, revealing no effect size in the enhanced results of HM1 images (Cohen's $d = 0.0441$) and small effect size for HM2 images (Cohen's $d = 0.3037$). Conversely, all competing methods exhibited substantial bias, with medium or large effect sizes.

We further investigated whether the effect size between Alzheimer's disease (AD) and normal cognition (NC) was preserved after reconstruction in Supplementary Note 2. We compared the volumes of several regions of interest (ROIs) related to AD, including the amygdala, hippocampus, thalamus, putamen, caudate and lateral ventricle, using 400+ images at baseline (aged between 70 and 85 years old) from the ADNI dataset. Supplementary Table 3 indicates that, after reconstruction, the effect size between AD and NC was preserved for the lateral ventricle and even increased for the amygdala, hippocampus, thalamus, putamen and caudate. In conclusion, the BME-X model exhibited superior performance in effectively removing real artefacts while consistently preserving essential tissue characteristics throughout the reconstruction process.

https://doi.org/10.1038/s41551-024-01283-7

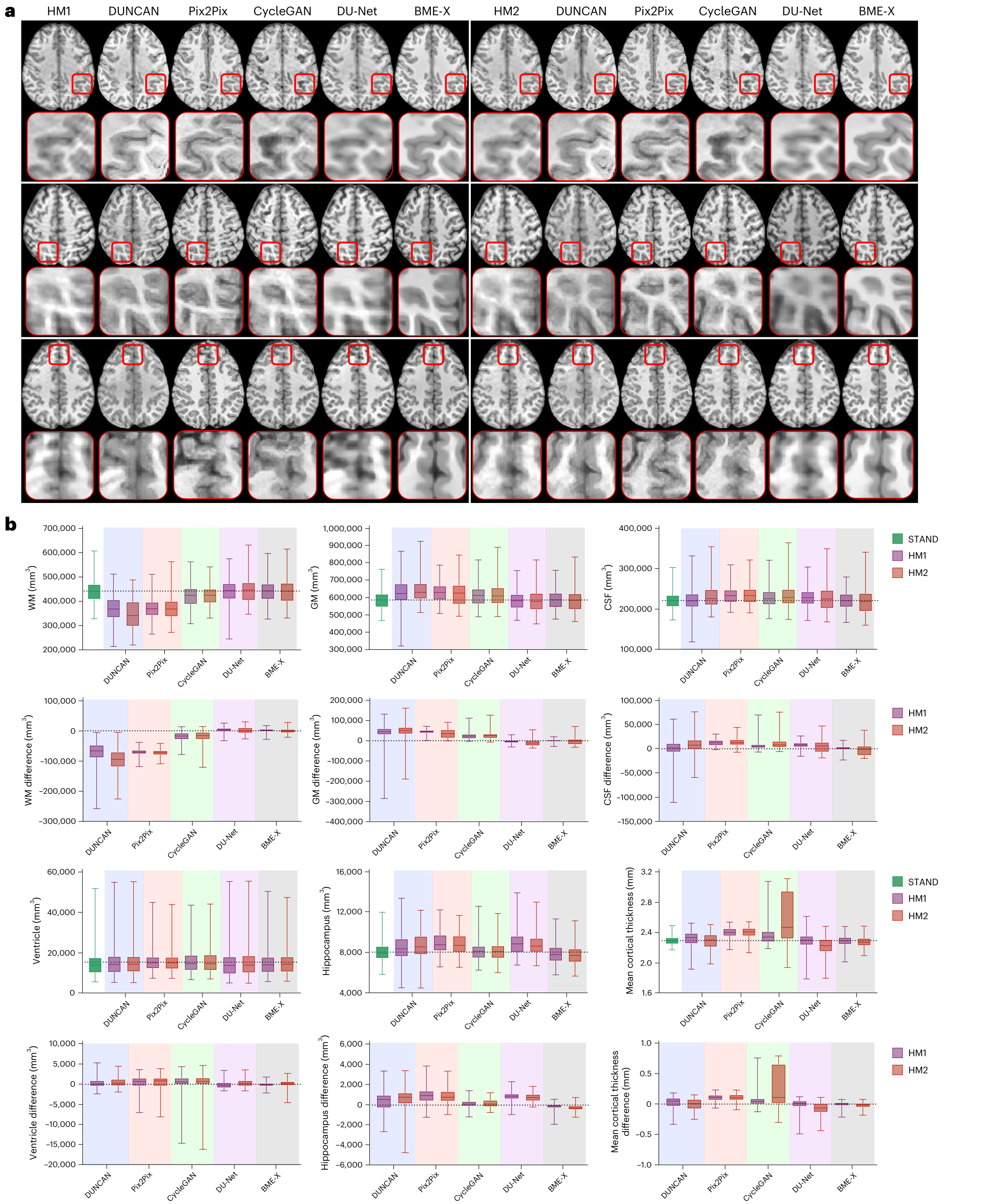

Fig. 6 | Enhancement results and the bias quantification for 280 in vivo corrupted T1w images from the MR-ART dataset, generated by competing methods and the BME-X model. a, A visual comparison of the enhanced results. The first and the seventh columns show in vivo corrupted images with two levels of real artefacts (that is, HM1 and HM2) acquired from the same participant. The remaining columns show the enhanced results generated by four competing methods and the BME-X model. **b**, A quantitative comparison on 280 in vivo corrupted images using tissue volumes (that is, WM, GM, CSF, ventricle and hippocampus), mean cortical thickness and the corresponding difference compared with STAND. In each box plot, the midline represents the median value, and its lower and upper edges represent the first and third quartiles. The whiskers go down to the smallest value and up to the largest. The dotted line represents the mean value or a zero value. The exact measurement values are provided as Source data. The corresponding effect sizes are listed in Supplementary Table 2.

https://doi.org/10.1038/s41551-024-01283-7

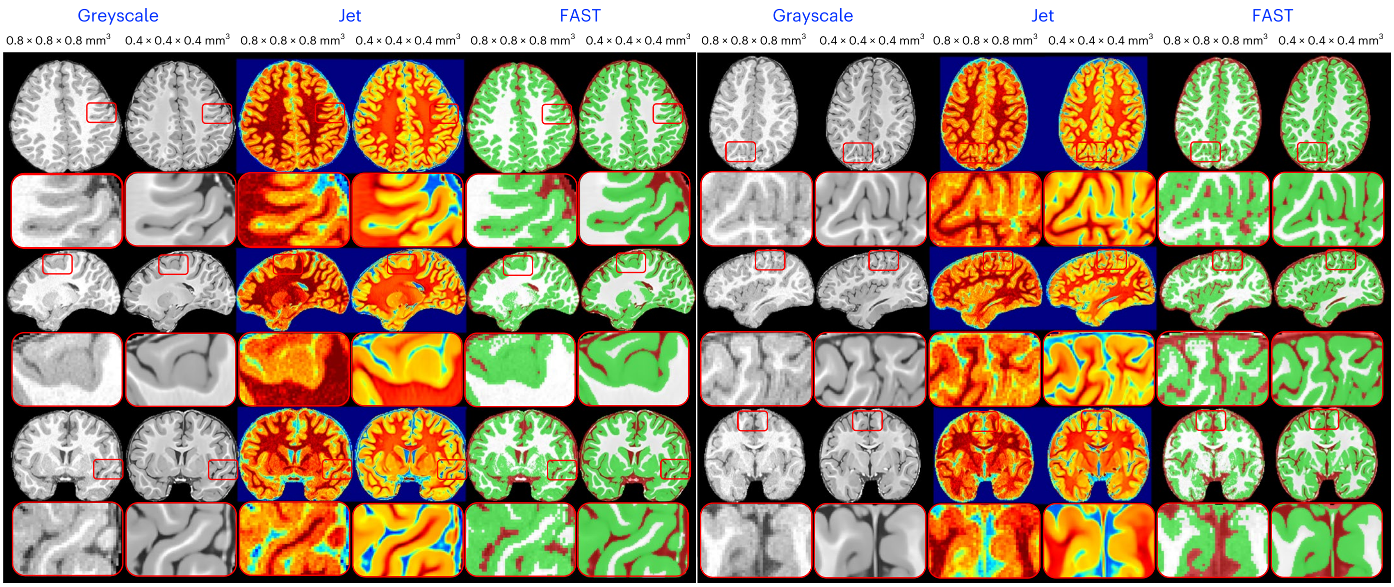

Fig. 7 | Ultrasuper-resolution reconstruction by the BME-X model. The corrupted images are randomly selected from the BCP dataset at 24 months of age, which were scanned with 3 T Siemens scanners (resolution $0.8 \times 0.8 \times 0.8$ mm^3). The enhanced results by the BME-X model are in a high-field and high-resolution space (7 T, resolution $0.4 \times 0.4 \times 0.4$ mm^3). Colour maps (that is, greyscale and Jet) of the corrupted and enhanced images are shown for a better visual comparison. To evaluate whether 7-T-like images could lead to more accurate tissue segmentations, we employed FAST to automatically perform tissue segmentations from the original 3 T and reconstructed 7-T-like images.

Application on reconstruction 7-T-like images from 3 T MRIs

High-field MRI systems like 7 T are known to yield images with superior spatial resolution and SNRs compared with lower-field strengths such as 3 T, enabling more precise visualization of small structures and subtle pathologies. However, they are limited by their higher costs and longer acquisition times[64]. In 'Motion correction and super resolution on 24-month-old images' and 'Robustness quantification' sections, we have presented evidence demonstrating the efficacy of the BME-X model in reconstructing high-resolution images (that is, 3 T images) from low-resolution inputs. Furthermore, we were motivated to apply the BME-X framework to estimate high-field-like (7-T-like) images from 3 T MRIs, given the advantages of high-field MRIs in providing high contrast and resolution to radiologists[65]. To achieve this, we trained an ultrasuper-resolution model using the proposed framework, using a set of training images from 28 participants acquired by a 7 T Siemens scanner (TR/TE = 5,000/2.45 ms, where TR is the repeating time and TE is the echo time) with a voxel resolution of $0.4 \times 0.4 \times 0.4$ mm^3 and a mean age of 25.9 ± 3.8 years[66]. We generated paired low- and high-quality training samples by degrading the 7 T images, similar to 'Simulated paired low- and high-quality images' section in Methods, and automatically generated the corresponding classification labels using iBEAT V2.0[47].

We used the proposed ultrasuper-resolution model to reconstruct images from $0.8 \times 0.8 \times 0.8$ mm^3 to $0.4 \times 0.4 \times 0.4$ mm^3, as shown in Fig. 7. To evaluate whether 7-T-like images facilitate more precise downstream tasks such as tissue segmentation compared with 3 T images, we randomly selected 3 T testing images from BCP[39] and employed the Functional Magnetic Resonance Imaging of the Brain (FMRIB)'s Automated Segmentation Tool (FAST)[67] to automatically perform tissue segmentation from the original 3 T and reconstructed 7-T-like images. Please note that the labels generated by FAST do not serve as ground truth, but as indirect evidence of image quality. Two representative images are shown in Fig. 7. It can be seen that the original 3 T images ($0.8 \times 0.8 \times 0.8$ mm^3) have blurred tissue boundaries, as seen in the zoomed views. Due to the limited voxel resolution, the CSF voxels are discontinuously buried in the deep GM, making it challenging to perform tissue segmentation, as also reflected in the Jet colormaps and segmentation maps. By contrast, after reconstruction, the 7-T-like images exhibit a much clearer appearance with fine-grained tissue structures, and the corresponding tissue segmentations are more accurate than those from 3 T images, as determined by visual inspection. In particular, these deeply buried CSF voxels could be precisely delineated from the 7-T-like images. Overall, these findings demonstrate the efficacy of BME-X in reconstructing high-field like images, thereby enabling more accurate tissue segmentation and improving the diagnostic performance.

Application on pathological brain MRIs with lesions or gliomas

Most results presented above were derived from typically developing brain MRIs. In this subsection, we validated the effectiveness of BME-X in removing artefacts from pathological MR images with different brain conditions. that is, MS and gliomas. Figure 8a shows nine exemplary T1w images with MS lesions, randomly selected from ref. 68. These lesions are typically small and appear as areas of hypointensity, appearing darker than the surrounding normal WM. Expert identification of brain lesions is denoted by red arrows for visualization purposes. Figure 8b shows three T1w brain images with gliomas, randomly chosen from ref. 69. In these images, glioma regions are larger and exhibit a darker appearance compared with WM or GM. Figure 8c illustrates post-contrast T1w images with gliomas randomly chosen from ref. 70, where the glioma boundaries exhibit bright contrast compared with WM or GM.

As shown in Fig. 8a, we synthesized the corrupted images with artefacts (that is, minor, moderate and severe) presented in the first, third and fifth rows. BME-X not only generated images with a clean and sharp appearance but also preserved small MS areas, as indicated by red arrows. Similarly, for large gliomas in T1w images (Fig. 8b) and post-contrast T1w images (Fig. 8c), BME-X effectively removed motion artefacts while retaining gliomas components, despite being trained exclusively with typically developing images. This finding suggested that the BME-X model could be applied to a broader range of brain images, including those with abnormalities. However, it is important to note that we assume the brain consists of WM, GM and CSF, and therefore, the lesions and gliomas were classified as one or mixed tissue

https://doi.org/10.1038/s41551-024-01283-7

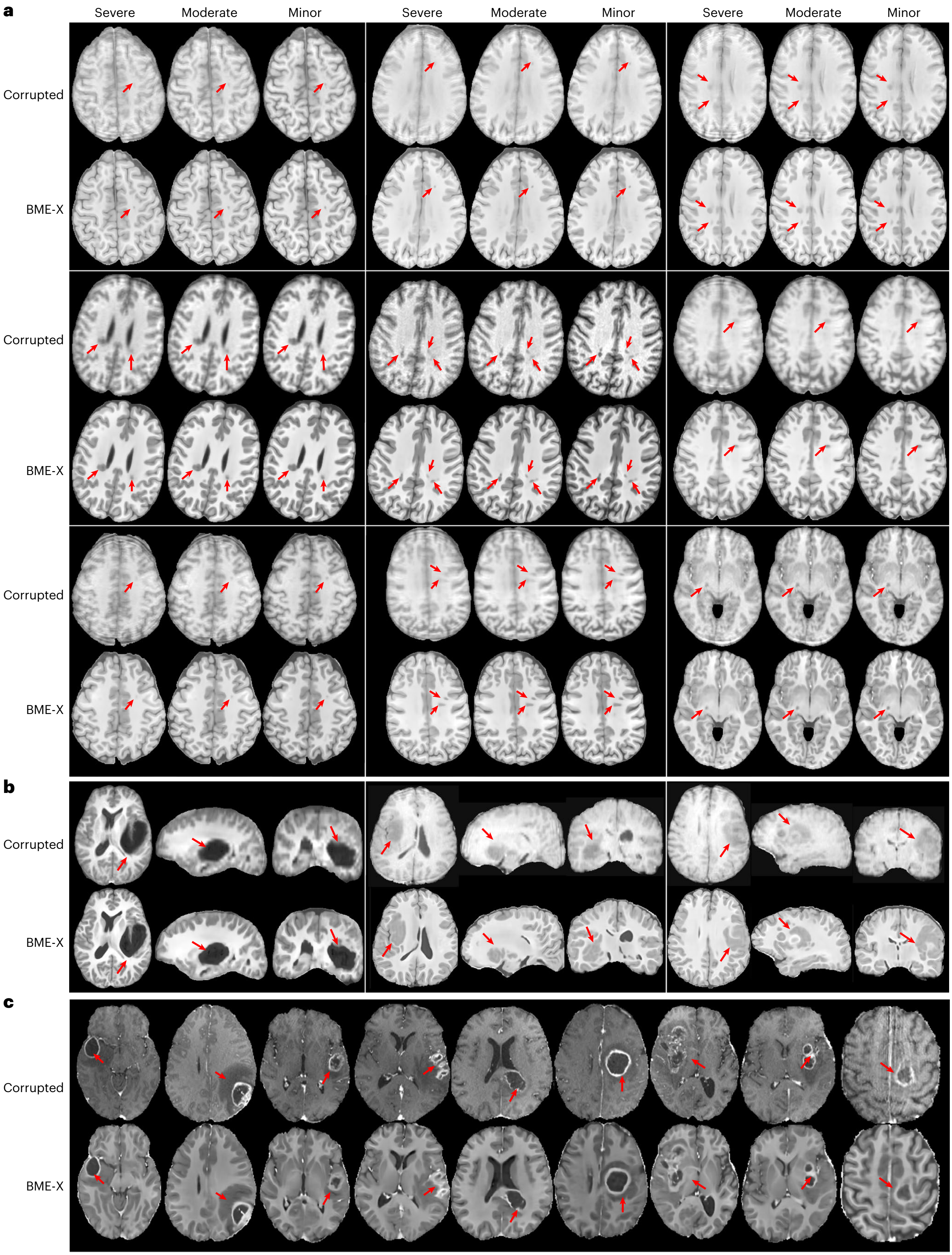

Fig. 8 | Enhanced results for the abnormal brain images with different brain conditions. a, T1w images were synthesized with simulated artefacts featuring MS lesions from nine participants. **b**, In vivo T1w brain images with gliomas. **c**, Post-contrast T1w brain images with gliomas. The lesions and gliomas are denoted by red arrows.

categories. To mitigate this issue, we plan to provide one additional label for the abnormal regions in the tissue classification module. This adjustment would enable the model to classify the lesions and gliomas as a separate class, thereby further enhancing image quality. In Supplementary Note 3, we further present a proof of concept demonstrating that incorporating an auxiliary lesion label enhances performance in preserving lesions.

Application on harmonization across scanners

The diversity in MRI acquisition protocols, encompassing variations in scanner types, magnetic field strengths and pulse sequences, introduces considerable disparities in image contrast and appearance. In Supplementary Fig. 3, the left section (blue) vividly portrays in vivo images acquired with Siemens, GE and Philips scanners using 1.5 T and 3 T field strengths, each displaying distinct imaging parameters. Clearly, the varying MRI scanner types and imaging parameters contribute to dynamic intensity histograms across scanners, emphasizing the challenges associated with harmonization. However, in the right section (orange) of the figure, the corresponding enhanced images produced by BME-X display more consistent contrast and appearance across different scanners. This improvement is evident in the resulting histogram distributions, which exhibit substantially enhanced uniformity. These results highlighted the capability of BME-X in mitigating the disparities introduced by diverse acquisition protocols. These findings strongly suggested that our proposed framework can not only effectively enhance images but also serve as a valuable harmonization technique across different MRI scanners, despite not being specifically designed for the harmonization task. By providing more consistent image characteristics, BME-X contributes to addressing the challenges posed by scanner variability in multi-centre studies and clinical settings. One compelling advantage of our harmonization is that it maps all scans, regardless of scanners or sites, to the same common space. In contrast, existing harmonization methods often require the assignment of a reference site[71], which may introduce site bias.

Downstream tasks

Our enhanced results hold substantial potential for various downstream tasks, including tissue segmentation, registration, parcellation and diagnosis. In the following, we will briefly introduce the results for each downstream task. For more details, please refer to Supplementary Notes 4–6.

Subcortical ROI segmentation. In this study, we first demonstrated the advantage of BME-X for subcortical ROI segmentation, including the thalamus, caudate, putamen, pallidus, hippocampus and amygdala, on 6-month-old infants from the BCP dataset. MRIs of 6-month-old infants present the most challenging labelling tasks among lifespan MRIs owing to the extremely low tissue contrast and also motion artefacts. As shown in Supplementary Fig. 6, the subcortical labels on the enhanced images are substantially better than those without enhancement.

Tissue segmentation. To evaluate the effectiveness of BME-X for the tissue segmentation task, we used synthetically corrupted images and applied the FAST[67] segmentation tool to both corrupted and enhanced images, with artefact-free images serving as the ground truth. As shown in Supplementary Fig. 7, the results demonstrated notable improvements in FAST segmentation accuracy on the enhanced images compared with those without enhancement, highlighting the efficacy of BME-X.

Furthermore, the BME-X model can facilitate end-to-end tissue segmentation by integrating existing architectures such as DU-Net[37] into the final layer of our framework, comprising the BME-X and segmentation modules (referred to as BME-S). As shown in Supplementary Fig. 7, BME-S demonstrated performance superior to that of a single segmentation model (DU-Net) for corrupted images, with higher Dice ratios and robustness to artefacts. Furthermore, BME-X substantially improved tissue segmentation on challenging isointense infant images, as shown in Supplementary Fig. 8. Enhanced images led to substantially improved segmentations compared with the original images, underscoring the effectiveness of our approach across various imaging scenarios.

Registration. To validate whether our enhancement can improve registration, we randomly selected 30 unique pairs of T1w images at baseline from the ADNI dataset. Each image underwent manual labelling to assess registration accuracy using the Dice ratio. To mimic real-world scenarios, we added varying motion artefacts as in 'Testing dataset' section in Methods and Rician noise to each image. Subsequently, within each pair, the two corrupted images were registered using the Demon's algorithm[72], and the resulting warped labels were employed to calculate the Dice ratio. Similarly, within each pair, we used BME-X enhanced images derived from the corrupted ones for paired registration. The registration performance, measured in terms of the Dice ratio, is presented in Supplementary Table 4. It demonstrates superior results for enhanced images, validating that BME-X enables more accurate registration. The statistical significance is evident with $P < 0.0001$ and Cohen's $d > 1.0$, indicating a highly substantial improvement.

Diagnosis. We extended our study to the diagnostic downstream task by using high-quality enhanced images in Supplementary Note 6. To illustrate the benefits of enhanced images, we trained a diagnostic model based on enhanced images, and compared it with a diagnostic model trained on the original images without enhancement. The results in Supplementary Table 5 demonstrate that, after enhancement, diagnostic accuracy improved substantially, along with sensitivity, specificity and F1 score.

Outlook

BME-X has shown outstanding performance in various tasks. However, there are several limitations and aspects to consider for future work. First, we plan to apply the foundation model to perform super-resolution reconstruction on ultralow-field images, such as scans with the Hyperfine machine[73]. This will help to make the foundation model more widely applicable. Second, our 24+ month model was trained only with typically developing infant participants aged 24+ months. Validation results on adult images (the MR-ART dataset in 'Bias quantification during reconstruction' section, the effect size between AD and NC in Supplementary Note 2 and the diagnostic task in Supplementary Note 6) have demonstrated the 24+ model's impressive generalization capability for adult brain scans. While promising, we acknowledge the potential bias for very older brains with substantial atrophy, as our training data are all from typically developing infants. To address this issue, we plan to train an ageing-dedicated model by exclusively including older brains with substantial atrophy as training data. It is straightforward to implement and will help mitigate the bias, enhancing its applicability to ageing populations. Third, the current implementation focused on only three major brain tissues: CSF, GM and WM. We will include more tissues and diseased regions as auxiliary labels in future work to address more complex medical scenarios and make the foundation model more valuable for clinical use. As a proof of concept, we have included lesions as an auxiliary label and trained a reconstruction model following the same framework as shown in Fig. 1. The comparison results without and with the auxiliary information are shown in Supplementary Fig. 5, demonstrating that small lesions are well preserved with auxiliary information. Fourth, BME-X was trained only on T1w/T2w images, which could limit its application range. Inspired by SynthSeg[74] and SynthSR[31], which can handle images with any contrast, we plan to enhance our approach by augmenting the training data through the synthesis of intensity images with different modalities. Overall, we believe that addressing these limitations will substantially improve the performance and applicability of the BME-X model.

Methods

To reconstruct high-quality images, the conventional strategy is to learn a mapping function f between a low-quality image I and a corresponding high-quality image I^0, that is, $f: I \rightarrow I^0$. For each voxel in the low-quality image, its intensity can be mapped to any intensity value in the high-quality image. Conversely, each voxel in the high-quality image can also be mapped to any intensity value in the low-quality image. This results in a complex many-to-many mapping, with a computational complexity of approximately $O(m \times n)$ for each voxel, where $[0, m]$ and $[0, n]$ are the intensity ranges of I and I^0, respectively. In this study, the BME-X model simplifies the complicated many-to-many reconstruction problem as two manageable tasks with notably lower time complexity: a tissue classification task and a tissue-aware enhancement task. Specifically, our foundation model first learns a tissue classification map L_c from I, that is, $f_1: I \rightarrow L_c$, where c is the number of tissues, then learns an enhancement mapping $f_2: L_c \rightarrow I^0$, with a total complexity of approximately $O(m+n)$. Since $O(m+n) \ll O(m \times n)$, the proposed strategy is much easier to learn an optimal solution than the conventional way. As shown in Fig. 1, the tissue classification module is trained to predict the tissue labels from low-quality images. Then, with guidance from the tissue labels, the tissue-aware enhancement module will be trained to estimate high-quality images from low-quality images.

Classification model

Numerous network architectures are suitable for this classification task, such as U-Net[75], densely connected convolutional network (DenseNet)[76], DU-Net[37], nnU-Net[77] and Transformer[78]. For this study, we selected the DU-Net architecture as our backbone classification model. The DU-Net architecture consists of an encoder and a decoder, as well as skip connections. Starting at a convolution layer (Conv), the encoder goes through three groups: dense block + Conv + batch normalization (BN) + rectified linear unit (ReLu), followed by a dense block to connect the encoder and decoder, and finally the decoder consists of three Deconv + BN +ReLu + dense block. At the end of the network, a Conv automatically generates class probabilities for each voxel. We use a cross-entropy loss L_1 to evaluate errors between the predicted tissue class probability and the expected class, defined as

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{c} x_j(i)\ln\hat{x}_j(i), \tag{2}$$

where $x_j(i)$ denotes the predicted tissue probability belonging to the jth class for the voxel i, $\hat{x}_j(i)$ is the corresponding ground truth tissue probability, c is the number of classes categories and N is the number voxels. Importantly, the tissue classification maps for both training and testing data are automatically predicted by the proposed classification model, rather than by iBEAT V2.0.

Enhancement model

The main contribution of this work lies in the integration of automated tissue classification with image enhancement, which provides a powerful framework for improving image quality. Specifically, the automated classification output from the classification module is concatenated with the low-quality input images to form a joint input for the enhancement module. This allows the model to exploit the tissue information to guide the enhancement process, resulting in high-quality images. To ensure the compatibility of the concatenated data, we apply a Conv + BN + ReLu block to pre-process both the tissue classification and intensity image data. The block facilitates feature extraction and transformation, which enhances the performance of the downstream enhancement model. To simplify the design and maintain consistency with the tissue classification network, we use the DU-Net[37] architecture as the backbone of the enhancement model. At the end of the network, a Conv layer is used to generate the high-quality image output. We use a MSE loss L_2 to calculate errors between predicted high-quality images and the corresponding ground truth high-quality images, defined as

$$L_2 = \frac{1}{N}\sum_{i=1}^{N}\left(y(i) - \hat{y}(i)\right)^2, \tag{3}$$

where $y(i)$ is the predicted intensity for the voxel i and $\hat{y}(i)$ is the corresponding ground truth intensity.

Simulated paired low- and high-quality images

To train the BME-X model in a supervised manner, paired low- and high-quality images are required. However, acquiring such paired images can be challenging and costly. One issue is that, even if a participant undergoes two scans within a short session, one being motion-free and the other motion-corrupted, there may still be intensity differences between the scans unrelated to the artefacts[13]. In addition, establishing voxel-to-voxel correspondence between motion-free and motion-corrupted images is difficult[9]. To address these challenges, we propose an alternative strategy in which simulated artefacts are added to real artefact-free MRIs. This allows us to use the artefact-corrupted and artefact-free MRIs as paired low- and high-quality images. By employing this strategy, we can jointly train the classification and enhancement modules, using the low-quality images as inputs and the corresponding tissue labels and high-quality images as the two targets. Furthermore, this strategy enables the generation of diverse artefact patterns, improving the framework's robustness to various types of image corruption.

To achieve this, we employed an image-based motion-simulation approach using a software[60]. Several sequence parameters are required to simulate realistic motion, such as the sampling trajectory, phase-encoding direction, TR, TE, total scan time and image resolution. For example, when simulating artefacts for BCP data[39], we used specific parameters, including phase-encoding direction (anterior–posterior), TR 2,400 ms, TE 2.2 ms, image resolution $0.8 \times 0.8 \times 0.8$ mm^3, random strength of motion amplitude $\mathrm{amp} = [0.01, 2.5]$ and frequency $\mathrm{hz} = [0, 0.0075] \times \mathrm{TR}$ (amp and hz are randomly chosen). Thus, the simulated motions, whether rotation, periodic, continuous or sudden, are derived from variable parameters, introducing the necessary complexity. Subsequently, we further added blurring through an artefact simulator[79], which employs Fourier transformation to obtain k-space data, followed by the application of a random phase shift to each k-space line along the readout direction. This process integrates random motion amounts by assuming a Gaussian distribution for phase shifts. The variability in blurring artefact severity is achieved by adjusting the standard deviation (σ) of this distribution. In this study, we randomly set σ within the range of [1, 2] for images with high tissue contrast (that is, images at 9, 12, 18 and 24+ months) and within the range of [0.5, 1.5] for images with low tissue contrast (that is, images at 0, 3 and 6 months), thereby ensuring dynamic variability in the artefact simulation function.

Implementations

The classification and enhancement networks were jointly trained using the Caffe[80] deep learning framework (Caffe 1.0.0-rc3), with a total loss $L = L_1 + \lambda L_2$, where λ was set as 10^{-7}. In the classification network, we set $c = 4$ in this work, including background, CSF, GM and WM. The kernels were initialized by Xavier. We used the stochastic gradient descent strategy for the network optimization. The learning rate was set as 0.005 and multiplied by 0.1 after each epoch. In this study, the models were trained for five epochs. All hyperparameters for the foundation model, such as the patch size, the loss weight λ, the number of epochs and the learning rate, were tuned on the basis of the validation set.

Training dataset

We used 52 foetal participants (21–36 gestational weeks) from our in-house collection and 464 participants (0–6 years old) from BCP[39] as our training data. Parents of all participants provided permission and informed consent before their participation. All procedures were approved by the University of North Carolina at Chapel Hill and the University of Minnesota Institutional Review Boards. All participants were normally developing without any pathology and their scans were artefact-free by careful visual inspection. To generate tissue labels for training the classification network, we leveraged a cerebrum-dedicated pipeline, iBEAT V2.0[47] (http://www.ibeat.cloud), which has successfully processed 50,000+ scans acquired with a wide variety of imaging protocols and scanners from 200+ institutions across the world and consistently outperforms competing tools, according to excellent feedbacks and comments provided by independent users. All pre-processed results by the iBEAT V2.0, including intensity inhomogeneity correction, skull stripping and cerebellum removal, passed the quality control. To generate paired low-quality images from the artefact-free images for training the tissue-aware enhancement network, we introduced simulated degradation. The artefact-free images were degraded by adding simulated rotation, periodic, continuous and sudden motions (via a motion simulation tool[60]), along with imaging noise and image blurring (via an artefact simulator[79]). It is important to note that the image degradation process was applied to the original brain images (before skull stripping).

In this study, to effectively correct artefacts for lifespan brain MRIs, we trained enhancement models at different developmental stages, including foetal phase, 0 months, 3 months, 6 months, 9 months, 12 months, 18 months and 24+ months of age. The rationale behind this training strategy is that brain images within each of these age groups exhibit highly representative and relatively consistent appearances, which has been extensively verified in previous studies[81–84]. Supplementary Table 1 presents the number of training data for each age group. We randomly extracted 2,000 patches (size: $40 \times 40 \times 40$ for images with a resolution of $0.8 \times 0.8 \times 0.8$ mm^3) from each training image. We divided the training samples, allocating 95% to the training set and 5% to the validation set. During testing, we employed the age-matched model for enhancement corresponding to the age of the input data. Notably, for participants aged from 24 months to 100 years old, we applied the same model (that is, 24+ months of age) for enhancement.

Testing dataset

In this study, we validated on 19 testing datasets acquired with various scanners and imaging protocols/parameters, including dHCP[52,53], NDAR[54], BCP[39], SALD[48], CCNP[49,50], DLBS[51], IXI (http://brain-development.org/ixi-dataset), Chinese Adult Brain[85], Autism Brain Imaging Data Exchange (ABIDE)[86], Aging Brain: Vasculature, Ischemia, and Behavior (ABVIB)[87], ADNI[58], Australian Imaging, Biomarkers and Lifestyle (AIBL)[88], HBN[55], HCP[56], International Consortium for Brain Mapping (ICBM)[89], OASIS3[57], Southwest University Longitudinal Imaging Multimodal (SLIM) Brain Data Repository[90], MR-ART[63] and T2w foetal images suffered from motion and imaging noise (from in-house collection). All sites of these testing datasets obtained study protocol approval from their respective Institutional Review Boards, and informed consent was provided by participants or their parents/legal guardians. We have diligently complied with all applicable ethical regulations during the utilization of all participants involved in this study. More imaging information about these datasets is listed in Supplementary Table 1. These datasets cover a broad age range, from foetuses to elderly individuals, enabling a comprehensive evaluation of the foundation model. Note that participants in the training and testing datasets are completely non-overlapping.

Synthesized data. To qualitatively evaluate the performance, we generated 2,088 synthesized corrupted images from BCP, NDAR, SALD, CCNP, DLBS and IXI datasets. Specifically, for the BCP dataset, we randomly selected 20 artefact-free 24-month-old T1w images and synthesized corrupted images with three different levels of motions: minor ($\mathrm{amp} = 1$, $\mathrm{hz} = 0.12 \times \mathrm{TR}$), moderate ($\mathrm{amp} = 2$, $\mathrm{hz} = 0.18 \times \mathrm{TR}$) and severe ($\mathrm{amp} = 3$, $\mathrm{hz} = 0.42 \times \mathrm{TR}$). We then downsampled these images into lower resolutions (that is, $0.8 \times 0.8 \times 0.8$ mm$^3 \rightarrow 1.6 \times 0.8 \times 0.8$ mm^3 and $2.4 \times 0.8 \times 0.8$ mm^3). This resulted in 180 corrupted images, with each artefact-free image corresponding to nine different levels of corruption (that is, three levels of motions and three levels of resolutions), which we used to evaluate the performance of our foundation model as well as competing methods. We chose images of 24-month-old participants for synthesis because scans at this age often suffer from substantial artefacts due to the difficulty of keeping the participants still during the scan[3–5]. For other testing datasets (that is, NDAR, SALD, CCNP, DLBS and IXI), we synthesized corrupted images with severe periodic motions (via a released software[60] with $\mathrm{amp} = 3$, $\mathrm{hz} = 0.42 \times \mathrm{TR}$).

In vivo data. We tested the foundation model on 10,963 in vivo images from the 19 testing datasets. The data were acquired using various scanners, such as Siemens, GE and Philips, with age phases ranging from mid-foetal to late adulthood. The MR-ART dataset[63], in particular, includes both motion-free and motion-corrupted images from the same participants, covering the adult lifespan (18–75 years old). Each participant was scanned under three conditions: staying still (STAND), slight head motion (HM1) and more excessive head motion (HM2). Of the MR-ART data, 140 participants had images for all three conditions, resulting in 280 in vivo corrupted images. Since establishing voxel-to-voxel correspondence between paired motion-free and motion-corrupted images, even from the same participant, is very difficult[9], STAND images are unsuitable as ground truth for the enhanced results of HM1 and HM2 images. Nonetheless, they serve to explore whether any potential bias was introduced during the reconstruction, such as variances in tissue volumes and cortical thickness between STAND and reconstructed data. For pre-processing, we resampled all testing images to $0.8 \times 0.8 \times 0.8$ mm^3 (the same resolution with training images) and used the toolbox[47] to perform skull stripping and cerebellum removal. More imaging information about the in vivo images is listed in Supplementary Table 1.

Additional testing images are presented in 'Results and discussion' to evaluate the performance, including (1) low-resolution images at 0 months, 3 months, 6 months, 9 months, 12 months and 18 months of age (from in-house collection) in 'Performance on 10,963 in vivo images, foetal to adulthood' section; (2) adult testing images with different brain conditions (MS and gliomas) from refs. 68–70 in 'Application on pathological brain MRIs with lesions or gliomas' section; and (3) isointense phase T1w images acquired with a 3 T GE scanner[91] in 'Downstream tasks' section.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The raw data generated in this study are available from dHCP[52,53] (https://biomedia.github.io/dHCP-release-notes), NDAR[54] (https://nda.nih.gov/edit_collection.html?id=19), BCP[39] (https://nda.nih.gov/edit_collection.html?id=2848), SALD[48] (http://fcon_1000.projects.nitrc.org/indi/retro/sald.html), CCNP[49,50] (https://ccnp.scidb.cn/en/detail?dataSetId=826407529641672704&version=V3&code=o00133), DLBS[51] (https://fcon_1000.projects.nitrc.org/indi/retro/dlbs.html), IXI (http://brain-development.org/ixi-dataset/), Chinese Adult Brain[85] (https://www.nitrc.org/projects/adultatlas), ABIDE[86] (https://fcon_1000.projects.nitrc.org/indi/abide/), ABVIB[87] (https://ida.loni.usc.edu/home/projectPage.jsp?project=ABVIB), ADNI[58] (https://ida.loni.usc.edu/home/projectPage.jsp?project=ADNI), AIBL[88] (https://ida.loni.usc.edu/home/projectPage.jsp?project=AIBL), HBN[55] (https://

data.healthybrainnetwork.org/main.php), HCP[56] (http://www.humanconnectomeproject.org/data), ICBM[89] (https://ida.loni.usc.edu/home/projectPage.jsp?project=ICBM), OASIS3[57] (https://sites.wustl.edu/oasisbrains/), SLIM[90] (https://fcon_1000.projects.nitrc.org/indi/retro/southwestuni_qiu_index.html) and MR-ART[63] (https://openneuro.org/datasets/ds004173/versions/1.0.2). Source data are provided with this paper.

Code availability

The source codes and trained models are available via GitHub at https://github.com/DBC-Lab/Brain_MRI_Enhancement.git. The network was trained using the Caffe deep learning framework (Caffe 1.0.0-rc3). The deployment was implemented with custom Python code (Python 2.7.17). The source codes of competing methods are available for DUNCAN (version 3.0) via Zenodo at https://doi.org/10.5281/zenodo.3742351 (ref. 38); Pix2Pix/CycleGan via GitHub at https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix; DU-Net at https://liwang.web.unc.edu/wp-content/uploads/sites/11006/2020/04/Anatomy_Guided_Densely_Connected_U_Net.txt; NLUP (version 2.0) at https://personales.upv.es/jmanjon/upsampling.htm; FAST at https://fsl.fmrib.ox.ac.uk/fsl/docs/#/structural/fast; a multiresolution non-local means filter (version 1.0) at https://personales.upv.es/jmanjon/res_denoising_NLM3D.htm; and Demon's algorithm at https://simpleitk.readthedocs.io/en/master/link_DemonsRegistration2_docs.html. The image pre-processing steps, including skull stripping and cerebellum removal, were performed by using a public cerebrum-dedicated pipeline (iBEAT V2.0, http://www.ibeat.cloud). The motion simulation tool is available via GitHub at https://github.com/Yonsei-MILab/MRI-Motion-Artifact-Simulation-Tool. The artefact simulator is available at https://ieeexplore.ieee.org/abstract/document/8759167. To quantitatively assess the significance of the results, we conducted statistical analyses using two-sided t-tests to obtain P values. Cohen's d was used as a key metric to quantify the magnitude of the observed effect, with calculations performed using the effect size calculators at https://lbecker.uccs.edu.

References

1. Frisoni, G., Fox, N., Jack, C., Scheltens, P. & Thompson, P. The clinical use of structural MRI in Alzheimer's disease. *Nat. Rev. Neurol.* **6**, 67–77 (2010).
2. Copeland, A. et al. Infant and child MRI: a review of scanning procedures. *Front. Neurosci.* **15**, 666020 (2021).
3. Thieba, C. et al. Factors associated with successful MRI scanning in unsedated young children. *Front. Pediatr.* **6**, 146 (2018).
4. Khan, J. et al. A program to decrease the need for pediatric sedation for CT and MRI. *Appl. Radiol.* **36**, 30–33 (2007).
5. Li, G. et al. Mapping longitudinal development of local cortical gyrification in infants from birth to 2 years of age. *J. Neurosci.* **34**, 4228–4238 (2014).
6. Havsteen, I. et al. Are movement artifacts in magnetic resonance imaging a real problem?—A narrative review. *Front. Neurol.* **8**, 232 (2017).
7. Zaitsev, M., Maclaren, J. & Herbst, M. Motion artefacts in MRI: a complex problem with many partial solutions. *J. Magn. Reson. Imaging* **42**, 887–901 (2015).
8. Gallichan, D., Marques, J. & Gruetter, R. Retrospective correction of involuntary microscopic head movement using highly accelerated fat image navigators (3D FatNavs) at 7T: 3D FatNavs for high-resolution retrospective motion correction. *Magn. Reson. Med.* **75**, 1030–1039 (2015).
9. Liu, S., Thung, K., Qu, L., Lin, W. & Yap, P.-T. Learning MRI artefact removal with unpaired data. *Nat. Mach. Intell.* **3**, 60–67 (2021).
10. Sommer, K. et al. Correction of motion artifacts using a multiscale fully convolutional neural network. *Am. J. Neuroradiol.* **41**, 416–423 (2020).
11. Cordero-Grande, L. et al. Motion-corrected MRI with DISORDER: distributed and incoherent sample orders for reconstruction deblurring using encoding redundancy. *Magn. Reson. Med.* **84**, 713–726 (2020).
12. Kecskemeti, S. et al. Robust motion correction strategy for structural MRI in unsedated children demonstrated with three-dimensional radial MPnRAGE. *Radiology* **289**, 509–516 (2018).
13. Duffy, B. et al. Retrospective motion artifact correction of structural MRI images using deep learning improves the quality of cortical surface reconstructions. *NeuroImage* **230**, 117756 (2021).
14. Tisdall, M. et al. Volumetric navigators for prospective motion correction and selective reacquisition in neuroanatomical MRI. *Magn. Reson. Med.* **68**, 389–399 (2011).
15. Stucht, D. et al. Highest resolution in vivo human brain MRI using prospective motion correction. *PLoS ONE* **10**, e0133921 (2015).
16. Pipe, J. Motion correction with PROPELLER MRI: application to head motion and free-breathing cardiac imaging. *Magn. Reson. Med.* **42**, 963–969 (1999).
17. Korin, H., Felmlee, J., Riederer, S. & Ehman, R. Spatial-frequency-tuned markers and adaptive correction for rotational motion. *Magn. Reson. Med.* **33**, 663–669 (1995).
18. Medley, M., Yan, H. & Rosenfeld, D. An improved algorithm for 2-D translational motion artifact correction. *IEEE Trans. Med. Imaging* **10**, 548–553 (1992).
19. Atkinson, D., Hill, D., Stoyle, P., Summers, P. & Keevil, S. Automatic correction of motion artifacts in magnetic resonance images using an entropy focus criterion. *IEEE Trans. Med. Imaging* **16**, 903–910 (1997).
20. Haskell, M., Cauley, S. & Wald, L. TArgeted Motion Estimation and Reduction (TAMER): data consistency based motion mitigation for MRI using a reduced model joint optimization. *IEEE Trans. Med. Imaging* **37**, 1253–1265 (2018).
21. Cordero-Grande, L., Hughes, E., Hutter, J., Price, A. & Hajnal, J. Three-dimensional motion corrected sensitivity encoding reconstruction for multi-shot multi-slice MRI: application to neonatal brain imaging. *Magn. Reson. Med.* **79**, 1365–1376 (2017).
22. Haskell, M. et al. Network Accelerated Motion Estimation and Reduction (NAMER): convolutional neural network guided retrospective motion correction using a separable motion model. *Magn. Reson. Med.* **82**, 1452–1461 (2019).
23. Jin, K., Mccann, M., Froustey, E. & Unser, M. Deep convolutional neural network for inverse problems in imaging. *IEEE Trans. Image Process.* **26**, 4509–4522 (2016).
24. Ahishakiye, E., Van Gijzen, M. B., Tumwiine, J., Wario, R. & Obungoloch, J. A survey on deep learning in medical image reconstruction. *Intell. Med.* **1**, 118–127 (2021).
25. Ravishankar, S., Ye, J. C. & Fessler, J. Image reconstruction: from sparsity to data-adaptive methods and machine learning. *Proc. IEEE* **108**, 86–109 (2019).
26. Lee, J., Kim, B. & Park, H. MC^2-Net: motion correction network for multi-contrast brain MRI. *Magn. Reson. Med.* **86**, 1077–1092 (2021).
27. Polak, D. et al. Motion guidance lines for robust data consistency-based retrospective motion correction in 2D and 3D MRI. *Magn. Reson. Med.* **89**, 1777–1790 (2023).
28. Wang, G., Shi, H., Chen, Y. & Wu, B. Unsupervised image-to-image translation via long-short cycle-consistent adversarial networks. *Appl. Intell.* **53**, 17243–17259 (2022).
29. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition* 5967–5976 (IEEE Computer Society, 2017); https://doi.org/10.1109/CVPR.2017.632
30. Manjon, J. et al. Non-local MRI upsampling. *Med. Image Anal.* **14**, 784–792 (2010).

31. Iglesias, J. et al. SynthSR: a public AI tool to turn heterogeneous clinical brain scans into high-resolution T1-weighted images for 3D morphometry. *Sci. Adv.* **9**, eadd3607 (2023).
32. Pham, C.-H. et al. Multiscale brain MRI super-resolution using deep 3D convolutional networks. *Comput. Med. Imaging Graph.* **77**, 101647 (2019).
33. Mohan, J., Krishnaveni, V. & Guo, Y. A survey on the magnetic resonance image denoising methods. *Biomed. Signal Process. Control* **9**, 56–69 (2014).
34. Liu, M. et al. Style transfer using generative adversarial networks for multi-site MRI harmonization. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* (eds de Bruijne, M. et al.) 313–322 (Springer, 2021); https://doi.org/10.1007/978-3-030-87199-4_30
35. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2006).
36. Kemenczky, P. et al. Effect of head motion-induced artefacts on the reliability of deep learning-based whole-brain segmentation. *Sci. Rep.* **12**, 1618 (2022).
37. Wang, L. et al. Volume-based analysis of 6-month-old infant brain MRI for autism biomarker identification and early diagnosis. *Med. Image Comput. Comput. Assist. Interv.* https://doi.org/10.1007/978-3-030-00931-1_47 (2018).
38. Liu, S. et al. eCode used in article "Learning MRI artefact removal with unpaired data". *Zenodo* https://doi.org/10.5281/zenodo.3742351 (2020).
39. Howell, B. et al. The UNC/UMN Baby Connectome Project (BCP): an overview of the study design and protocol development. *NeuroImage* **185**, 891–905 (2018).
40. Coupé, P., Manjon, J., Robles, M. & Collins, L. Adaptive multiresolution non-local means filter for three-dimensional magnetic resonance image denoising. *Image Process. IET* **6**, 558–568 (2012).
41. Wang, Z., Bovik, A., Sheikh, H. R. & Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612 (2014).
42. Wang, Z., Simoncelli, E. P. & Bovik, A. C. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers* 1398–1402 (IEEE, 2003); https://doi.org/10.1109/ACSSC.2003.1292216
43. Wang, Z. & Bovik, A. A universal image quality index. *IEEE Signal Process Lett.* **9**, 81–84 (2002).
44. Sheikh, H. R. & Bovik, A. C. Image information and visual quality. *IEEE Trans. Image Process.* **15**, 430–444 (2006).
45. Mason, A. et al. Comparison of objective image quality metrics to expert radiologists' scoring of diagnostic quality of MR images. *IEEE Trans. Med. Imaging* **39**, 1064–1072 (2020).
46. Duffy, B. A. et al. Retrospective correction of motion artifact affected structural MRI images using deep learning of simulated motion. In *Medical Imaging with Deep Learning* (2018); https://openreview.net/forum?id=H1hWfZnjM
47. Wang, L. et al. iBEAT V2.0: a multisite-applicable, deep learning-based pipeline for infant cerebral cortical surface reconstruction. *Nat. Protoc.* **18**, 1488–1509 (2023).
48. Wei, D. et al. Structural and functional brain scans from the cross-sectional Southwest University adult lifespan dataset. *Sci. Data* **5**, 180134 (2018).
49. Liu, S. et al. Chinese Color Nest Project: an accelerated longitudinal brain–mind cohort. *Dev. Cogn. Neurosci.* **52**, 101020 (2021).
50. Gao, P. et al. A Chinese multi-modal neuroimaging data release for increasing diversity of human brain mapping. *Sci. Data* **9**, 286 (2022).
51. Park, D. C. & Festini S. B. in *Cognitive Neuroscience of Aging: Linking Cognitive and Cerebral Aging* 363–388 (Oxford Univ. Press, 2016); https://doi.org/10.1093/acprof:oso/9780199372935.003.0015
52. Hughes, E. et al. A dedicated neonatal brain imaging system. *Magn. Reson. Med.* **78**, 794–804 (2017).
53. Cordero-Grande, L. et al. Sensitivity encoding for aligned multishot magnetic resonance reconstruction. *IEEE Trans. Comput. Imaging* **2**, 266–280 (2016).
54. Payakachat, N., Tilford, J. M. & Ungar, W. National Database for Autism Research (NDAR): big data opportunities for health services research and health technology assessment. *PharmacoEconomics* **34**, 127–138 (2015).
55. Alexander, L. et al. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Sci. Data* **4**, 170181 (2017).
56. Elam, J. S. et al. The Human Connectome Project: a retrospective. *NeuroImage* **244**, 118543 (2021).
57. LaMontagne, P. J. et al. OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. Preprint at *medRxiv* https://doi.org/10.1101/2019.12.13.19014902 (2019).
58. Weiner, M. W. et al. The Alzheimer's Disease Neuroimaging Initiative 3: continued innovation for clinical trial improvement. *Alzheimer's Dement.* **13**, 561–571 (2017).
59. Wang, L. et al. Benchmark on automatic six-month-old infant brain segmentation algorithms: the iSeg-2017 Challenge. *IEEE Trans. Med. Imaging* **38**, 2219–2230 (2019).
60. Lee, S., Jung, S., Jung, K.-J. & Kim, D.-H. Deep learning in MR motion correction: a brief review and a new motion simulation tool (view2Dmotion). *Invest. Magn. Reson. Imaging* **24**, 196 (2020).
61. Coupé, P. et al. Robust Rician noise estimation for MR images. *Med. Image Anal.* **14**, 483–493 (2010).
62. Nowak, R. D. Wavelet-based Rician noise removal for magnetic resonance imaging. *IEEE Trans. Image Process.* **8**, 1408–1419 (1999).
63. Nárai, Á. et al. Movement-related artefacts (MR-ART) dataset of matched motion-corrupted and clean structural MRI brain scans. *Sci. Data* **9**, 630 (2022).
64. Plenge, E. et al. Super-resolution methods in MRI: can they improve the trade-off between resolution, signal-to-noise ratio, and acquisition time? *Magn. Reson. Med.* **68**, 1983–1993 (2012).
65. Wang, J., Chen, Y., Wu, Y., Shi, J. & Gee, J. Enhanced generative adversarial network for 3D brain MRI super-resolution. In *2020 IEEE Winter Conference on Applications of Computer Vision* 3616–3625 (IEEE, 2020); https://doi.org/10.1109/WACV45572.2020.9093603
66. Tardif, C. et al. Open Science CBS Neuroimaging Repository: sharing ultra-high-field MR images of the brain. *NeuroImage* **124**, 1143–1148 (2015).
67. Zhang, Y. Y., Brady, M. & Smith, S. A. Segmentation of brain MR images through a hidden Markov random field model and the expectation maximization algorithm. *IEEE Trans. Med. Imaging* **20**, 45–57 (2001).
68. Styner, M. et al. 3D segmentation in the clinic: A Grand Challenge II: MS lesion segmentation. *MIDAS J.* https://doi.org/10.54294/lmkqvm (2007).
69. Sayah, A. et al. Enhancing the REMBRANDT MRI collection with expert segmentation labels and quantitative radiomic features. *Sci. Data* **9**, 338 (2022).
70. Menze, B. H. et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans. Med. Imaging* **34**, 1993–2024 (2015).

71. Roca, V. et al. A three-dimensional deep learning model for inter-site harmonization of structural MR images of the brain: extensive validation with a multicenter dataset. *Heliyon* **9**, e22647 (2023).
72. Thirion, J.-P. Image matching as a diffusion process: an analogy with Maxwell's demons. *Med. Image Anal.* **2**, 243–260 (1998).
73. Deoni, S. et al. Accessible pediatric neuroimaging using a low field strength MRI scanner. *NeuroImage* **238**, 118273 (2021).
74. Billot, B. et al. SynthSeg: segmentation of brain MRI scans of any contrast and resolution without retraining. *Med. Image Anal.* **86**, 102789 (2023).
75. Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (eds Navab, N., Hornegger, J., Wells, W. & Frangi, A.) 234–241 (Springer, 2015); https://doi.org/10.1007/978-3-319-24574-4_28
76. Huang, G., Liu, Z., Maaten, L. V. D. & Weinberger K. Q. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition* 2261–2269 (IEEE Computer Society, 2017); https://doi.org/10.1109/CVPR.2017.243
77. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2020).
78. Vaswani, A. et al. Attention is all you need. In *Proc. 31st International Conference on Neural Information Processing Systems* 6000–6010 (Curran Associates, 2017); https://api.semanticscholar.org/CorpusID:13756489
79. Zhang, Q. et al. Frnet: flattened residual network for infant MRI skull stripping. In *2019 IEEE 16th International Symposium on Biomedical Imaging* 999–1002 (IEEE, 2019); https://doi.org/10.1109/ISBI.2019.8759167
80. Jia, Y. et al. Caffe: convolutional architecture for fast feature embedding. In *Proc. 22nd ACM International Conference on Multimedia* 675–678 (Association for Computing Machinery, 2014); https://doi.org/10.1145/2647868.2654889
81. Nie, D. et al. 3-D fully convolutional networks for multimodal isointense infant brain image segmentation. *IEEE Trans. Cybern.* **49**, 1123–1136 (2018).
82. Makropoulos, A., Counsell, S. & Rueckert, D. A review on automatic fetal and neonatal brain MRI segmentation. *NeuroImage* **170**, 231–248 (2017).
83. Wang, L. et al. LINKS: learning-based multi-source integration framework for segmentation of infant brain images. *NeuroImage* **108**, 160–172 (2014).
84. Li, G. et al. Computational neuroanatomy of baby brains: a review. *NeuroImage* **185**, 906–925 (2018).
85. Zhu, J. & Qiu, A. Chinese adult brain atlas with functional and white matter parcellation. *Sci. Data* **9**, 352–362 (2022).
86. Di Martino, A. et al. Enhancing studies of the connectome in autism using the Autism Brain Imaging Data Exchange II. *Sci. Data* **4**, 170010 (2017).
87. Rodriguez, F., Zheng, L. & Chui, H. C. Psychometric characteristics of cognitive reserve: how high education might improve certain cognitive abilities in aging. *Dement. Geriatr. Cogn. Disord.* **47**, 1–10 (2019).
88. Lai, M. et al. Relationship of established cardiovascular risk factors and peripheral biomarkers on cognitive function in adults at risk of cognitive deterioration. *J. Alzheimer's Dis.* **74**, 1–9 (2020).
89. JC, M. et al. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). *Philos. Trans. R. Soc. Lond. Ser. B* **356**, 1293–1322 (2001).
90. Wei, L. et al. Longitudinal test-retest neuroimaging data from healthy young adults in southwest China. *Sci. Data* **4**, 170017 (2017).
91. Sun, Y. et al. Multi-site infant brain segmentation algorithms: The iSeg-2019 Challenge. *IEEE Trans. Med. Imaging* **40**, 1363–1376 (2021).

Acknowledgements

Y.S., Limei Wang and Li Wang were supported by the National Institute of Mental Health under award numbers MH133845, MH117943, MH123202 and MH116225. G.L. was supported by the National Institutes of Health (NIH) under award numbers MH133845, MH117943, MH123202, MH116225, AG075582 and NS128534. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. This work also uses approaches developed by NIH grants (U01MH110274 and R01MH104324) and the efforts of the UNC/UMN Baby Connectome Project Consortium. We acknowledge M. M. Pangelinan for her valuable contribution in providing the in vivo low-resolution data used for super-resolution validation. We express our sincere gratitude to all those who have supported us in the validation: J. Bernal, J. Kim, K. A. Vaughn, J. Tuulari, K. Oishi, A. Tapp, Y. Chen, X. Geng, T. F. Vaz and Z. Zariry. We also deeply appreciate all participants who contributed to the datasets involved in this work.

Author contributions

Y.S. and Li Wang designed and implemented the pipeline. Y.S. and Li Wang carried out the application, performed the experiments and analysed the data. Y.S. and Li Wang performed result validation. Y.S. wrote the paper. Limei Wang, W.L., G.L. and Li Wang revised the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41551-024-01283-7.

Correspondence and requests for materials should be addressed to Li Wang.

Peer review information *Nature Biomedical Engineering* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Li Wang

Last updated by author(s): Oct 13, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	1. The source codes and trained models are available at https://github.com/DBC-Lab/Brain_MRI_Enhancement.git. The network was trained using the Caffe deep learning framework (Caffe 1.0.0-rc3). The deployment was implemented with the custom Python code (Python 2.7.17). 2. The source codes of competing methods are available: DUNCAN: https://zenodo.org/record/4474599#.ZD8DGOzML0o, version 3; Pix2Pix/CycleGan: https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix; DU-Net: https://liwang.web.unc.edu/wp-content/uploads/sites/11006/2020/04/Anatomy_Guided_Densely_Connected_U_Net.txt; NLUP: https://personales.upv.es/jmanjon/upsampling.htm, version 2.0 FAST: https://fsl.fmrib.ox.ac.uk/fsl/docs/#/structural/fast A multiresolution non-local means filter: https://personales.upv.es/jmanjon/res_denoising_NLM3D.htm, version 1.0 Demon’s algorithm: https://simpleitk.readthedocs.io/en/master/link_DemonsRegistration2_docs.html# 3. The image pre-processing steps, including skull stripping, cerebellum removal, were performed by using a public cerebrum-dedicated pipeline (iBEAT V2.0 Cloud, http://www.ibeat.cloud). The motion simulation tool: https://github.com/Yonsei-MILab/MRI-Motion-Artifact-Simulation-Tool The artifact simulator: https://ieeexplore.ieee.org/abstract/document/8759167 4. To generate degraded images, the simulated rotation/periodic/continuous/sudden motions were added by a released software (https://github.com/Yonsei-MILab/MRI-Motion-Artifact-Simulation-Tool). 5. Codes for evaluation metrics: MSE: https://www.mathworks.com/help/images/ref/immse.html

PSNR: https://www.mathworks.com/help/images/ref/psnr.html
SSIM: https://www.mathworks.com/help/images/ref/ssim.html
MS-SSIM: https://www.mathworks.com/help/images/ref/multissim3.html
UQI: https://viewer.mathworks.com/?viewer=plain_code&url=https%3A%2F%2Fin.mathworks.com%2Fmatlabcentral%2Fmlc-downloads%2Fdownloads%2Fsubmissions%2F29500%2Fversions%2F1%2Fcontents%2FimageQualityIndex.m&embed=web
TCT: https://github.com/DBC-Lab/Brain_MRI_Enhancement/blob/main/For_reference/TCT.py
VIF: https://github.com/sattarab/image-quality-tools/blob/master/metrix_mux/metrix/vif/vifvec_release/vifvec.m
6. To quantitatively assess the significance of the results, we conducted statistical analyses using two-sided t-tests to obtain P values. Cohen's d was used as a key metric to quantify the magnitude of the observed effect, with calculations performed using the Effect Size Calculators (https://lbecker.uccs.edu).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The raw data generated in this study are available from dHCP [51, 52] (https://biomedia.github.io/dHCP-release-notes), NDAR [53] (https://nda.nih.gov/edit_collection.html?id=19), BCP [38] (https://nda.nih.gov/edit_collection.html?id=2848), SALD [47] (http://fcon_1000.projects.nitrc.org/indi/retro/sald.html), CCNP [48, 49] (https://ccnp.scidb.cn/en/detail?dataSetId=826407529641672704&version=V3&code=o00133), DLBS [50] (https://fcon_1000.projects.nitrc.org/indi/retro/dlbs.html), IXI (http://brain-development.org/ixi-dataset/), Chinese Adult Brain [84] (https://www.nitrc.org/projects/adultatlas), ABIDE [85] (https://fcon_1000.projects.nitrc.org/indi/abide/), ABVIB [86] (https://ida.loni.usc.edu/home/projectPage.jsp?project=ABVIB), ADNI [57] (https://ida.loni.usc.edu/home/projectPage.jsp?project=ADNI), AIBL [87] (https://ida.loni.usc.edu/home/projectPage.jsp?project=AIBL), HBN [54] (https://data.healthybrainnetwork.org/main.php), HCP [55] (http://www.humanconnectomeproject.org/data), ICBM [88] (https://ida.loni.usc.edu/home/projectPage.jsp?project=ICBM), OASIS3 [56] (https://sites.wustl.edu/oasisbrains/), SLIM [89] (https://fcon_1000.projects.nitrc.org/indi/retro/southwestuni_qiu_index.html), and MR-ART [62] (https://openneuro.org/datasets/ds004173/versions/1.0.2).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	We do not include sex or gender information in this study.
Reporting on race, ethnicity, or other socially relevant groupings	We do not include information on race, ethnicity or other socially relevant groupings in this study.
Population characteristics	In total, 11,479 participants (19 datasets, including all training and testing participants) covering a broad age range from foetuses to elderly individuals, were included in this study for training and testing, as shown in Supplementary Table 4. We do not involve sex information.
Recruitment	1. According to Edwards, A.D. et al. paper [Frontiers in neuroscience, 16: 886772, 2022], for the dHCP dataset, the participants were recruited at St Thomas' Hospital, London, and imaged at the Evelina Newborn Imaging Centre, Centre for the Developing Brain, King's College London, United Kingdom. The study population includes participants born at term equivalent age without any known pregnancy or neonatal problems and are regarded as healthy. 2. According to information related to the NDAR dataset (https://depts.washington.edu/uwautism/research-projects/infantbrain-imaging-study/), participants were recruited from families with (1) a child between the ages of 4 and 12, (2) an older sibling who is EITHER typically developing OR has an Autism Spectrum Disorder, and (3) who have previously participated in Infant Brain Imaging Study. 3. According to Howell, B.R. et al. paper [Neuroimage. 185: 891-905, 2019], for the BCP dataset, participants were recruited from existing registries at UNC and UMN based on state-wide birth records as well as from broader community resources (e.g., community centers and targeted day-care centers) to ensure the sample approximates the racial/ethnic and socio-economic diversity of the US census. To augment recruitment of the youngest cohort of participants, the investigators may recruit participants perinatally by approaching expectant and new mothers at "The Birthplace" at UMN and the UNC Hospitals Newborn Nursery. 4. According to Wei, D. et al. paper [Scientific data. 5: 180134, 2018], for the SALD dataset, all participants included in this study were selected from a large dataset of individuals who have participated in the ongoing BCM data collection initiative. The young adults (18-25 years old) of the lifespan sample enrolled as college students of Southwest University in Chongqing, China. Many of the mid adults (26 to 40 years old) were recruited directly from the staff of Southwest University. The rest of the adults sample were recruited from communities close to the university campus. We primarily recruited participants through leaflets, online advertisements, and face-to-face propaganda. 5. According to Gao, P. et al. paper [Scientific data. 9: 286, 2022], for the 3R-BRAIN dataset, all participants were recruited by advertisements posted in colleges and communities. Specifically, to protect personal information, we applied the face masking toolkit to achieve privacy anonymization by blurring facial information while all the participants agreed to share their anonymized data with the public. 6. According to Kristen M.K. et al. paper [Neuroimage, 104: 21-34, 2015], for the DLBS dataset, the participants were

recruited through media advertisements and flyers and underwent health history screening via a health questionnaire as well as telephone and personal interviews. All participants were screened against cardiovascular, neurological, and psychiatric disorders, head injury with loss of consciousness N10 min, and drug/alcohol abuse. Participants were native English speakers and strongly right-handed (on the Edinburgh Handedness Questionnaire (Oldfield, 1971). The participants were well-educated (mean 16.55 ± 2.68 years) and scored highly (28.35 ± 1.27) on the Mini Mental State Examination (MMSE; Folstein et al., 1975).
7. For the IXI dataset (https://brain-development.org/ixi-dataset/), this study collected images from three different hospitals in London.
8. According to Yan C. Z. et al. paper [Frontiers in systems neuroscience, 4: 1377, 2010], for the Chinese Adult Brain dataset, all participants are young healthy volunteers with head motion less than 2.0 mm displacement in any of the x, y, or z directions or 2.0° of any angular motion throughout the resting-state scan. All are righthanded and had no history of neurological and psychiatric disorders.
9. According to Di, M. A. et al. paper [Scientific data, 4(1): 1-15, 2017], for the ABIDE dataset, this study followed independent data collections across multiple sites. They can range from inclusion/exclusion criteria, recruitment/sampling strategies, MRI scanner types, data acquisition parameters, and instructions (e.g., eyes open versus closed).
10. For the ABVIB dataset (https://adrc.usc.edu/scientists-researchers/), all participants were selected to represent (1) a spectrum of vascular risk, ranging from low to high, and (2) a spectrum of cognitive impairment, ranging from no to mild cognitive impairment. Neuropsychological testing was performed every year, and MRI scans were done every 2 years.
11. According to R.C. Petersen et al. paper [Neurology, 74(3): 201-209, 2010], for the ADNI dataset, all participants had to have a Hachinski Ischemic Score of less than or equal to 4; permitted medications stable for 4 weeks prior to screening; a Geriatric Depression Scale score of less than 6; a study partner with 10 hours per week of contact either in person or on the telephone and who could accompany the participant to the clinical visits; visual and auditory acuity adequate for neuropsychological testing; good general health with no diseases precluding enrollment; 6 grades of education or work history equivalent; and ability to speak English or Spanish fluently. Women had to be sterile or 2 years past childbearing potential. Participants had to be able to complete a 3-year imaging study (2 years for participants with AD). Participants agreed to DNA extraction for APOE testing and banking and agreed to blood and urine examination for biomarkers.
12. According to Ellis K.A. et al. paper [International psychogeriatrics, 21(4): 672-687, 2009], for the AIBL dataset, a total of 1166 volunteers were recruited. Participants with AD (211) had neuropsychological profiles which were consistent with AD and were more impaired than participants with mild cognitive impairment (133) or healthy controls (768), who performed within expected norms for age on neuropsychological testing. PiB PET scans were performed on 287 participants, 100 had DEXA scans and 91 participated in ActiGraph monitoring.
13. According to Ellis K.A. et al. paper [Scientific data, 4: 170181, 2017], for the HBN dataset, this study adopted a community-referred recruitment model using advertisements to encourage the participation of families who have concerns about psychiatric symptoms in their children. With few exceptions, the presence of psychiatric, medical, or neurological illness does not exclude participation. All individuals meeting inclusion criteria, without any reasons for exclusion, are invited to participate in the study.
14. According to Alexander, L. et al. paper [Neuroimage, 62(4): 2222-2231, 2012], for the HCP dataset, the participants were young adult sibships, with most of these sibships including a MZ or DZ twin pair, aged between 22 and 35 years old.
15. According to Maude, S. et al. paper [American Journal of Psychiatry, 171(6): 627-639, 2014], for the ICBM dataset, the participants were recruited across 15 sites. They were included if they had a genetically confirmed 22q11.2 microdeletion, were between the ages of 6 and 70 years, and underwent a comprehensive structured psychiatric assessment using a validated instrument.
16. According to Pamela J LaMontagne et al. paper [MedRxiv, 2019.12. 13.19014902], for the OASIS3 dataset, the following cohorts were recruited: (1) Individuals who were generally healthy, were cognitively normal (CDR = 0), and had a family history of AD, defined as being a biologic child of at least one parent with a reported history of AD dementia with onset age 80 years; (2) individuals who were generally healthy, were cognitively normal (CDR = 0), and had no family history of AD for either biological parent and lived at least to age 70 years; and (3) healthy individuals 65 and older, both those who were cognitively normal (CDR 0) and this with very mild-mild symptomatic AD (CDR 0.5 and 1). Participants were recruited from the community via flyers, word of mouth, and community engagements. Participants from all cohorts agreed to submit an initial blood sample for genetic testing and complete regular cognitive testing, neuroimaging, and lumbar punctures approximately every 2-3 years. Each participant was enrolled along with a collateral source, someone who knew the participants well (e.g., the spouse of the adult child) and could report whether the participant's current cognitive and functional performance was or was not at previously attained levels.
17. For the SLIM dataset (http://fcon_1000.projects.nitrc.org/indi/retro/southwestuni_qiu_index.html), a total of 167 healthy undergraduate students from the local community of Southwest University in China participated in this study as a part of our ongoing project investigating the associations among brain imaging, mental health, and creativity. All participants were right-handed and had no history of neurological or psychiatric diseases.
18. According to Nárai et al. [10.1038/s41597-022-01694-8], the brain scans have been acquired between 2019 and 2021, and the data collection included 148 healthy adult volunteers (95 female; age range: 18–75 years; median age: 25.16 years; IQR: 10.50 years) with no reported history of neurological or psychiatric diseases. All participants provided written, informed consent before participation.

Ethics oversight

1. According to Edwards, A. D. et al. paper [Frontiers in neuroscience, 16: 886772, 2022], for the dHCP dataset, the study was approved by the United Kingdom Health Research Authority (Research Ethics Committee reference number: 14/LO/1169) and written parental consent was obtained in every case for imaging and open data release of the anonymized data.
2. According to Hazlett, H. et al. paper [Nature. 542: 348-351, 2017], for the NDAR dataset, data collection sites had studyprotocols approval from their Institutional Review Boards (IRB), and all enrolled participants had informed consent provided byparent/guardian.
3. According to Howell, B. R. et al. paper [Neuroimage. 185: 891-905, 2019], for the BCP dataset, all procedures were approved by the University of North Carolina at Chapel Hill and the University of Minnesota Institutional Review Boards.
4. According to Wei, D. et al. paper [Scientific data. 5: 180134, 2018], for the SALD dataset, the dataset collection was approved by the Research Ethics Committee of the Brain Imaging Center of Southwest University.
5. According to Gao, P. et al. paper [Scientific data. 9: 286, 2022], for the 3R-BRAIN dataset, all data were approved by the Beijing Municipal Science and Tech Commission (Grant No.Z161100002616023f).
6. According to Kristen, M. K. et al. paper [Neuroimage, 104: 21-34, 2015], for the DLBS dataset, all participants provided

written informed consent and were debriefed in accord with university human investigations committee guidelines.
7. For the IXI dataset (https://brain-development.org/ixi-dataset/), the data are approved by the European Research Council (ERC), Engineering and Physical Sciences Research Council (EPSRC), Medical Research Council (MRC), and Action Medical Research.
8. According to Yan, C. Z. et al. paper [Frontiers in systems neuroscience, 4: 1377, 2010], for the Chinese Adult Brain dataset, written informed consent was obtained from each participant, and the study was approved by the Institutional Review Board of Beijing Normal University Imaging Center for Brain Research.
9. According to Di, M. A. et al. paper [Scientific data, 4(1): 1-15, 2017], for the ABIDE dataset, all data are required to confirm that their local Institutional Review Board (IRB) or ethics committee has approved both the initial data collection and the retrospective sharing of a fully de-identified version of the datasets (i.e., after the removal of the 18 protected health information identifiers including facial information from structural images as identified by the HealthInsurance Portable and Accountability Act [HIPAA]).
10. For ABVIB dataset (https://ida.loni.usc.edu/login.jsp), the project represented a collaborative effort among investigators at the University of Southern California, University of California at Davis, University of California at San Francisco, University of California at Los Angeles, and University of California at Berkeley.
11. According to Petersen, R. C. et al. paper [Neurology, 74(3): 201-209, 2010], for ADNI dataset, this study was approved by the Institutional Review Boards of all of the participating institutions. Informed written consent was obtained from all participants at each site.
12. According to Ellis, K. A. et al. paper [International psychogeriatrics, 21(4): 672-687, 2009], for AIBL dataset, the study was approved by the institutional ethics committees of Austin Health, St Vincent's Health, Hollywood Private Hospital and Edith Cowan University, and all volunteers gave written informed consent before participating in the study.
13. According to Ellis K.A. et al. paper [Scientific data, 4: 170181, 2017], for the HBN dataset, the study was approved by the Chesapeake Institutional Review Board (https://www.chesapeakeirb.com/). Written informed consent is obtained from participants ages 18 or older. For participants younger than 18, written consent is obtained from their legal guardians and the participant.
14. According to Van Essen, D. C. et al. paper [Neuroimage, 62(4): 2222-2231, 2012], for the HCP dataset, all participant recruitment procedures and informed consent forms, including consent to share de-identified data, were approved by the Washington University institutional review board.
15. According to Maude, S. et al. paper [American Journal of Psychiatry, 171(6): 627-639, 2014], for the ICBM dataset, each site received approval from its local ethics committee (institutional review board). Each participant and his or her legal caregiver, where appropriate, gave written informed consent for study participation.
16. According to Pamela, J. L. et al. paper [MedRxiv, 2019.12. 13.19014902], for OASIS3 dataset, all participants were consented into Knight ADRC-related projects following procedures approved by the Institutional Review Board of Washington University School of Medicine.
17. For the SLIM dataset (http://fcon_1000.projects.nitrc.org/indi/retro/southwestuni_qiu_index.html), the experimental protocol was approved by the Institutional Review Board of the Southwest University Brain Imaging Center. All participants provided informed consent prior to engaging in the experiment and received payment for their participation.
18. According to Nárai et al. [10.1038/s41597-022-01694-8], the research protocol used for collecting the dataset was designed and conducted in accordance with the Hungarian regulations and laws, and was approved by the National Institute of Pharmacy and Nutrition (file number: OGYÉI/70184/2017).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

There is currently no established theoretical guidance for estimating sample size in deep learning models. As such, our selection was informed by practical considerations and the availability of data for training and validation.
1. For training, we used 52 foetal participants (21-36 GW) from our in-house collection and 464 participants (0-6 years old) from UMN/UNC Baby Connectome Project (BCP) as our training participants In this study, to effectively correct artifacts for lifespan brain MRIs, we trained enhancement models at different developmental stages, including fetal phase (52 participants), 0 month (24 participants), 3 months (40 participants), 6 months (51 participants), 9 months (80 participants), 12 months (73 participants), 18 months (102 participants), and 24+ months (94 participants) of age.
2. For testing, we validated on 19 testing datasets (10,963 in vivo images and 2,448 synthesized corrupted images) acquired with various scanners and imaging protocols/parameters, including the dHCP (395 participants), NDAR (600 participants), BCP (20 participants), SALD (494 participants), CCNP (181 participants), DLBS (305 participants), IXI (539 participants), Chinese Adult Brain (180 participants), ABIDE (685 participants), ABVIB (228 participants), ADNI (427 participants), AIBL (1,293 participants), HBN (1,091 participants), HCP (459 participants), ICBM (345 participants), OASIS (2,662 participants), SLIM (497 participants), MR-ART (420 participants) and T2-weighted foetal images (142 participants, from in-house collection). More imaging information about these datasets is listed in Supplementary Table 4.

Data exclusions

1. According to Edwards, A.D. et al. paper [Frontiers in neuroscience, 16: 886772, 2022], for the dHCP dataset, the principal exclusion criteria are: (1) Mothers or infants with contraindication to MR imaging, including implanted metallic devices such as orthopedic devices or non-compatible clips to close patent ductus areteriosus. (2) Preterm infants who are too unwell to tolerate the scanning period, despite full intensive neonatal care. This decision would be made by the attending paediatric doctor performing the scan, in conjunction with the medical team caring for the patient. (3) Language difficulties preventing proper communication about the trial and the consent process.

2. According to Hazlett, H. et al. paper [Nature. 542: 348-351, 2017], for the NDAR dataset, the exclusion criteria for both groups included the following: (1) diagnosis or physical signs strongly suggestive of a genetic condition or syndrome (for example, fragile X syndrome) reported to be associated with ASDs, (2) a significant medical or neurological condition affecting growth, development or cognition (for example, CNS infection, seizure disorder, congenital heart disease), (3) sensory impairment such as vision or hearing loss, (4) low birth weight (< 2,000 g) or prematurity (< 36 weeks gestation), (5) possible perinatal brain injury from exposure to in utero exogenous compounds reported to likely affect the brain adversely in at least some individuals (for example, alcohol, selected prescription medications), (6) non-English speaking families, (7) contraindication for MRI (for example, metal implants), (8) adopted participants, and (9) a family history of intellectual disability, psychosis, schizophrenia or bipolar disorder in a first-degree relative. The sample for this analysis included all children with longitudinal imaging data processed until 31 August 2015. During our data analysis, we have performed tissue segmentation on all the archived participants from BCP and NDAR. To chart accurate growth trajectory, if any of two raters considered the segmentation quality for one scan as "poor", this scan will be excluded. Finally, we excluded 34 of 493 scans (7%) from BCP and NDAR for charting the growth trajectory.
3. According to Howell, B.R. et al. paper [Neuroimage. 185: 891-905, 2019], for the BCP dataset, participants are excluded if they were born prior to 37 weeks gestation, had a birth weight lower than 2,000 grams, or if they had any major delivery complications. Major delivery complications may include neonatal hypoxia or neonatal illness requiring a greater than two day NICU stay. They are also excluded if they: (1) are adopted, (2) have a first degree relative with autism, intellectual disability, schizophrenia, or bipolar disorder, (3) have any significant medical and/or genetic conditions affecting growth, development, or cognition, or (4) have any contraindication to MRI. Additional exclusion criteria include major pre- and/or perinatal issues including: (1) maternal pre-eclampsia, placental abruption, maternal HIV status, and maternal alcohol or illicit drug use during pregnancy. Finally, children are excluded from the study if their caregivers are unable to communicate in English at a level to provide informed consent.
4. According to Wei, D. et al. paper [Scientific data. 5: 180134, 2018], for the SALD dataset, the exclusion criteria included: (1) MRI related exclusion criteria, which included claustrophobia, metallic implants, Meniere's Syndrome and a history of fainting within the previous 6 months; (2) current psychiatric disorders and neurological disorders; (3) use of psychiatric drugs within the three months prior to scanning; (4) pregnancy; or (5) a history of head trauma. Informed written consent was obtained from each participant. Besides, we required participants to refrain from drinking during the day before the scanning and the scanning day.
5. According to Gao, P. et al. paper [Scientific data. 9: 286, 2022], for the 3R-BRAIN dataset, the participants were excluded if they met any of the following criteria: regular smoker, pregnancy, abnormal color discrimination, a history of alcohol or drug abuse, currently any medication, MRI contraindications, neuropsychiatric or severe somatic disorder and sedative-hypnotic medication within a month or any medication for major neuropsychiatric disorders.
6. According to Kristen, M.K. et al. paper [Neuroimage, 104: 21-34, 2015], for the DLBS dataset, this study did not disclose specific exclusion criteria.
7. For the IXI dataset (https://brain-development.org/ixi-dataset/), this study did not disclose specific exclusion criteria.
8. According to Yan, C. Z. et al. paper [Frontiers in systems neuroscience, 4: 1377, 2010], for the Chinese Adult Brain dataset, the participants with excessive head motion need to be excluded from urther analysis.
9. According to Di, M. A. et al. paper [Scientific data, 4(1): 1-15, 2017], for the ABIDE dataset, the specific inclusion/exclusion criteria used for selecting controls (e.g., IQ range, first degree relative with ASD) or ASD (e.g., absence of reported seizure and genetic syndromes) varied across collections. Each collection narrative on the ABIDE II website provides details regarding these criteria.
10. For ABVIB dataset, this study did not disclose specific exclusion criteria.
11. According to Petersen, R.C. et al. paper [Neurology, 74(3): 201-209, 2010], for ADNI dataset, all participants could not have any medical contraindications to MRI and could not be enrolled in other trials or studies concurrently.
12. According to Ellis, K.A. et al. paper [International psychogeriatrics, 21(4): 672-687, 2009], for the AIBL dataset, 54 participants were excluded from further study due to comorbid disorders which could affect cognition or because of withdrawal of consent.
13. According to Ellis K.A. et al. paper [Scientific data, 4: 170181, 2017], for the HBN dataset, the primary causes for exclusion center on the presence of acute safety concerns (e.g., danger to self or others), cognitive or behavioral impairments that could interfere with participation (e.g., being non-verbal, IQ less than 66), or medical concerns that are expected to confound brain-related findings.
14. According to Van Essen, D. C. et al. paper [Neuroimage, 62(4): 2222-2231, 2012], for the HCP dataset, the exclusion criteria are: (1) exclude sibships with individuals having severe neurodevelopmental disorders (e.g. autism), documented neuropsychiatric disorders (e.g. schizophrenia or severe recurrent depression) or neurologic disorders (e.g. Parkinson's disease), but will include individuals who are smokers, are overweight, or have a history of heavy drinking or recreational drug use without having experienced severe symptoms (Supplemental Table S1 lists the full set of inclusion and exclusion criteria under consideration); (2) Twins born prior to 34 weeks gestation and non-twins born prior to 37 weeks gestation will be excluded; (3) conduct an additional extensive, reliable, and valid psychiatric interview, the Semi-Structured Assessment for the Genetics of Alcoholism (SSAGA, Bucholz et al., 1994), to confirm the absence of significant psychiatric illness.
15. According to Maude, S. et al. paper [American Journal of Psychiatry, 171(6): 627-639, 2014], for the ICBM dataset, the exclusion criteria were applied: normal or inconclusive genetic testing, missing demographic information, age less than 6 years, nonstructured clinical psychiatric assessment, assessment for a single disorder (e.g., autism spectrum disorder), and assessment with self- or parent-report questionnaire.
16. According to Pamela, J. M. et al. paper [MedRxiv, 2019.12. 13.19014902], for the OASIS3 dataset, the exclusion criterion included medical conditions that precluded longitudinal participation (e.g., end-stage renal disease requiring dialysis) or medical contraindications for the study arms (e.g., pacemaker for MRI, anticoagulant use for lumbar puncture).
17. For the SLIM dataset (http://fcon_1000.projects.nitrc.org/indi/retro/southwestuni_qiu_index.html), four participants were excluded because of problems with the imaging data (two participants with excessive artifacts) and incomplete behavioral data (2 participants). All participants were right-handed and had no history of neurological or psychiatric diseases.
18. According to Nárai et al. [10.1038/s41597-022-01694-8], the data collection included
148 healthy adult volunteers (95 female; age range: 18–75 years; median age: 25.16 years; IQR: 10.50 years) with no reported history of neurological or psychiatric diseases.

Replication

The code used for training the deep-learning models are made publicly available for the reproducibility purpose (https://github.com/DBC-Lab/Brain_MRI_Enhancement.git).

Randomization

The samples were allocated into experimental groups (training and testing) randomly.

Blinding

For all training and testing datasets in this study, the investigators were not blinded to the group allocation (only age related) since the proposed foundation model is designed to improve the medical image quality covering the lifespan from fetuses to elderly individuals (with varying intensity distribution related to age).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Magnetic resonance imaging

Experimental design

Design type

We do not engage in data collection; instead, we use previously collected data from other researchers to support our analysis.

Design specifications

We do not engage in data collection; instead, we use previously collected data from other researchers to support our analysis.

Behavioral performance measures

We do not engage in data collection; instead, we use previously collected data from other researchers to support our analysis.

Acquisition

Imaging type(s)

Structural MRI.

Field strength

1.5 Tesla, 3 Tesla and 7 Tesla.

Sequence & imaging parameters

1. We have reported the scanner, modality, TR/TE (ms), and resolution information of data from the 19 datasets in Supplementary Table 4.
2. For the 7T images (Siemens scanner), the TR/TE (ms) is 5000/2.45 for T1w images, with a voxel resolution of 0.4 x 0.4 x 0.4 mm^3.

Area of acquisition

A whole brain scan.

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software

The image preprocessing steps, including skull stripping, cerebellum removal were performed by using a public cerebrum-dedicated pipeline (iBEAT V2.0 Cloud, http://www.ibeat.cloud). To generate degraded images, the simulated rotation/periodic/continuous/sudden motions were added by a released software (https://github.com/Yonsei-MILab/MRI-Motion-Artifact-Simulation-Tool).

Normalization

For image preprocessing, the resolution of all images in Supplementary Table 4 was resampled into 0.8 x 0.8 x 0.8 mm^3, and the resolution of images in the subsection 'Application on reconstruction 7T-like images from 3T MRIs' was resampled into 0.4 x 0.4 x 0.4 mm^3.

Normalization template

1. For foetuses, we randomly selected a participant from the in-house collection as a template to perform histogram matching for remaining participants.
2. For each age group from newborn infants to elderly individuals, we randomly selected a participant from the BCP as a template to perform histogram matching for remaining participants.
Please refer to https://github.com/DBC-Lab/Brain_MRI_Enhancement/tree/main/Templates for all templates.

Noise and artifact removal

We do not engage in data collection; instead, we use previously collected data from other researchers to support our analysis.

Volume censoring

We do not engage in data collection; instead, we use previously collected data from other researchers to support our analysis.

Statistical modeling & inference

Model type and settings: No statistical modeling or reference was used in this study.

Effect(s) tested: No effects were tested in this study.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☒ Both

Anatomical location(s): We performed extraction of the cerebrum by leveraging an infant cerebrum-dedicated processing pipeline (http://www.ibeat.cloud).

Statistic type for inference (See Eklund et al. 2016): No inference was used in this study.

Correction: No correction was used in this study.

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☒	☐ Multivariate modeling or predictive analysis