UltimateSynth: MRI Physics for Pan-Contrast AI

Rhea Adams[1], Walter Zhao[1], Siyuan Hu[1], Wenjiao Lyu[2,3], Khoi Minh Huynh[2,3], Sahar Ahmad[2,3], Dan Ma[1,✉], & Pew-Thian Yap[2,3,✉]

[1]*Department of Biomedical Engineering, Case Western Reserve University, Cleveland, OH, USA*
[2]*Department of Radiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA*
[3]*Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA*

Magnetic resonance imaging (MRI) is commonly used in healthcare for its ability to generate diverse tissue contrasts without ionizing radiation. However, this flexibility complicates downstream analysis, as computational tools are often tailored to specific MRI types and lack generalizability across the full spectrum of scans used in healthcare. Here, we introduce a versatile framework for the development and validation of pan-contrast AI models that can exhaustively cater to the full spectrum of scans achievable with MRI, enabling model deployment across scanner models, scan types, and age groups. Core to our framework is UltimateSynth, a technology that combines tissue physiology and MR physics in synthesizing realistic images across a comprehensive range of contrasts to bolster the AI development life cycle through efficient data labeling, generalizable model training, and thorough performance benchmarking. UltimateSynth is a platform for pan-contrast generalization of contrast-specific tools. We showcase the effectiveness of UltimateSynth by training an off-the-shelf U-Net to generalize anatomical segmentation across over 150,000 unique MRI contrasts, achieving robust tissue volumetric quantification with exceptionally low variability below 2%.

MAGNETIC resonance imaging (MRI) offers a myriad of soft-tissue contrasts that are useful in physics, biology, and medicine. By altering the contrast, radiologists can enhance specific tissue properties, such as water content, fat, or blood flow, facilitating the identification of conditions like tumors, inflammation, bleeding, and other pathologies. Different MRI contrast types, such as T1-weighted, T2-weighted, and diffusion-weighted images, provide complementary information that aids comprehensive clinical assessment and diagnosis. However, undesirable contrast variations may also arise from differences in imaging equipment, acquisition sequences, or subjective preferences[1–4]. Contrast variations can cause inconsistent interpretation, diagnostic errors, and reduced reproducibility[5,6]. They also pose challenges in automated analysis—MRI computational tools that rely on specific contrast settings may struggle to extract true tissue properties when faced with varying contrast. In large-scale datasets, contrast variability complicates the integration and analysis of data from different sources, reducing their collective utility.

AI networks for MRI often lack generalizability beyond their training data[7,8]. This poses challenges in distinguishing actual tissue characteristics from contrast variations. Insufficient consideration of contrast deviations throughout the AI development life cycle—including data annotation, network learning, and model deployment—limits AI accuracy, reliability, and adaptability. Given the practical impossibility of gathering *in vivo* training images encompassing all conceivable contrasts, contemporary approaches to improve MR image generalization mainly revolve around synthesizing contrasts for data augmentation during network training. These

✉ Corresponding authors: Dan Ma (dxm302@case.edu) and Pew-Thian Yap (ptyap@med.unc.edu)

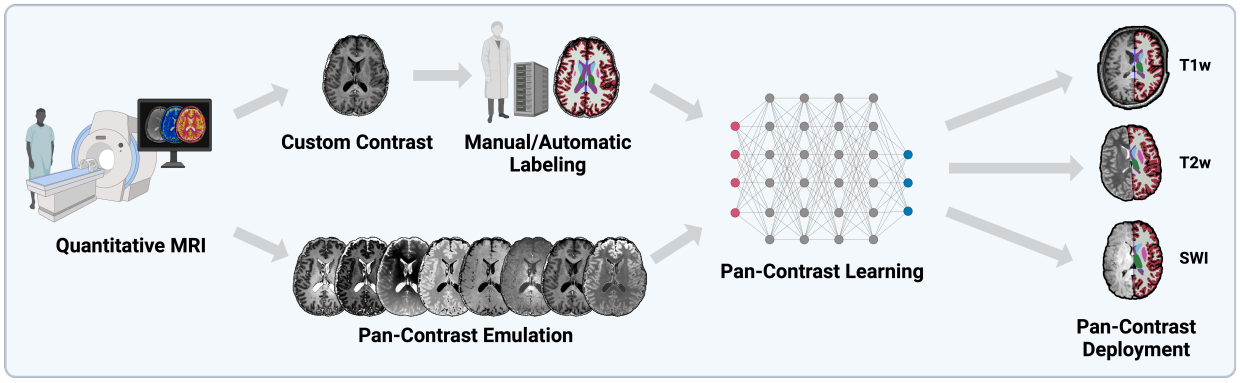

Fig. 1 | Pan-contrast AI from minimal labeling. UltimateSynth facilitates the learning of pan-contrast AI models by using data labeled only with as few as one contrast. It generates contrasts customized to software tools or annotator preferences, offering wide-ranging versatility in data labeling. UltimateSynth substantially reduces data acquisition and labeling efforts while boosting model output consistency by employing a diverse range of label-consistent contrasts during model training. In addition, UltimateSynth allows for the explicit validation of how well models handle the full range of MRI contrasts or any specific MRI contrast—a critical feature missing in existing MRI AI frameworks.

methods utilize either generative models[9–11] or MR signal models[12–17]. Generative model-based approaches, such as SynthSeg[9,10], employ domain randomization to generate images with unrealistic but extreme contrast variations to enhance generalizability. Alternatively, MR signal model-based approaches, exemplified by PhysSeg[12], generate contrasts from quantitative tissue maps using signal equations associated with MR sequences.

These approaches, while demonstrated to be effective, have important limitations. Firstly, generative models often overlook the impact of MR imaging mechanisms and natural tissue transitions on image contrasts. This opens the door to generating unrealistic image contrasts, posing a risk of misleading network training and hindering the network's ability to generalize across real-world data. Secondly, existing MR model approaches[12–17] are constrained to one or two sequences that can be represented as static signal equations, only scratching the surface of possible contrasts. PhysSeg for example deals primarily with T1-weighted images generated with magnetization-prepared rapid gradient-echo (MPRAGE) and spoiled gradient echo (SPGR) sequences. This calls into question their generalizability to unforeseen contrasts. Thirdly, synthesized contrasts find predominant use solely in model training, sidelining their potential impact on the development life cycle of an AI model. Notably, their application in data labeling and performance benchmarking is conspicuously overlooked, despite the undeniable significance of these steps in AI development. These limitations undermine the potential of synthetic contrasts in developing *pan-contrast* AI models that are capable of handling all possible variations of MRI scans.

Here, we introduce UltimateSynth, a method developed from the ground-up based on the fundamental principles of MR physics to generate any possible MR image contrast on demand, with the ambitious goal of achieving pan-contrast generalizability for AI networks (Fig. 1). Rather than relying on MR sequence-dependent models, UltimateSynth utilizes classic spin dy-

namics equations to produce a full spectrum of image contrasts across the complete range of tissue properties and scanner parameters. This encompasses typical image contrasts from clinical scans, extreme contrasts attainable only through aggressive scanner settings, and suboptimal contrasts that may occur during routine acquisitions. Common contrasts such as T1-weighted, T2-weighted, diffusion-weighted, susceptibility-weighted, FLAIR, and BOLD represent only a small fraction of what can be generated with UltimateSynth. Additionally, UltimateSynth can synthesize contrasts beyond the physical limits of MRI scanners and sequences, equipping AI models to handle unforeseen imaging variations. The authenticity of the contrasts generated with UltimateSynth is a direct result of explicitly integrating intrinsic tissue properties, quantified through nuclear spin relaxation, and extrinsic imaging factors, like the static magnetic field, radiofrequency pulse, and acquisition timings. UltimateSynth's ability to generate diverse and genuine images enables training AI models compatible with any MRI scan.

Importantly, UltimateSynth's pan-contrast capability bolsters all three key stages of the AI development life cycle: Labeling, Emulation, and Deployment (LED). UltimateSynth offers customizable contrasts for efficient data labeling, facilitates widely generalizable model training to avoid costly retraining, and provides a platform for exhaustive performance benchmarking:

1. Labeling: The first step in AI development involves annotating a training dataset, either manually or using automated tools like FreeSurfer[18]. However, conventional approaches often label only one image per anatomy, necessitating considerable time to label large volumes of data for training. UltimateSynth substantially improves labeling efficiency by generating a variety of images from the same anatomy. This allows for the augmentation of training data and improves reliability by enforcing labeling consistency across different contrasts. UltimateSynth also enhances manual annotation by improving the visibility of structures with poor contrast, like the cerebellum, using customized image contrasts.

2. Emulation: The next stage of AI development involves training network models with labeled data. Training is typically confined to labeled contrasts, limiting model generalizability. UltimateSynth significantly broadens the training data by adding contrasts beyond the initially labeled ones. This is achieved by creating images across common, uncommon, and even unachievable contrasts, thus facilitating training of pan-contrast models and maximizes the value of labeled data.

3. Deployment: Before deployment, models must be thoroughly evaluated across various image contrasts. By simulating diverse and realistic images, UltimateSynth offers a platform for stress-testing AI models and quantifying uncertainty, ensuring model effectiveness across diverse datasets, vendors, and age groups.

We demonstrate with independent datasets spanning across scanners, acquisition protocols, and ages that a single U-Net[19] segmentation model trained with UltimateSynth, using only a modest number of acquired images, yields unprecedented generalizability and consistency across all cases. Notably, when stress-tested with more than 150,000 unique MR contrasts, the model yields high consistency in volumetric quantification with exceptionally low volumetric variability of 1.67%, exceeding 7-fold improvement over the state-of-the-art SynthSeg[9].

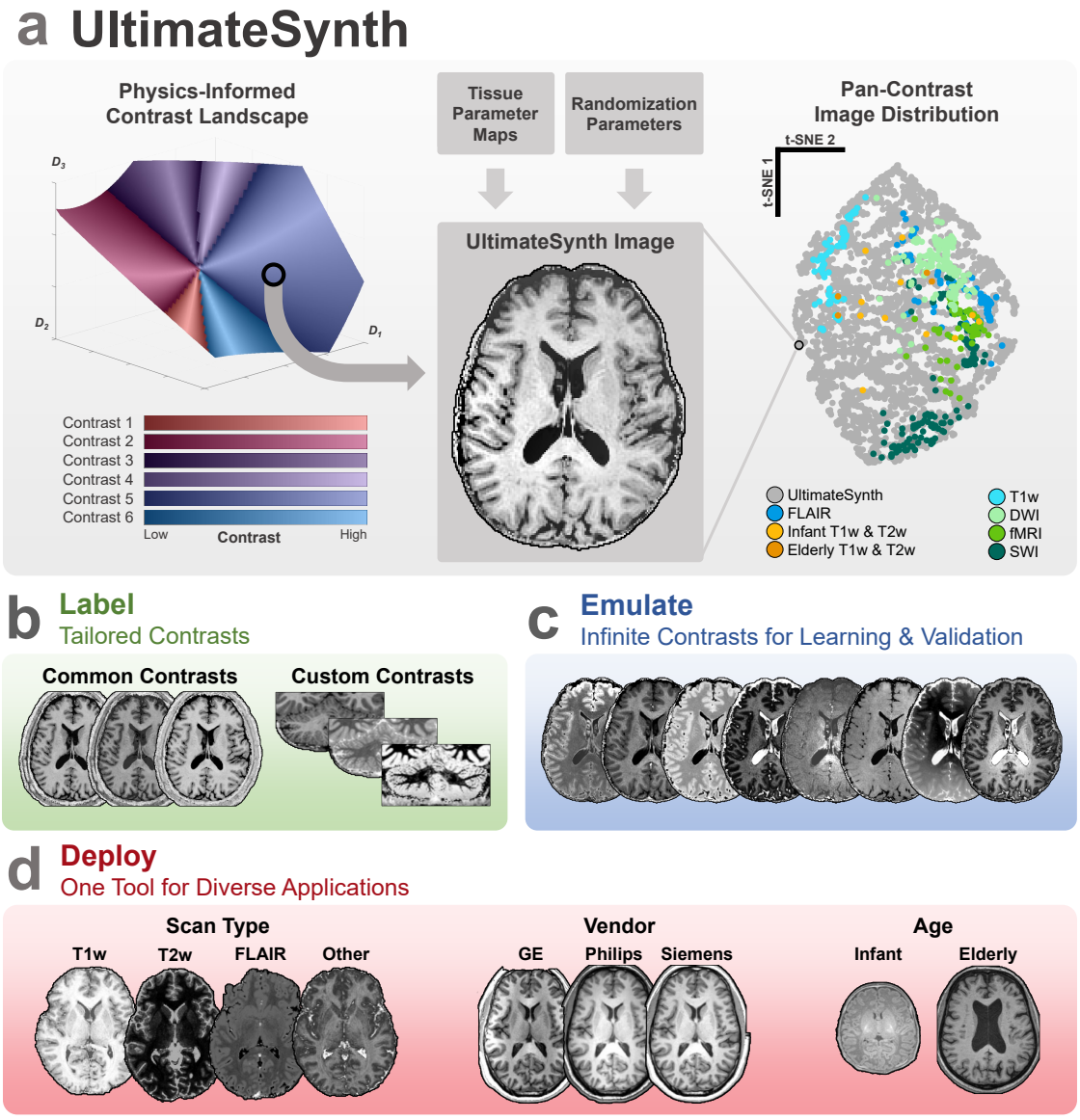

Fig. 2 | **UltimateSynth and the AI LED life cycle. a**, UltimateSynth generates an MR contrast landscape using physics-informed combination of diverse tissue and scan parameters. Combining sample locations from this landscape with tissue maps and random perturbations produces unique UltimateSynth images. Radiomics contrast features of possible UltimateSynth images (gray in t-SNE plot) completely encompass the same features from common and uncommon MR scans (blue, cyan, teal, light green, dark green, yellow, and orange in t-SNE plot). **b**, UltimateSynth generates tailored contrasts for automatic labeling using existing software or enhanced contrasts for easier manual delineation of specific structures. **c**, UltimateSynth generates realistic and diverse images from a single underlying anatomy for contrast-insensitive model learning. **d**, UltimateSynth allows pan-contrast evaluation to ensure model generalizability across different contrasts, vendors, and ages.

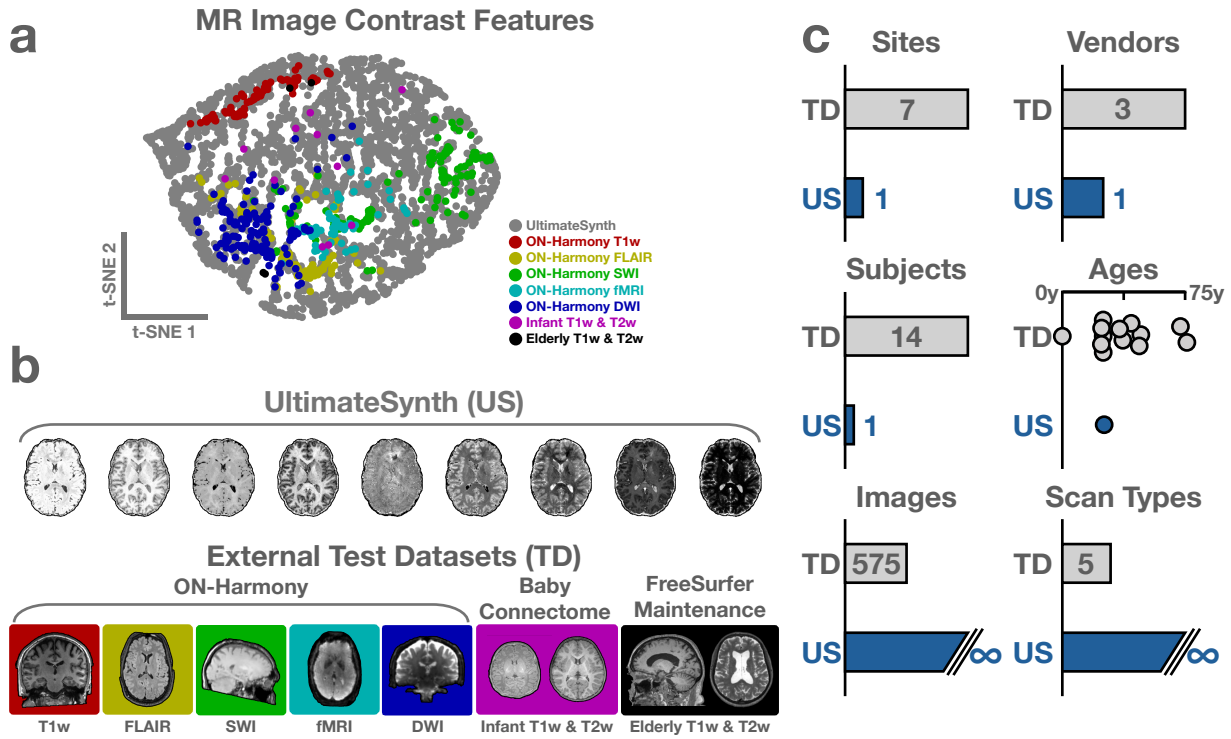

Fig. 3 | **Pan-contrast completeness of UltimateSynth. a**, t-SNE visualization of 15 first-order radiomics image intensity features extracted from 3,200 UltimateSynth images (gray) and 575 external test images from the ON-Harmony[20] (red, yellow, green, teal, blue), Baby Connectome Project[21] (purple), and FreeSurfer Maintenance[22] (black) datasets. **b**, Representative images of the datasets plotted in **a**, spanning three major age groups and five MR scan types (T1w, T2w/DWI, FLAIR, SWI, and fMRI). **c**, Quantitative comparison of the number of sites, vendors, subjects, ages, images, and scan types included in the UltimateSynth and external images in **a**.

RESULTS

Realistic Pan-Contrast Synthesis UltimateSynth simulates a comprehensive dictionary of signal magnetizations that reflect the combined effects of intrinsic tissue parameters (such as T1- and T2-relaxation times and proton density) and extrinsic acquisition parameters that determine contrast mechanisms (such as initial magnetization, pulse timings, and radiofrequency excitations). The simulation models classical nuclear spin dynamics without assumptions about anatomy or scan details. The UltimateSynth dictionary is then projected to a multidimensional subspace using singular value decomposition (SVD)[23], yielding a physics-informed contrast landscape. Each coordinate in the subspace is associated with an eigenvector of the dictionary, and each element of the eigenvector corresponds to a unique combination of tissue parameters. In Fig. 2a, we visualize this subspace using the first three primary eigenvectors with the landscape colored according to different relative tissue contrasts for brain MRI, as explained in Contrast Optimization. With additional information from tissue property maps, any desired MR contrast can be generated for an anatomy. Repeating this process with different subspace coordinates generates a diverse set of unique UltimateSynth images covering all possible MRI scan types, as illustrated by the

pan-contrast image distribution in Fig. 2a.

In this study, we utilized tissue parameter maps acquired from *in vivo* human brain scans to transfer voxel-wise tissue parameter combinations into a large number of uniquely weighted images, spanning common, uncommon, and unachievable contrasts across high- and low-end scanners with spatial inhomogeneities caused by varied scanning conditions or random perturbations. Fig. 3a illustrates that 3,200 images produced by UltimateSynth from a single subject fully encompass the diversity found in a set of 575 independent MR images. UltimateSynth images match the contrast variations seen across multiple datasets (Fig. 3b). As highlighted in Fig. 3c, achieving a similar range of contrasts with traditional MRI would typically require data from multiple sites, subjects, ages, and demographics. We demonstrate in Movie S1 the pan-contrast diversity of UltimateSynth for one subject. Leveraging the infinite range of possible UltimateSynth contrasts reflecting underlying tissue parameters, we aim to improve the entire AI development LED life cycle (Figs. 2b–d).

Efficient Labeling for Pan-Contrast Learning Utilizing UltimateSynth for both reference labeling and training data emulation, we trained off-the-shelf network models (nnU-Net[19]) for two segmentation tasks: whole-brain segmentation (UltBrainNet) with automatically labeled training samples and cerebellum segmentation (UltCerebNet) with manually labeled training samples.

For UltBrainNet, reference labels were generated automatically by segmenting multiple synthetic images with varied contrasts of the same anatomy like those in Fig. 2b using FreeSurfer (Fig. 4a), a whole-brain segmentation tool that is frequently used as a silver-standard reference in lieu of manual segmentation. The FreeSurfer labels across 32 subjects, each with 8 repeat segmentations, have an average of $4.66 \pm 10.95\%$ percent disagreement from majority (PDM, Fig. 4b) and an average of $6.67 \pm 7.16\%$ label volume variation (LVV, Fig. 4c) for non-background labels. Note that FreeSurfer may fail considerably with certain contrasts. FreeSurfer segmentations can deviate from the majority-voted white matter volume by as much as 19.65% in particularly poor images. For each subject, majority voting was performed across segmentations to generate a refined segmentation with increased resilience to contrast-dependent inconsistencies for model training.

For UltCerebNet, we generated customized tissue contrasts for the cerebellum to enhance the WM-GM tissue boundary for manual segmentation, as seen in Fig. 2b. UltCerebNet was pretrained using two manual segmentations and then applied to unlabeled data with manual correction to augment training data for further model refinement.

UltBrainNet and UltCerebNet were based on the off-the-shelf nnU-Net. Both models were trained using 6,400 pan-contrast images sampled from the SVD manifold using the eigencontrast technique explained in Efficient Contrast Sampling and Uniform Sampling of the Contrast Landscape. An example of the inter-tissue contrast diversity in these pan-contrast training images is provided in Fig. 2c.

Pan-Contrast Performance Benchmarking UltimateSynth provides a platform for comprehensive performance and generalizability assessment of AI networks prior to model dissemination. This platform allows for rigorous assessment of models, regardless of whether they are trained

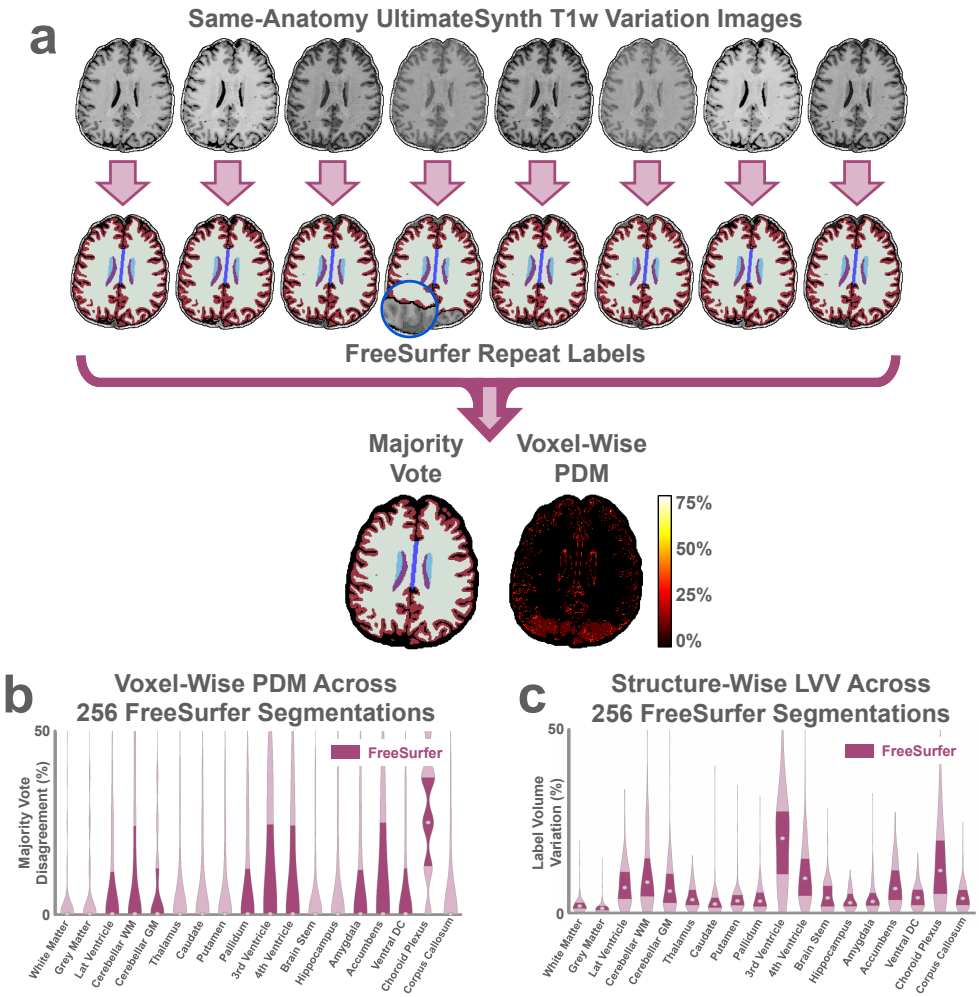

Fig. 4 | **Refined labeling using UltimateSynth.** **a**, UltimateSynth generates multiple images per subject, allowing for the combination of FreeSurfer segmentations to reduce contrast-dependent errors and major failures (red arrow). **b**, Voxel-wise percent disagreement from majority (PDM) across 256 (32 subjects $\times$ 8 repeats) segmentations. **c**, Structure-wise label volume variation (LVV) across 256 segmentations.

with or without UltimateSynth, across a broad spectrum of image contrasts. The flexible control over image generation provided by UltimateSynth also facilitates investigation of potential failure modes across the full contrast landscape. We show that pan-contrast models trained and validated using UltimateSynth (UltBrainNet and UltCerebNet) can be deployed with wide generalizability for segmentation tasks across diverse datasets unseen in training, encompassing varying contrasts, vendors, and ages (Fig. 2d).

Using 155,520 synthetic whole brain images with a wide spectrum of contrasts unseen in training, we compared the segmentation performance between UltBrainNet and the state of the art (SOTA) SynthSeg[9] (Fig. 5). For each subject, 19,400 images were generated using eigencontrast sampling and individually fed to the models, resulting in 19,400 segmentations for the same anatomy. Consensus segmentation labels were established through voxel-wise majority voting,

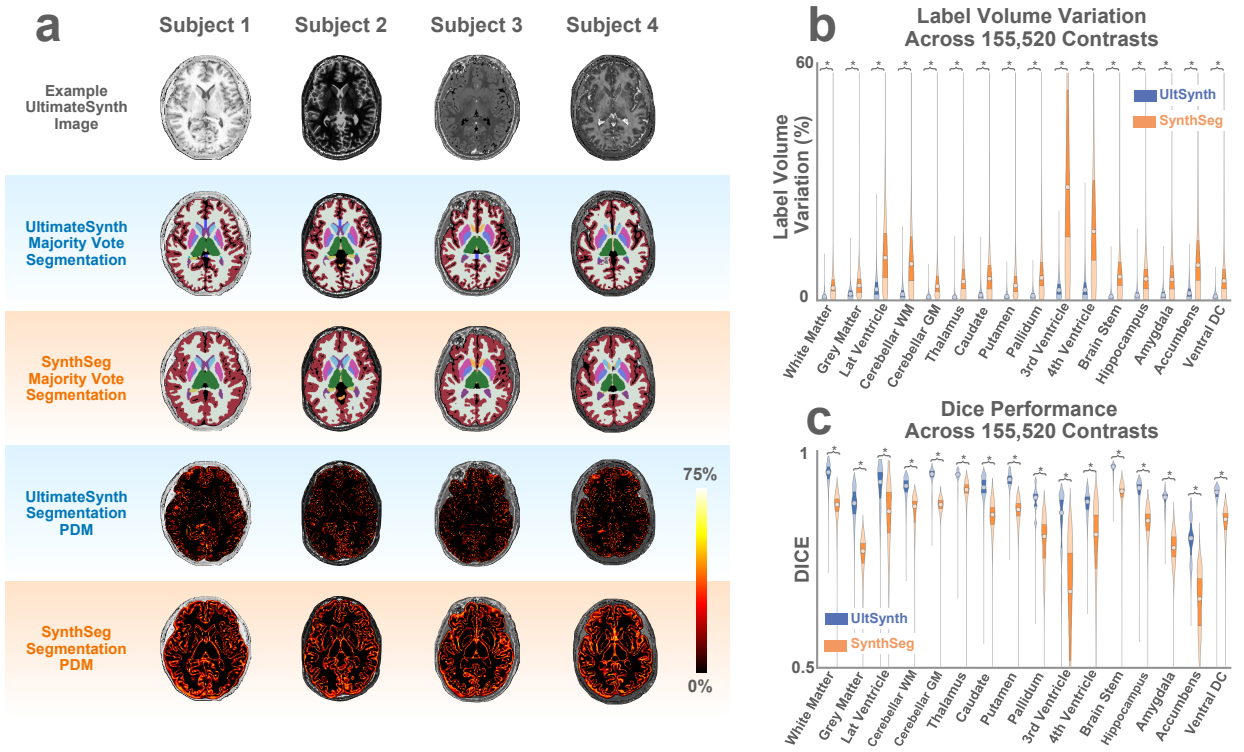

Fig. 5 | Brain segmentation performance evaluation using 155,520 unique contrasts generated via UltimateSynth. **a**, Four example subjects used in the evaluation of UltBrainNet and SynthSeg with 19,440 unique UltimateSynth contrasts per subject, uniformly sampled from the contrast landscape. Segmentation majority votes were calculated for same-subject images and disagreement frequency with the majority was evaluated using PDM on a voxel-wise basis. **b**, The percent variation in label volumes within each subject were evaluated using LVV across all 155,520 images of eight subjects (four additional subjects in Fig. S1). UltBrainNet consistently demonstrates less than 5% median volume variation across all examined structure labels, a significant improvement over SynthSeg ($P < .001$). **c**, Structure-specific Dice scores were also evaluated across all test images. Compared with SynthSeg, UltBrainNet yields significantly higher Dice scores with respect to the FreeSurfer consensus labels ($P < .001$).

and performance uncertainty was measured via the PDM metric. Across the eight subjects, UltBrainNet's PDM for non-background labels is $4.49 \pm 10.13\%$, compared with SynthSeg's $9.99 \pm 13.74\%$, which is notably higher around tissue boundaries, especially for the basal ganglia.

We further compared UltBrainNet and SynthSeg in terms of the LVV of all 16 shared structure segmentation labels for each subject (Fig. 5b). Across all 155,520 images, we measured a remarkably low LVV for UltBrainNet, never exceeding 5% median LVV across the 16 structure labels, and a mean LVV of $1.92 \pm 1.67\%$, compared to the SynthSeg's mean LVV of $11.93 \pm 14.06\%$. The structure-specific LVV for both methods are summarized in Table S1. The performance improvement given by UltBrainNet over SynthSeg remains unchanged across different types of MR scans (Fig. S2 and Table S2). Structure-specific Dice scores of all 155,520 segmentations compared to FreeSurfer consensus labels indicate that UltBrainNet yields a significantly higher mean Dice score of 0.897 ± 0.032 over SynthSeg's 0.805 ± 0.051 (Fig. 5c and Table S1). When directly comparing image-wise performance between both methods for each label (Fig. S3 and Fig. S4),

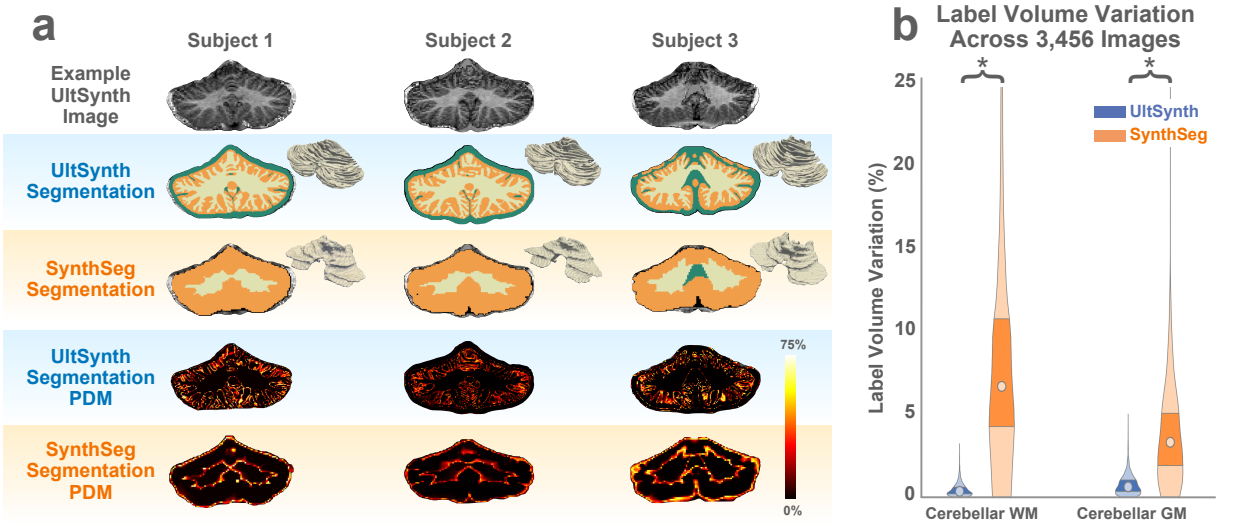

Fig. 6 | Cerebellum segmentation performance evaluation using 3,456 unique contrasts generated with UltimateSynth. **a**, Qualitative performance comparison of pan-contrast majority vote segmentation, reconstructed white matter surfaces, and voxel-wise PDM between UltCerebNet and SynthSeg on 1,152 unique contrasts per subject. **b**, LVV of cerebellar white matter (CWM) and gray matter (CGM) for all 3,456 images. UltCerebNet mean LVV is below 1% for both tissues, a statistically significant improvement over SynthSeg's 8.62% mean LVV for CWM ($P < .001$) and 3.88% mean LVV for CGM ($P < .001$).

UltimateSynth has higher Dice scores and lower LVV in at least 84.2% and 76.7% of all labels, respectively.

Comparing model performance of UltCerebNet and SynthSeg using 3,456 UltimateSynth images of the cerebellum generated from three subjects, Fig. 6a highlights UltCerebNet's detailed segmentation of the cerebellar WM in contrast to SynthSeg's more conservative estimations. UltCerebNet demonstrates lower voxel-wise PDM than SynthSeg ($4.65 \pm 10.47\%$ versus $6.54 \pm 12.03\%$). Fig. 6b compares the cerebellum GM and WM LVV values across all images of each subject, showing a nearly 18-fold lower LVV for UltCerebNet than SynthSeg ($0.55 \pm 0.43\%$ versus $9.61 \pm 9.09\%$, as shown in Table S3.

Pan-Contrast Assessment with External Datasets We compared the performance of UltBrainNet with two SOTA contrast-agnostic algorithms, SynthSeg and PhysSeg[12]. For this comparison, we utilized the publicly available ON-Harmony dataset[20], which includes 560 3D MRI images from 10 subjects. The dataset is comprehensive, featuring data from 3 different scanner vendors, collected across 5 different sites and involving 6 different scanners. It covers 5 types of MRI scans: T1-weighted imaging, FLAIR, diffusion-weighted imaging (DWI), susceptibility-weighted imaging (SWI), and functional MRI (fMRI).

Figure 7 provides a summary of the segmentation outcomes of UltBrainNet, SynthSeg, and PhysSeg specifically for T1-weighted and FLAIR images, amounting to a total of 160 images. Even with the presence of mild to severe contrast variations in repeat scans of the same subject, both T1-weighted and FLAIR, UltBrainNet consistently demonstrated superior segmentation per-

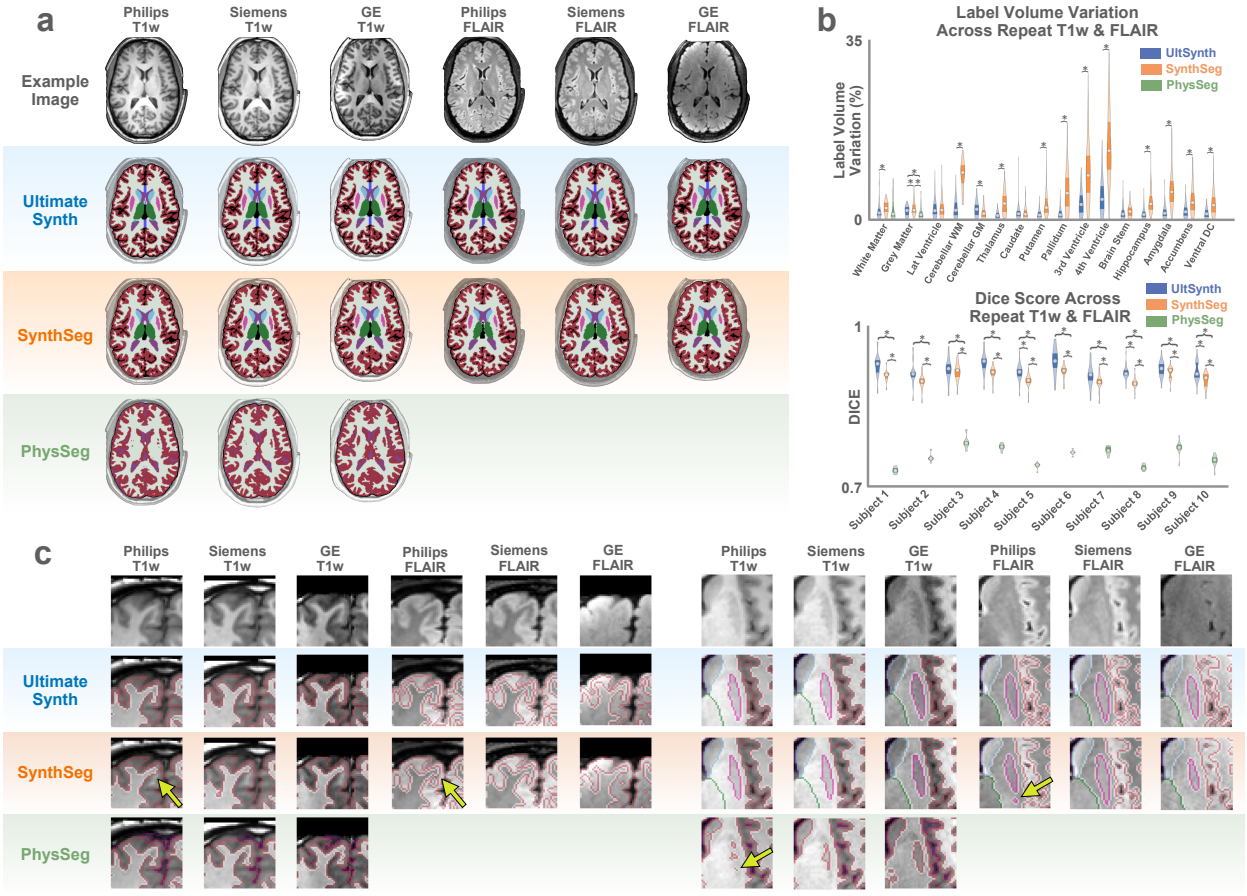

Fig. 7 | UltBrainNet segmentation generalization to the external ON-Harmony dataset spanning multiple vendors, sites, scanners and scan types. **a**, Single-subject comparison of UltimateSynth, SynthSeg, and PhysSeg across three major scanner vendors and two weighted contrasts. PhysSeg results for FLAIR images are excluded, as the model is designed solely for T1-weighted images. **b**, Quantitative performance evaluation of 160 images spanning 10 subjects, three vendors, five sites, six scanners, and two scan types. UltimateSynth demonstrates less than 5% LVV across 16 labels, with statistically significant improvements over the SOTA for 10 of 16 structures ($P < .001$). UltimateSynth and SynthSeg yield comparable Dice scores with respect to FreeSurfer reference labels. PhysSeg is notably limited by its restriction to three-tissue labeling. **c**, Examination of regions with segmentation difficulties. SynthSeg struggles to delineate the folded sulcus in the left frontal lobe, whereas PhysSeg is inconsistent in its putamen label continuity. UltimateSynth produces plausible tissue boundaries across all six contrasts for both regions.

formance compared to SynthSeg and PhysSeg (Fig. 7a). This consistent improvement suggests that models generated by UltimateSynth are capable of producing "harmonized" measurements that remain robust and accurate under varying imaging conditions.

Whole-brain volume-weighted Dice scores for co-registered scans of the same anatomy show that UltBrainNet and SynthSeg achieve similar performance when compared to FreeSurfer references, with scores of 0.901 ± 0.007 and 0.889 ± 0.007, respectively. In contrast, PhysSeg performs worse, with a score of 0.767 ± 0.015, partly due to its three-tissue labeling approach (Fig. 7b). Detailed Dice scores for individual subjects are provided in Table S4.

UltBrainNet also demonstrates significantly lower label volume variance (LVV) than either SOTA method in 12 out of 16 anatomical structures, as detailed in Table S5. While UltBrainNet and SynthSeg achieve comparable Dice scores, UltBrainNet shows a marked advantage in reducing LVV. Furthermore, UltBrainNet consistently identifies critical morphological features, such as the spatial continuity of folded gyri and the putamen, as illustrated in Fig. 7c.

We also extended our evaluation to include other types of MR scans, such as susceptibility-weighted imaging (SWI) (Fig. S5 and Table S6), diffusion-weighted imaging (DWI) (Fig. S6 and Table S7), and functional MRI (fMRI) (Fig. S7 and Table S8), as presented in the Supplementary Materials. These images vary significantly in resolution and quality and often exhibit geometric distortions and cropped fields of view (FOVs). Despite not being specifically trained on these types of variations, UltBrainNet significantly outperforms SynthSeg in maintaining label consistency across repeated scans.

Finally, we investigated a particularly challenging segmentation task with infant brains, which have unique intra-tissue and cross-age intensity changes due to gradual myelination processes[24]. We apply UltBrainNet, SynthSeg, and Infant FreeSurfer[25] (infantFS) to a series of same-subject longitudinal MR images acquired in an independent study[21] (Fig. 8a). In contrast to the obvious fail cases for non-UltimateSynth based methods for the 17-day-old image, we note appreciable consistency in the repeated segmentation of anatomical features by UltBrainNet, despite age-related contrast changes driven by myelination (Fig. S8). Additional time points and T2-weighted images of the same subject, shown in Fig. S9, confirms the segmentation consistency of UltBrainNet both longitudinally and across types of MR scans.

We reiterate UltimateSynth's utility as an AI performance assessment platform by creating a single-subject 1,152 image pan-contrast stress-test set from a 6-month-old with incomplete WM myelination. Multi-contrast MRI data for infants at this age is scarce, making our evaluation a valuable opportunity to rigorously test AI models in rare conditions (Fig. 8b). UltBrainNet yields an average PDM of $8.29 \pm 13.39\%$ across all 1,152 images, compared with SynthSeg's $15.04 \pm 14.68\%$. As seen in Fig. S10, UltBrainNet's LVV for this infant set is similar to its performance on adult pan-contrast images, with only $2.45 \pm 1.86\%$ LVV on average. Driven by uncertainty around the white matter and basal ganglia, SynthSeg's label variation is substantially larger, with a mean LVV of $19.62 \pm 18.19\%$. The exceptional performance of UltBrainNet on even challenging cases like developing infants despite using only a few adult training anatomies highlights the effectiveness of UltimateSynth in training pan-contrast models. The pan-contrast LVV for each structure can be found in Table S9.

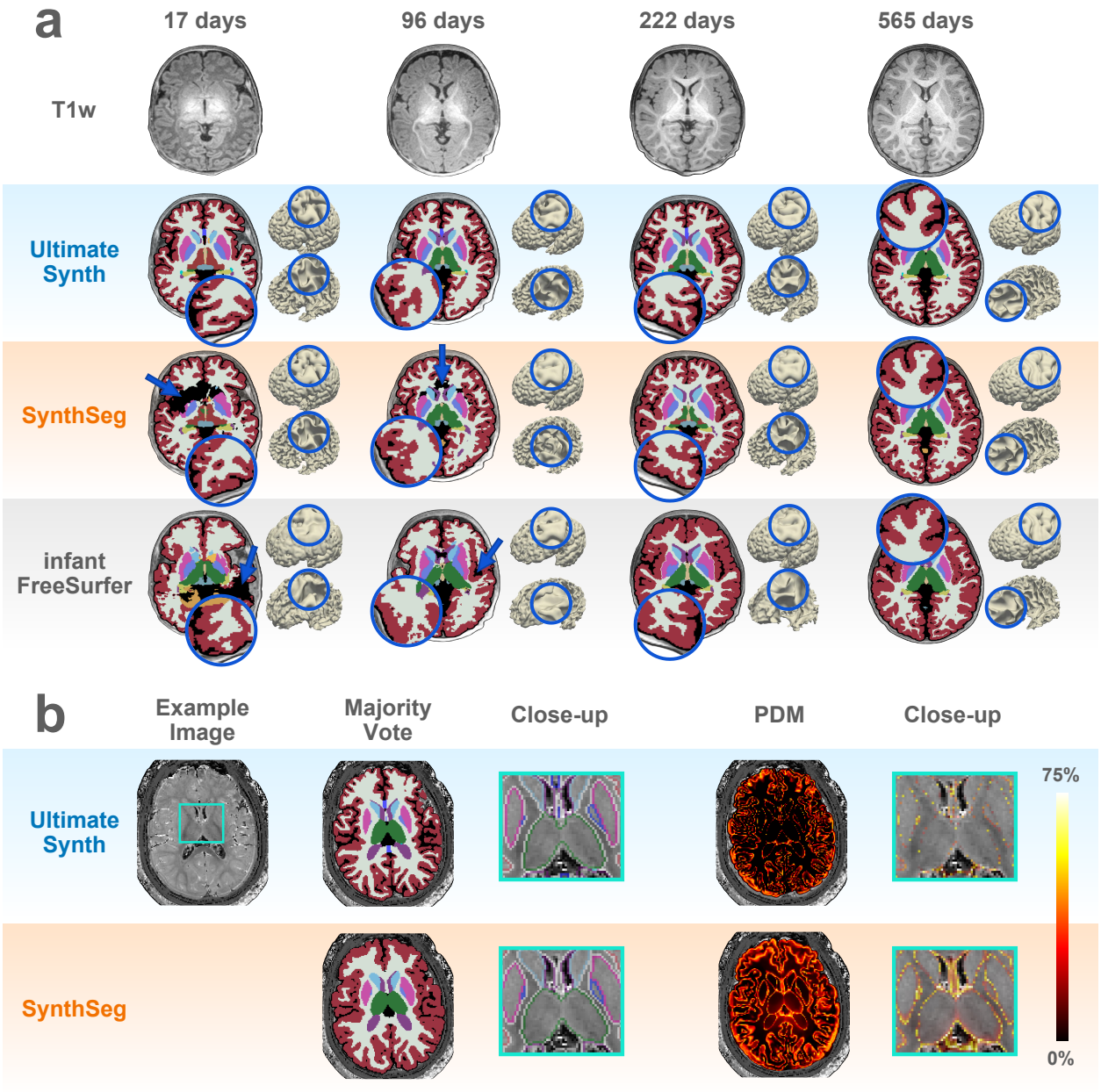

Fig. 8 | **UltBrainNet segmentation generalization to age-related tissue contrast variations.** **a**, Longitudinal segmentation consistency from birth to 18 months of life. UltBrainNet consistently segments small sulci, unlike SynthSeg and Infant FreeSurfer, which predict drastically different tissue boundaries across age-related contrast changes despite the same underlying anatomy. These inconsistencies are highlighted across the pial (top) and white matter (bottom) surface reconstructions at each time point. At very young ages like 17 days and 96 days, existing methods may produce serious segmentation failures (blue arrows). **b**, Quantitative evaluation of segmentation performance across 1,152 UltimateSynth images from a 6-month-old subject. UltimateSynth demonstrates lower voxel-wise PDM than SynthSeg, especially around subcortical areas like the thalamus, as emphasized in the close-up images.

DISCUSSION

We introduced UltimateSynth—a versatile framework that integrates MR physics and tissue physiology to generate a wide spectrum of realistic MR contrasts without repeated scanning—to support the labeling, emulation, and deployment stages of AI development. UltimateSynth facilitates the generation of labeled data for training networks that are highly generalizable, covering over 150,000 unique contrasts with exceptionally low variability below 2%.

Physics-Informed Pan-Contrast Emulation UltimateSynth utilizes MR spin dynamic simulation for unparalleled versatility in generating diverse and authentic image contrasts. The UltimateSynth contrast dictionary transcends the limitations of physical scanners and covers a vast spectrum of tissue parameters. It simulates essential contrast mechanisms that govern any MR sequences, including initial magnetization (stemming from magnetization preparation methods and radiofrequency excitations), timing (crucial for any MR scan), and phase (affected by field inhomogeneity, spin dynamics and interactions, or gradients), and supports adding random measurement errors and spatial variations for robust model training.

Unlike recent physics-based approaches that are limited to one or two MR sequences focusing on healthy tissue, UltimateSynth offers a comprehensive pan-contrast simulation, providing a much broader spectrum of imaging contrasts. This extensive range allows for more diverse and realistic training data. In contrast to domain-randomization methods like SynthSeg, which utilize unrealistic Gaussian-distributed signals, UltimateSynth generates authentic contrasts that accurately reflect in vivo tissue transitions and partial volume effects. This realism significantly enhances the ability of networks trained with UltimateSynth to segment small structures, particularly in regions with low signal-to-noise ratios (Fig. S11, blue arrows). As a result, UltimateSynth-trained models demonstrate a notable advantage over SynthSeg in handling complex and subtle anatomical details.

Focusing on T1 and T2 relaxation times——key for generating common MR contrasts——our simulations span a range from no spin dynamics to pure distilled water, covering all healthy and pathological tissue scenarios. Given that multiple tissue parameters can have shared influence on spin dynamics, T1 and T2 variations also capture magnetization effects related to diffusion, $T2^*$, susceptibility, and multi-tissue exchanges. We show that UltimateSynth-trained networks perform well in segmenting alternative MR scan types, such as susceptibility-weighted images with very low contrasts and reduced through-plane resolutions (Fig. S5), as well as diffusion and functional MR images (Figs. S6 and S7). Fig. S13a visually demonstrates UltimateSynth's comprehensiveness through a t-SNE plot[26], showing its coverage of contrasts across diverse MR scan types and age groups.

Customizable Synthesizer for Efficient Labeling UltimateSynth addresses two fundamental challenges in medical image labeling—scarcity and uncertainty—by generating an infinite number of contrast variations for each anatomical structure. Traditional methods often require extensive manual labeling for each new contrast, which is both time-consuming and prone to inconsistencies. UltimateSynth overcomes this by providing a virtually limitless supply of contrast variants derived from a single labeled anatomy. This capability not only eliminates the need for contrast-specific relabeling but also significantly increases the quantity and diversity of training

samples available. Consequently, it minimizes inter-contrast labeling inconsistencies, ensuring that the models trained on these datasets are more robust and reliable.

One of the most powerful features of UltimateSynth is its ability to propagate labels created for a specific contrast across an infinite range of other contrasts. This allows for the development of pan-contrast models that are capable of generalizing beyond the contrast-specific limitations of traditional tools. Essentially, UltimateSynth enables AI models to learn from a wide spectrum of imaging contrasts without requiring separate, manually labeled datasets for each one. This ability to generalize contrast-specific tools into pan-contrast applications is crucial for developing robust AI models that perform consistently well across different imaging environments. As a result, researchers and developers can accelerate the creation of advanced AI models that are not only versatile but also more cost-effective and easier to train.

UltimateSynth customizes imaging contrasts to align with the input requirements of various labeling software, enabling efficient, automated, and diverse labeling processes. This capability significantly broadens the range of tools available for generating the large volumes of labeled data needed for effective model training. In contrast, domain-randomized methods like SynthSeg generate unusual contrasts that are often incompatible with standard labeling tools, limiting their usefulness. Additionally, UltimateSynth enhances anatomical visibility through tailored contrasts, making manual labeling easier and facilitating the creation of new labels. Conversely, extending methods like SynthSeg to include new labels is difficult due to their limited support for manual labeling, posing challenges for expanding their applicability.

UltimateSynth provides label-time augmentation, allowing for the evaluation of label quality by comparing multiple labeling outcomes for the same anatomical structure. This process aids in identifying and correcting inaccuracies in labels. Techniques such as majority voting can be applied to refine and consolidate these labels, essential for training robust and high-performing models. By ensuring labels are more accurate and consistent, UltimateSynth reduces the risk of models learning from erroneous data, thereby enhancing their reliability and performance. Additionally, this approach contributes to the creation of more precise ground truth labels for model evaluation, offering a more dependable benchmark for assessing model efficacy and reducing errors in practical applications.

Platform for Exhaustive AI Benchmarking UltimateSynth provides diverse or customized imaging contrasts, enabling comprehensive benchmarking of AI models under varied conditions. Until now, the lack of a dedicated platform to thoroughly test these models across different contrasts has raised concerns about their ability to generalize effectively. Robust evaluation requires a large dataset of labeled images that cover a wide spectrum of contrasts, which is difficult to obtain in practice. The process of manually labeling such a vast number of images is extremely labor-intensive and time-consuming. Furthermore, images acquired from the same anatomical structure are often not perfectly aligned due to variations in patient positioning, acquisition angles, or slight movements during scanning. These misalignments introduce errors and inconsistencies into the analysis process, contaminating the evaluation results. Consequently, this could lead to inaccurate conclusions about an algorithm's performance and its generalizability across different imaging conditions.

UltimateSynth addresses these challenges by providing a standardized and controlled platform for reliable and accurate assessment of model robustness, enabling the rapid development of pan-contrast AI models. This platform supports test-time augmentation, which allows for comprehensive quantitative analysis of segmentation performance on the same anatomical structures across a wide range of perfectly aligned contrasts. This offers a rigorous test of a model's ability to generalize across varying imaging conditions. We demonstrated the utility of UltimateSynth by segmenting over 19,000 unique contrast images per subject, providing a detailed assessment of model performance. This extensive evaluation allows us to measure volumetric variations and voxel-wise segmentation disagreements, which are critical for identifying model weaknesses, anticipating potential points of failure, and guiding the development of targeted corrective solutions to improve model robustness and reliability.

Wide Generalizability with Limited Training Data UltimateSynth achieves extensive generalizability with a relatively modest training sample size, demonstrating its effectiveness in training robust AI models. For example, UltBrainNet, an AI model developed using UltimateSynth, was trained on a dataset of only 32 healthy volunteers with an average age of 25.6 ± 4.3 years. Despite the limited number of training samples, the model exhibited consistent and reliable performance across a wide range of conditions and throughout the entire human lifespan, as shown in Fig. S14. Remarkably, UltBrainNet also generalized well to an entirely independent dataset that presented significant variability, including data from three different MR vendors, six different scanners, five distinct imaging sites, five types of MR scans, and a range of voxel resolutions. This ability to perform robustly across such diverse conditions underscores the model's strong generalizability and its potential for application in various clinical and research settings without the need for extensive retraining or adjustment.

To assess the impact of training sample size on model performance, we trained six neural networks, each using a fixed set of 6,400 UltimateSynth-generated images. However, the images were derived from varying numbers of unique subjects, ranging from 1 to 32 (referred to as US1 through US32). The networks trained with fewer unique anatomical examples (e.g., US1) faced challenges when confronted with unexpected anatomical variations during testing, especially in regions such as the lateral ventricles (Fig.S12a, indicated by blue arrows). As the number of unique subjects increased, the networks' ability to handle anatomical variability improved. However, the performance gains began to plateau after including more than 16 subjects (Fig.S12c and Table S10).

Remarkably, the network trained with 16 subjects (US16) outperformed SynthSeg——which was trained on over 1,000 subjects——in terms of Dice similarity coefficient for 11 anatomical labels and lateral ventricle volume (LVV) for 14 labels. This demonstrates that a well-curated, smaller dataset can achieve competitive or even superior performance compared to much larger datasets. All six networks exhibited strong pan-contrast capabilities and performed consistently well across different imaging contrasts without showing significant trends in volume variation for specific anatomical structures (Fig.S12b and TableS11). Furthermore, visualization using t-SNE (Fig. S13b) showed that the degree of pan-contrast generalization was similar between the model trained with 32 subjects (US32) and the one trained with only 1 subject (US1). This suggests that while increasing the diversity of training subjects improves robustness to anatomical variation, it does not necessarily affect the model's ability to generalize across different imaging contrasts.

The ability to achieve broad generalizability with a small training sample size has profound implications for the development of AI models, particularly in the field of medical imaging and other data-intensive domains. Training models effectively with fewer data samples means that the process becomes less dependent on large, meticulously labeled datasets, which are often difficult and costly to obtain. This reduction in data requirements significantly lowers the number of iterations needed for human-in-the-loop model training, where human experts are required to refine labels, provide feedback, and guide the model's learning process.

By minimizing the need for extensive manual labeling and continuous data acquisition, the development cycle for AI models becomes more efficient. It accelerates the pace of AI development by reducing the time spent on label refinement and model retraining, which are traditionally labor-intensive and time-consuming. Furthermore, the streamlined process facilitates more straightforward integration of human feedback, allowing for faster adjustments and improvements based on real-world data and expert insights. This not only makes the development of robust AI models more accessible and manageable but also more cost-effective, enabling smaller research teams and organizations with limited resources to build high-performance AI models that are well-suited for practical applications. Overall, this efficiency enhances the scalability and adaptability of AI technologies, broadening their potential impact across various fields.

Future Work UltimateSynth's current application to healthy human brain MRI is merely one example of the platform's broad potential. Its highly adaptable framework is designed to support a wide range of use cases, encompassing various diseases, anatomical structures, and even different species. By generating high-quality, diverse, and well-labeled datasets, UltimateSynth serves as an ideal tool for training MRI foundation models. These foundation models are capable of learning generalized representations from extensive datasets, allowing them to achieve strong performance across a variety of tasks with only minimal fine-tuning. This transfer learning approach significantly accelerates AI development by reducing the dependence on large, meticulously annotated datasets, which are often challenging and costly to produce.

The platform's ability to provide a broad diversity of imaging contrasts further enhances the versatility of the models trained on it. Models developed using UltimateSynth can be leveraged as foundation models for a range of advanced tasks, such as super-resolution imaging, which enhances the quality and resolution of MRI scans beyond their original acquisition parameters; spatial normalization, which aligns images to a standard anatomical template for comparative analysis across studies or populations; and anomaly detection, which identifies unusual patterns that may indicate disease or abnormalities. By facilitating the training of robust, contrast-agnostic AI models, UltimateSynth helps bridge the gap between research and clinical application, enabling the development of more effective and adaptable AI tools for medical imaging. This flexibility is crucial for advancing AI-driven diagnostics, treatment planning, and personalized medicine across various healthcare domains.

METHODS

MRF Data MR Fingerprinting (MRF)[27], an efficient and clinically feasible technique[27–29], was used to obtain the tissue quantitative maps from 32 healthy volunteers (25.6±4.3 years of age) at a single site[30]. Scanning was approved by the local Institutional Review Board and were conducted using a 3T Siemens Prisma scanner with a 20-channel head coil. Prior to scanning, all participants provided written informed consent. The tissue property maps from each subject included co-registered quantitative T1, T2, and proton density maps, all simultaneously acquired from a single MRF scan. Three-dimensional head scans were acquired with whole brain coverage using an axial acquisition, a $300\times300\times144\,mm^3$ field of view, 1.0 mm isotropic image resolution, and a scan time of 10 minutes. A fast B1 mapping scan (20 seconds) was acquired from the same subject to measure the transmit field inhomogeneity. Quantitative T1, T2, and proton density maps were generated using a low-rank algorithm.

Traveling Phantom Data The ON-Harmony[20] traveling heads dataset was utilized for testing purposes. The data consists of ten subjects (33.2±9.9 years of age) each with between 6 and 11 scan sessions across three vendors, five sites, and six scanners, for a total of 560 combined T1-weighted, FLAIR, SWI, DWI, and fMRI images. The data were downloaded from the OpenNEURO public database (`https://openneuro.org/datasets/ds004712`). Same-subject anatomical scans were rigidly co-registered for subsequent evaluation using FSL's FLIRT[31–33]. Original modality resolutions were 1 mm isotropic for T1-weighted and FLAIR, 2 mm isotropic for DWI, 2.4 mm isotropic for fMRI, and $0.8\times0.8\times3\,mm^3$ for SWI. All images were resampled to 1 mm isotropic using linear interpolation as a preprocessing step.

Baby Data Two longitudinal T1-weighted and T2-weighted MRI scans of 2 subjects enrolled as part of the Baby Connectome Project (BCP)[21] were used in testing the UltBrainNet. The images were acquired using a 3T Siemens Prisma MRI scanner equipped with a Siemens 32-channel head coil. T1-weighted MR images were acquired with 208 sagittal slices, TR/TE = 2,400/2.24 ms, flip angle = 8°, acquisition matrix = 320 × 320, and resolution = $(0.8\,mm)^3$. T2-weighted MR images were acquired with 208 sagittal slices, TR/TE = 3,200/564 ms, variable flip angles, acquisition matrix = 320 × 320, and resolution = $(0.8\,mm)^3$. One six-month-old subject scanned using a $(0.8\,mm)^3$ isotropic resolution 5-minute MRF protocol[34] was also included in testing to evaluate infant UltimateSynth images.

Elderly Data T1-weighted and T2-weighted MRI scans for two elderly subjects (76 years of age and 72 years of age) demonstrating appreciable lateral ventricle enlargement were obtained from the open-source FreeSurfer Maintenance Dataset[22] for testing purposes. The data were downloaded from the OpenNEURO public database (`https://openneuro.org/datasets/ds004958`) and intra-subject scans were rigidly co-registered using FLIRT[31–33].

Physics-Informed Contrast Generation In nuclear magnetic resonance (NMR) and magnetic resonance imaging (MRI), the acquired signal is determined by both intrinsic tissue properties, such as relaxation times, proton density, diffusion, and hemodynamics, and extrinsic imaging factors, including the static magnetic field (B0), radiofrequency (RF) field (B1), and acquisition timings (echo time, TE; repetition time, TR). Here, we applied the classic Bloch equations[35] to synthesize diverse image contrasts, covering the full range of combinations of imaging factors

(initial magnetization, RF pulses, TR, TE and phase) from a set of tissue properties (T1, T2 and proton density). Specifically, the spin magnetization $M_r \in \mathbb{R}^3$ at a given echo time (TE) is described by a state equation[36] of a single isochromat r:

$$M_r = A_r(\Lambda, \theta) M_r^0 + B_r(\Lambda, \theta), \tag{1}$$

where M_r^0 is the initial magnetization prior to the RF excitation. This magnetization is in the form of $[0\ 0\ M_z]^\top$, where the value of $M_z \in [-1, 1]$ can be modified by various preparation mechanisms, such as T1 inversion, T2 preparation, or diffusion preparation pulses. We simulated the M_r^0 magnetization from an inversion preparation pulse with randomly varied inversion time (TI) without physical limitation. θ represents a set of tissue properties (e.g., T1, T2, and proton density) and experiment-specific parameters (e.g., B1 inhomogeneity), whereas Λ comprises imaging factors, including RF pulses and TE times. A_r is a system matrix $A_r(\Lambda, \theta) = R(\mathrm{T1}, \mathrm{T2}, \mathrm{TE}) Q(\alpha, \phi)$, with $R(\mathrm{T1}, \mathrm{T2}, \mathrm{TE})$ modeling spin relaxation:

$$R(\mathrm{T1}, \mathrm{T2}, \mathrm{TE}) = \begin{bmatrix} e^{-\mathrm{TE}/\mathrm{T2}} & 0 & 0 \\ 0 & e^{-\mathrm{TE}/\mathrm{T2}} & 0 \\ 0 & 0 & e^{-\mathrm{TE}/\mathrm{T1}} \end{bmatrix}, \tag{2}$$

and $Q(\alpha, \phi)$ modeling RF excitation:

$$Q(\alpha, \phi) = \begin{bmatrix} \cos(\phi) & \sin(\phi) & 0 \\ -\sin(\phi) & \cos(\phi) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\beta\alpha) & \sin(\beta\alpha) \\ 0 & -\sin(\beta\alpha) & \cos(\beta\alpha) \end{bmatrix} \begin{bmatrix} \cos(\phi) & -\sin(\phi) & 0 \\ \sin(\phi) & \cos(\phi) & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

where α is the RF excitation flip angle, ϕ is the RF pulse phase, and β is a spatially-dependent scaling bias field influenced by B1 inhomogeneity. Lastly, $B_r(\Lambda, \theta)$ is the input matrix:

$$B_r(\Lambda, \theta) = \begin{bmatrix} 0 & 0 & M_z(1 - e^{-\mathrm{TE}/\mathrm{T1}}) \end{bmatrix}^\top. \tag{4}$$

The resulting image contrast landscape can be encoded via a two-dimensional dictionary matrix $D^\star$. The rows encode all combinations of T1 and T2 values. The columns encode all combinations of imaging factors: TI, RF, and TE. The T1 and T2 ranges were selected from 0 (no spin dynamics) to 5000 milliseconds (T1 and T2 limits of pure distill water at room temperature). Because network training was based on the signal magnitude, the signal phase was not encoded in the dictionary. However, MR is sensitive to parts per million (p.p.m.) level deviations in the B0 field, and so the signal phase induced from field inhomogeneities can be simulated in the dictionary in the future as well. For imaging factors, we simulated TI values ranging from 10 to 5000 milliseconds, RF values from 0 to 90 degrees, and TE values from 1 to 400 milliseconds. These ranges were chosen to ensure that magnetization at TE for any tissue type spans the full range from -1 to 1. A total of 122,080,000 dictionary entries were generated in 1.1 seconds on a standard desktop computer. The relative GM-WM-CSF contrasts for UltimateSynth dictionary entries with unique TI-TE combinations and a fixed 90 degree RF pulse are visualized as a 2D plot in Fig. S15.

Efficient Contrast Sampling To effectively sample in the contrast landscape, we applied singular vector decomposition (SVD) on the mean-subtracted $D^\star$ to extract the most dominant contrasts. These eigencontrasts, encoded in matrix D, established a coordinate system that can facilitate contrast sampling. Identical to $D^\star$, the rows of D encompass all T1 and T2 combinations ranging from 0 to 5000 ms. We selected the top $K = 28$ sorted singular values to preserve over 99.999 percent of the total energy.

For each pair of T1 and T2 maps obtained from a quantitative MR scan, a subject-specific eigencontrast space was generated by selecting the corresponding rows from D based on voxel-wise T1 and T2 values. Contrasts were generated with random coefficients sampled from a normal distribution. The coefficients were weighted by the square roots of the corresponding singular values and then were multiplied with the left singular vectors $\{U_k\}_{k=1}^{K}$ from the SVD of $D^\star$. We partition the contrast landscape into 6 regions based on the intensity differences between white matter, gray matter, and cerebrospinal fluid (as defined in Contrast Optimization). By balanced sampling from each of these regions, we ensured contrast diversity in network training.

Uniform Sampling of the Contrast Landscape To investigate pan-contrast segmentation, it is useful to have a method of uniformly sampling the high-dimensional contrast landscape to ensure all possible UltimateSynth contrasts are sampled. As the monotonically decreasing singular values S from the SVD decomposition of $D^\star$ scale the magnitude of $\{U_k\}_{k=1}^{K}$ to produce D, it is natural that the larger singular values have a greater influence on the resulting UltimateSynth contrasts, and thus the dimensions associated with them in the contrast landscape need to be more densely sampled than the dimensions of smaller, less influential singular values. As the smallest discernible resolution in one dimension of an eigencontrast space is two points at the normalized coordinates $[-1, 1]$, we define a set of P density vectors $\{V_1, V_2, \ldots, V_P\}$ sampling different landscape dimensions, each with entries linearly spaced on the range of $[-1, 1]$, where the number of entries in the pth vector V_p is defined as the ratio S_p/S_P rounded to the nearest integer.

From the set $\{V_1, V_2, \ldots, V_P\}$, a new set of coordinate vectors evenly sampling the first P dimensions of the contrast landscape can be extracted, where each coordinate vector is a unique combination of one entry from each of the P density vectors. This process results in $n = \prod_{p=1}^{P} \lfloor S_p/S_P \rceil$ unique eigencontrast coefficients spanning the first P dimensions of the contrast landscape. We use two such evenly distributed eigencontrast coefficient sets in this work: one with $n = 155,520$ and $P = 9$ as seen in Fig. 5, and a second with $n = 1,152$ and $P = 6$ as seen in Figs. 6 and 8. For all higher dimensions $\{P + 1, \ldots, K\}$, the contrast landscape was sampled at each dimension's origin.

LVV As a Function of Contrast Type To investigate the possibility of performance varying as a function of inter-tissue relative voxel intensity, the pan-contrast dataset of 155,520 UltimateSynth images used in Fig. 5 was divided into four major contrast types in Fig. S2: 48,280 "T1-weighted" images where the relative ordering of voxel intensities for the three major tissue types of white matter, gray matter, and CSF were WM > GM > CSF, 18,304 "T2-weighted" images where the voxel intensities were ordered CSF > GM > WM, 48,936 "FLAIR" images where the ordering was GM > WM > CSF, and 40,000 "Other" images that encompass the other three possible relative orderings of the three tissues of interest. For these comparisons, we consider the same tissue

definitions as outlined in Contrast Optimization.

Bias Field The MRI bias field, also known as intensity inhomogeneity or shading, refers to non-uniformities in the intensity of the MRI signal across the image. This phenomenon arises due to variations in the scanner performance (field inhomogeneity), sensitivity of the MRI scanner's radiofrequency coil, and other factors in the imaging process. The bias field is scanner- and subject-dependent, making it challenging to model and correct. We implemented simulations of synthetic bias fields for U-Net training based on the approach in Billot et al's SynthSeg[9]. Voxel-wise intensities in a 3^3 low-resolution bias volume B' were sampled from a randomized Gaussian distribution, $\mathcal{N}_{3\times3\times3}(0, \sigma_B^2)$, where each voxel's standard deviation σ_B was randomly drawn from a uniform distribution $\mathcal{U}(0, b_B)$. The final synthetic bias field B was generated by taking the voxel-wise exponential after upsampling B' using a cubic spline interpolation, to ensure local smoothness and intensity scaling values on the range of $[0, 1]$. The synthetic field B spatially scales the uncorrupted image G to produce a bias-corrupted UltimateSynth image, $G_B = G \times B$.

Contrast Optimization MRI contrast optimization for manual labeling involves adjusting imaging parameters to enhance the visibility of anatomical structures or abnormalities in the images. This process is crucial when performing manual labeling of MRI data, as it helps annotators accurately identify and delineate regions of interest. Using UltimateSynth, we derive an automated process for generating images with maximized relative tissue intensities for a set of structure labels of interest $\mathcal{L}$:

$$\mathcal{L} = \{\text{WM}, \text{GM}, \text{CSF}\}, \tag{5}$$

defining white matter (WM) as T1=850 ms, T2=40 ms, gray matter (GM) as T1=1,400 ms, T2=70 ms, and cerebrospinal fluid (CSF) as T1=3,000 ms, T2=300 ms. These structure labels correspond to three rows of eigencontrast matrix D, where the three values in each column represent the voxel intensities of these principle tissues for a given combination of extrinsic parameters, Λ_i. For a given column i of D, we can compare the set of voxel intensities of the structure labels $I_{\mathcal{L}}$ as inter-tissue contrast distances:

$$\begin{aligned} d_l^{\min} &= \texttt{min\_dist}(I_l, I_{\mathcal{L}-\{l\}}), \quad l \in \mathcal{L}, \\ d_l^{\max} &= \texttt{max\_dist}(I_l, I_{\mathcal{L}-\{l\}}), \quad l \in \mathcal{L}, \end{aligned} \tag{6}$$

and use this extreme maximum intensity difference to define the contrast range R of the tissues of interest:

$$R = \max_l \left(d_l^{\max}\right). \tag{7}$$

From here, the previously computed minimum distance $d_l^{\min}$ and R can be used to optimize for a single tissue of interest compared to the other two tissues in (8), or maximizing an equal distance between all three tissues to optimize for a contrast with clear distinctions between all tissues in the set $\mathcal{L}$ in (9):

- Maximizing the relative contrast between a single tissue and two other tissues:

$$C_l = \frac{d_l^{\min}}{R}, \quad C_l \in [0, 1]. \tag{8}$$

- Optimizing for even voxel intensity differences between all three tissue classes:

$$C = \frac{|\mathcal{L}| - 1}{R} \min_l d_l^{\min}, \quad C \in [0, 1]. \tag{9}$$

Computing C (or C_l for the single-tissue case) across all possible Λ and finding the column Λ_{max} where C is maximized thus identifies the largest possible contrast for those tissues, given the exhaustive range of Λ.

Furthermore, the relative intensities of the brain structure labels when sorted in monotonically decreasing order can be considered to observe different classes of similar UltimateSynth images. For the $\mathcal{L}$ defined in (5) with three structure labels of interest, there exist six unique permutations of decreasing relative signal intensity between tissues, and thus six UltimateSynth image contrast classes:

- $I_{CSF} > I_{GM} > I_{WM}$
- $I_{GM} > I_{CSF} > I_{WM}$
- $I_{CSF} > I_{WM} > I_{GM}$
- $I_{WM} > I_{CSF} > I_{GM}$
- $I_{WM} > I_{GM} > I_{CSF}$
- $I_{GM} > I_{WM} > I_{CSF}$

Data Processing and Consensus Labeling The T1-weighted-like region of eigencontrast matrix D (WM > GM > CSF w.r.t. relative tissue voxel intensities) was uniformly sampled, and each location's multi-tissue relative contrast according to (9) was recorded. The region was divided into two subsections according to the 60th percentile of C across all sampled locations. Eight coefficient vectors randomly selected from the higher-contrast subsection were used to generate eight T1-weighted images per subject with varying inter-tissue intensity differences while all being T1-weighted. These synthetic T1-weighted images were normalized between the 1st and 99th voxel intensity percentiles on a per-image basis and processed using the `recon-all` pipeline of FreeSurfer[18], including skull stripping and tissue parcellation. For non-T1-weighted images, the FSL Brain Extraction Tool[37] was used for skull stripping. The resulting skull-stripped brain masks were dilated using a spherical structural element of radius 3, to ensure the mask encompassed all soft brain tissues. Parcellation labels were condensed into 18 regions of interest, plus two additional non-brain labels for non-zero intensity voxels unlabeled by FreeSurfer. For each subject, voxel-wise majority voting across the eight FreeSurfer segmentations was performed to create a single consensus segmentation. Consensus labeling reconciles differences between multiple labels to ensure labeling quality and reliability.

Labeling Outside the Brain It is common for some unlabeled non-brain tissues, such as the meninges and CSF in the subarachnoid space, to still be present in the skull-stripped brain mask. To reduce confusion in network training, we augment the 18-label condensed FreeSurfer structure label set with two additional labels: a 19th "subarachnoid space" label assigned to

unlabeled nonzero intensity voxels in the brain mask and a 20th "skull, scalp, and other tissues" label assigned to nonzero intensity voxels outside the brain mask. Non-skull-stripped whole images where used for all train-time and test-time segmentation tasks, with binary brain masks as defined above being applied to the resulting label maps, so as to focus on segmentation performance in the primary region of interest.

Cerebellum Labeling The cerebellum was manually labeled as WM, GM, and CSF using human-in-the-loop semi-supervised learning. Two manually labeled samples were used to train a segmentation network, which was then applied to label initially unlabeled samples. The machine-predicted labels were manually inspected and corrected before being included to augment the training data for network retraining. Samples with poor quality labels were excluded from training. This process was iterated to progressively label more samples and was halted when segmentation converged to a satisfactory outcome.

UltBrainNet Training Two hundred 1 mm isotropic resolution UltimateSynth images were prepared per training subject by randomly sampling the contrast space D as described in Efficient Contrast Sampling after reference labeling FreeSurfer labels as in Data Processing and Consensus Labeling. Per subject, these 200 images were evenly divided into four categories, each of 50 UltimateSynth images: (1) whole-brain uncorrupted images, (2) whole-brain bias-corrupted images as described in Bias Field, (3) uncorrupted subcortical structure images, and (4) bias-corrupted subcortical structure images. For categories 3 and 4, the whole-brain UltimateSynth images are cropped by a bounding box encompassing the non-cortical, non-cerebellar structure labels. This image modification helps data labeling imbalance during training, as the smallest subcortical labels (choroid plexus) encompass up to 1,000 times smaller volumes than the largest labels (gray matter). In total, this data preparation results in a training set of 6,400 UltimateSynth images encompassing all manner of MR contrasts. A 3d_fullres nnUNetv2 was trained with a summed Cross Entropy and DICE loss comparing network predictions to eight-repeat majority voted FreeSurfer reference labels. Training was performed for 1,000 epochs on the CWRU HPC using an Nvidia V100 GPU (32GB memory), taking a total of 52.4 hours.

The resulting UltBrainNet was evaluated on 155,520 UltimateSynth images evenly distributed throughout the contrast space generated from 8 unseen test subjects' MRF scan data to assess same-anatomy segmentation performance across extreme contrast variations. The network was also tested on the ON-Harmony traveling heads dataset, the Baby Connectome Project dataset, 1,152 UltimateSynth images derived from quantitative parameter maps from an additional 6-month-old, and the FreeSurfer Maintenance dataset, to assess generalizability across different subjects, ages, and scanners.

UltCerebNet Training Two hundred UltimateSynth images were each prepared for two training subjects with manually delineated reference labels by randomly sampling the contrast space D as described in Efficient Contrast Sampling. Per subject, the 200 images were divided into two categories, each with 100 UltimateSynth images: (1) whole-brain uncorrupted images and (2) whole-brain bias-corrupted images as described in Bias Field, Images were then linearly upsampled to 0.5 mm isotropic resolution and cropped around the cerebellar white matter, cerebellar gray matter, and cerebellar CSF labels as indicated by the manual segmentations. The resulting set of 400 training images were used to train a 3d_fullres nnUNetv2 trained using a summed Cross

Entropy and Dice loss between predictions and manually annotated references for 1,000 epochs on the CWRU HPC using an Nvidia V100 GPU (32GB memory), taking a total of 27.8 hours.

This initial UltimateSynth-based Cerebellum segmentation network was test-time tasked with segmenting the cerebellums of 29 additional training subjects for whom high-quality reference labels were not available (one subject excluded due to exceptionally poor cerebellum qMR maps), using image contrasts optimized between white matter and gray matter as defined in Contrast Optimization. After visual quality inspections, these test-time label predictions were then treated as reference labels, and 200 UltimateSynth images were randomly generated per additional training subject as before, for an expanded training dataset of 6,200 total cerebellum images. This dataset was employed to train the 3d_fullres nnUNetv2 model that we deem UltCerebNet for 1,000 epochs on the CWRU HPC using an Nvidia V100 GPU (32GB memory), taking a total of 57.23 hours.

Pan-Contrast Completeness of UltimateSynth Training Data While the qualitative and quantitative results presented here demonstrate that UltimateSynth-trained networks achieve pan-contrast segmentation capabilities, the mechanism for this ability could be explained as learning an insensitivity to inter-tissue contrasts or some other process through which an incomplete representation of MR image contrast in the training data is extrapolated to out-of-distribution predictions during test time. To demonstrate that UltBrainNet and other UltimateSynth-based networks learn contrast invariance through truly pan-contrast examples during training, we observe the voxel intensity features of the 3,200 whole-brain images in UltBrainNet's training dataset, as well as 575 images used for testing.

For each image, non-background voxels are normalized to the 1st and 99th percentiles and the minimum and maximum voxel intensities are set to 0 and 255, for fair comparisons across all images. Then, for each of the 18 structure labels in each image given the reference segmentation (or UltBrainNet segmentation if reference is unavailable), 15 first-order intensity-based Radiomics features are calculated, as defined by PyRadiomics[38]: energy, minimum, 10th percentile, 90th percentile, maximum, mean, median, interquartile range, range, mean absolute deviation, robust mean absolute deviation, root mean squared, standard deviation, skewness, and kurtosis. This results in 270 radiomics features per image which, when compared against each other in Supplementary Fig. S13, summarize the voxel intensity differences between different tissues in the image, and thus describe the image's contrast.

Evaluation Metrics Segmentation performance was quantified using a volume-weighted Sørensen-Dice coefficient:

$$\text{Dice} = \sum_{k=1}^{K} \left(\frac{N_{X,k}}{N_X} \right) \left(\frac{2|X_k \cap Y_k|}{|X_k| + |Y_k|} \right), \tag{10}$$

where K is the number of labels in images X and Y, N_X is the number of voxels in X, $N_{X,k}$ is the number of voxels assigned label k in X, and X_k and Y_k are binary equivalents of X and Y for k.

We additionally introduce two quantitative evaluation tools designed to leverage UltimateSynth's ability to create variable-contrast images from the same underlying anatomy, with different spatial scales of interest: Percent Disagreement from the Majority Vote (PDM) which is a voxel-wise metric of segmentation inconsistency, and mean absolute Label Volume Variation

(LVV), which is a label-wise measure of segmentation consistency.

To define PDM, let us consider a set of images $\{1, 2, 3, \ldots, J\}$, with each image composed of N voxels and each voxel assigned one discrete label k from a set of $\{1, 2, 3, \ldots, K\}$ possible labels. For the nth voxel shared across all images, there exists a majority vote MV_n

$$\mathrm{MV}_n = \texttt{mode}\left[k_{n,1}, k_{n,2}, \ldots, k_{n,J}\right] \tag{11}$$

and the PDM for voxel n is

$$\mathrm{PDM}_n = \frac{1}{J}\left\{J - \sum_{j=1}^{J}\left[k_{n,j} = \mathrm{MV}_n\right]\right\} \times 100\%. \tag{12}$$

For LVV, consider the volume of a given label k in an image j,

$$\mathrm{Vol}_{k,j} = \mathrm{Vol}_{\mathrm{vox}} \sum_{n=1}^{N}\left[k_n = k\right], \tag{13}$$

where $\mathrm{Vol}_{\mathrm{vox}}$ is the volume of one voxel. Then, the mean volume of label k across the set of J images is

$$\mathrm{Vol}_{\mathrm{mean},k} = \frac{1}{J}\sum_{j=1}^{J}\mathrm{Vol}_{k,j} \tag{14}$$

and the percent mean label volume variation (LVV) is

$$\mathrm{LVV} = \frac{1}{J}\sum_{j=1}^{J}\left|\mathrm{Vol}_{k,j} - \mathrm{Vol}_{\mathrm{mean},k}\right| \times 100\%. \tag{15}$$

Visualization of Quantitative Metrics In Figs. 4, 5, 6, 7, S2, S5, S6, S7, S10, and S12, data of quantitative segmentation performance metrics are presented as violin plots[39,40] to better visualize data distributions. In each violin, the data median is depicted as a circle and the tops and bottoms of the darkly shaded regions indicate the first and third quartiles of the data, respectively. Violin width indicates the density of data points around that value as indicated on each plot's vertical axis, with dense regions in the data corresponding to locally wide violins and values sparsely represented in the data corresponding to narrow violins.

Statistical Analysis For all same-subject repeat segmentations, two-sided Welch's unequal variances t-tests[41] were performed between quantitative Dice and Volume Variation metrics between the investigated UltimateSynth models and SOTA alternatives as a test of statistically significant improvement over the status quo segmentation performance. In all tests, statistical significance was defined using a P value threshold of .05. For experiments where tests were performed between more than two methods, P value thresholds were Bonferroni corrected. When reporting numerical performances, results are presented as the Mean $\pm$ the Standard Deviation of the data.

Supplementary Information The manuscript contains supplementary material.

Acknowledgements This work was supported in part by the United States National Institutes of Health (NIH) under grants R01 CA269604 (D. Ma), R01 CA282516 (D. Ma), R01 NS109439 (D. Ma), R01 EB035160 (P.-T. Yap), R01 MH125479 (P.-T. Yap), R01 EB008374 (P.-T. Yap), R01 NS134849 (P.-T. Yap), and R01 HD112923 (D. Ma and P.-T. Yap).

Author Contributions R.A.: methodology, software, investigation, visualization, writing - original draft, writing – review and editing. W.Z.: methodology, resources. S.H.: resources, data curation. W.L.: resources, data curation. K.M.H.: resources, data curation, visualization. S.A.: resources, data curation, visualization. D.M.: conceptualization, methodology, supervision, funding acquisition, investigation, validation, writing – review and editing. P.-T.Y.: conceptualization, methodology, supervision, funding acquisition, investigation, validation, writing – review and editing.

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to D.M. and P.-T.Y.

1. Huppertz, H.-J., Kröll-Seger, J., Klöppel, S., Ganz, R. E. & Kassubek, J. Intra- and inter-scanner variability of automated voxel-based volumetry based on a 3D probabilistic atlas of human cerebral structures. *NeuroImage* **49**, 2216–2224 (2010). URL https://linkinghub.elsevier.com/retrieve/pii/S1053811909011379.

2. Fortin, J.-P., Sweeney, E. M., Muschelli, J., Crainiceanu, C. M. & Shinohara, R. T. Removing inter-subject technical variability in magnetic resonance imaging studies. *NeuroImage* **132**, 198–212 (2016). URL https://linkinghub.elsevier.com/retrieve/pii/S1053811916001452.

3. Hedges, E. P. *et al.* Reliability of structural MRI measurements: The effects of scan session, head tilt, inter-scan interval, acquisition sequence, FreeSurfer version and processing stream. *NeuroImage* **246**, 118751 (2022). URL https://linkinghub.elsevier.com/retrieve/pii/S1053811921010235.

4. Filippi, M. *et al.* Interscanner variation in brain MRI lesion load measurements in MS: Implications for clinical trials. *Neurology* **49**, 371–377 (1997). URL https://www.neurology.org/doi/10.1212/WNL.49.2.371.

5. Clark, K. A., Woods, R. P., Rottenberg, D. A., Toga, A. W. & Mazziotta, J. C. Impact of acquisition protocols and processing streams on tissue segmentation of T1 weighted MR images. *NeuroImage* **29**, 185–202 (2006).

6. Carré, A. *et al.* Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics. *Scientific Reports* **10**, 12340 (2020). URL https://www.nature.com/articles/s41598-020-69298-z.

7. Vasiliuk, A., Frolova, D., Belyaev, M. & Shirokikh, B. Limitations of Out-of-Distribution Detection in 3D Medical Image Segmentation. *Journal of Imaging* **9**, 191 (2023). URL https://www.mdpi.com/2313-433X/9/9/191.

8. Mårtensson, G. *et al.* The reliability of a deep learning model in clinical out-of-distribution MRI data: A multicohort study. *Medical Image Analysis* **66**, 101714 (2020). URL https://linkinghub.elsevier.com/retrieve/pii/S1361841520300785.

9. Billot, B. *et al.* SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Medical Image Analysis* **86**, 102789 (2023). URL https://linkinghub.elsevier.com/retrieve/pii/S1361841523000506.

10. Billot, B. *et al.* Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain MRI datasets. *Proceedings of the National Academy of Sciences* **120**, e2216399120 (2023). URL https://pnas.org/doi/10.1073/pnas.2216399120.

11. Iglesias, J. E. *et al.* Joint super-resolution and synthesis of 1 mm isotropic MP-RAGE volumes from clinical MRI exams with scans of different orientation, resolution and contrast. *NeuroImage* **237**, 118206 (2021). URL https://linkinghub.elsevier.com/retrieve/pii/S1053811921004833.

12. Borges, P. *et al.* Acquisition-invariant brain MRI segmentation with informative uncertainties. *Medical Image Analysis* **92**, 103058 (2024). URL https://linkinghub.elsevier.com/retrieve/pii/S1361841523003183.

13. Kumar, S., Saber, H., Charron, O., Freeman, L. & Tamir, J. I. Correcting synthetic MRI contrast-weighted images using deep learning. *Magnetic Resonance Imaging* **106**, 43–54 (2024).

14. Jog, A., Hoopes, A., Greve, D. N., Van Leemput, K. & Fischl, B. PSACNN: Pulse sequence adaptive fast whole brain segmentation. *NeuroImage* **199**, 553–569 (2019). URL https://linkinghub.elsevier.com/retrieve/pii/S1053811919304252.

15. Qiu, S. *et al.* Direct synthesis of multi-contrast brain MR images from MR multitasking spatial factors using deep learning. *Magnetic Resonance in Medicine* **90**, 1672–1681 (2023). URL https://onlinelibrary.wiley.com/doi/10.1002/mrm.29715.

16. Jacobs, L. *et al.* Generalizable synthetic MRI with physics-informed convolutional networks (2023). URL http://arxiv.org/abs/2305.12570. ArXiv:2305.12570 [physics].

17. Wang, K. *et al.* High-fidelity direct contrast synthesis from magnetic resonance fingerprinting. *Magnetic Resonance in Medicine* **90**, 2116–2129 (2023). URL https://onlinelibrary.wiley.com/doi/10.1002/mrm.29766.

18. Fischl, B. FreeSurfer. *NeuroImage* **62**, 774–781 (2012). URL https://linkinghub.elsevier.com/retrieve/pii/S1053811912000389.

19. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* **18**, 203–211 (2021). URL https://www.nature.com/articles/s41592-020-01008-z.

20. Warrington, S. *et al.* A resource for development and comparison of multi-modal brain 3 T MRI harmonisation approaches. *Imaging Neuroscience* **1**, 1–27 (2023). URL https://direct.mit.edu/imag/article/doi/10.1162/imag_a_00042/118214/A-resource-for-development-and-comparison-of.

21. Howell, B. R. *et al.* The UNC/UMN Baby Connectome Project (BCP): An overview of the study design and protocol development. *NeuroImage* **185**, 891–905 (2019). URL https://linkinghub.elsevier.com/retrieve/pii/S1053811918302593.

22. Greve, D. N. & Fischl, B. The FreeSurfer Maintenance Dataset (2024). URL https://openneuro.org/datasets/ds004958/versions/1.0.0.

23. Yang, M. *et al.* Low rank approximation methods for MR fingerprinting with large scale dictionaries. *Magnetic Resonance in Medicine* **79**, 2392–2400 (2018). URL https://onlinelibrary.wiley.com/doi/10.1002/mrm.26867.

24. Ahmad, S. *et al.* Multifaceted atlases of the human brain in its infancy. *Nature Methods* **20**, 55–64. URL https://www.nature.com/articles/s41592-022-01703-z.

25. Zöllei, L., Iglesias, J. E., Ou, Y., Grant, P. E. & Fischl, B. Infant FreeSurfer: An automated segmentation and surface extraction pipeline for T1-weighted neuroimaging data of infants 0–2 years. *NeuroImage* **218**, 116946 (2020). URL https://linkinghub.elsevier.com/retrieve/pii/S1053811920304328.

26. Van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *Journal of machine learning research* **9** (2008).

27. Ma, D. *et al.* Magnetic resonance fingerprinting. *Nature* **495**, 187–192 (2013). URL https://www.nature.com/articles/nature11971.

28. Ma, D. *et al.* Fast 3D magnetic resonance fingerprinting for a whole-brain coverage. *Magnetic Resonance in Medicine* **79**, 2190–2197 (2018). URL https://onlinelibrary.wiley.com/doi/10.1002/mrm.26886.

29. Cao, X. *et al.* DTI-MR fingerprinting for rapid high-resolution whole-brain T_1, T_2, proton density, ADC, and fractional anisotropy mapping. *Magnetic Resonance in Medicine* **91**, 987–1001 (2024). URL https://onlinelibrary.wiley.com/doi/10.1002/mrm.29916.

30. Ma, D. *et al.* Development of High-Resolution 3D MR Fingerprinting for Detection and Characterization of Epileptic Lesions. *Journal of Magnetic Resonance Imaging* **49**, 1333–1346 (2019). URL https://onlinelibrary.wiley.com/doi/full/10.1002/jmri.26319.

31. Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage* **17**, 825–841 (2002). URL https://linkinghub.elsevier.com/retrieve/pii/S1053811902911328.

32. Jenkinson, M. & Smith, S. A global optimisation method for robust affine registration of brain images. *Medical Image Analysis* **5**, 143–156 (2001). URL https://linkinghub.elsevier.com/retrieve/pii/S1361841501000366.

33. Greve, D. N. & Fischl, B. Accurate and robust brain image alignment using boundary-based registration. *NeuroImage* **48**, 63–72 (2009). URL https://linkinghub.elsevier.com/retrieve/pii/S1053811909006752.

34. Ma, D. *et al.* Motion Robust MR Fingerprinting Scan to Image Neonates With Prenatal Opioid Exposure. *Journal of Magnetic Resonance Imaging* **59**, 1758–1768 (2024). URL https://onlinelibrary.wiley.com/doi/10.1002/jmri.28907.

35. Bloch, F. Nuclear Induction. *Physical Review* **70**, 460–474 (1946). URL https://link.aps.org/doi/10.1103/PhysRev.70.460.

36. Zhao, B. *et al.* Optimal Experiment Design for Magnetic Resonance Fingerprinting: Cramér-Rao Bound Meets Spin Dynamics. *IEEE Transactions on Medical Imaging* **38**, 844–861 (2019). URL https://ieeexplore.ieee.org/document/8481484/.

37. Jenkinson, M. BET2: MR-based estimation of brain, skull and scalp surfaces. *Annual Meeting of the Organization for Human Brain Mapping (OHBM)* **17**, 167 (2005). URL https://cir.nii.ac.jp/crid/1573950400559824000.

38. Van Griethuysen, J. J. *et al.* Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research* **77**, e104–e107 (2017). URL https://aacrjournals.org/cancerres/article/77/21/e104/662617/Computational-Radiomics-System-to-Decode-the.

39. Hintze, J. L. & Nelson, R. D. Violin Plots: A Box Plot-Density Trace Synergism. *The American Statistician* **52**, 181–184 (1998). URL http://www.tandfonline.com/doi/abs/10.1080/00031305.1998.10480559.

40. Bechtold, B., Fletcher, P., Seamusholden & Gorur-Shandilya, S. Violin Plots for Matlab (2021). URL https://zenodo.org/record/4559847.

41. Welch, B. L. The generalization of 'student's' problem when several different population varlances are involved. *Biometrika* **34**, 28–35 (1947). URL https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/34.1-2.28.

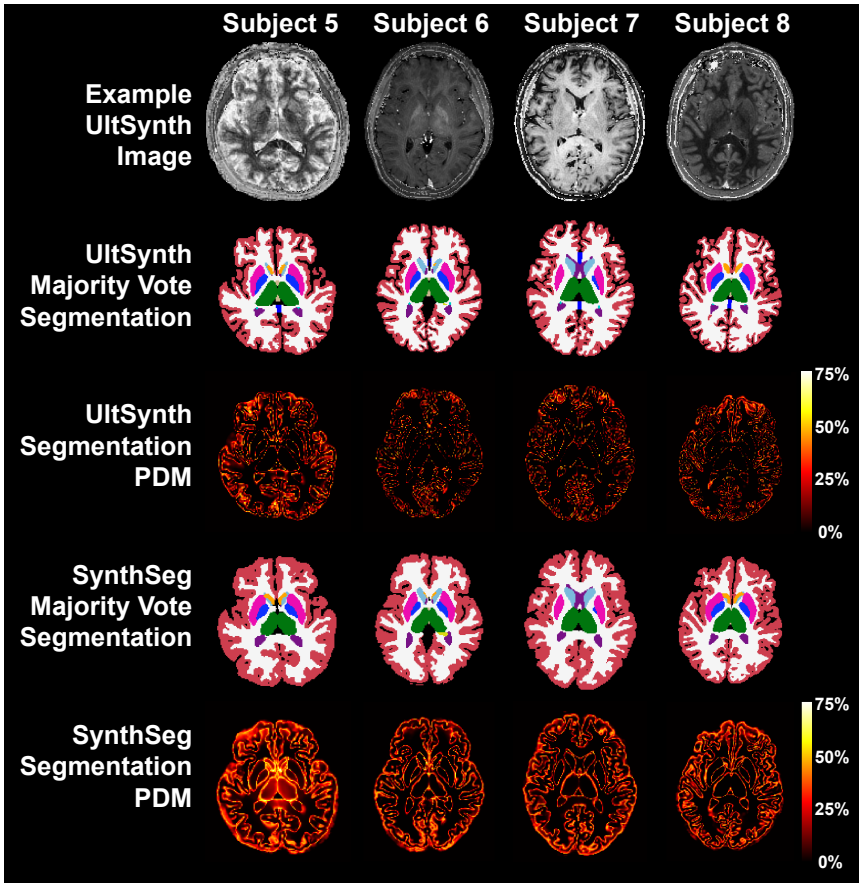

Fig. S1 | Example images, majority vote labels, and PDM values for four additional subjects used in the model performance evaluation for Figure 5. Trends from the first four subjects continue here, with UltimateSynth consistently demonstrating lower PDM and more refined predictions of gray matter compared with SynthSeg.

Tab. S1 | Structure-specific Dice scores and LVV values across 155,520 UltimateSynth images analyzed in Figures 5b and c. UltimateSynth performs significantly better than SynthSeg for both metrics and all labels ($P < .001$).

Structure Label	Mean Volume [%]	UltSynth Dice	SynthSeg Dice	US-SS Dice *P* Value	UltSynth Volume Variation [%]	SynthSeg Volume Variation [%]	US-SS Volume Variation *P* Value
White Matter	41.31	**0.93 ± 0.027**	0.85 ± 0.048	**< .001**	**1.01 ± 0.96**	4.15 ± 5.02	**< .001**
Grey Matter	39.09	**0.86 ± 0.043**	0.75 ± 0.049	**< .001**	**1.72 ± 1.28**	4.37 ± 3.76	**< .001**
Lateral Ventricle	1.21	**0.90 ± 0.049**	0.81 ± 0.098	**< .001**	**3.43 ± 2.97**	13.62 ± 12.10	**< .001**
Cerebellar White Matter	2.28	**0.90 ± 0.029**	0.81 ± 0.149	**< .001**	**1.86 ± 1.73**	15.83 ± 19.78	**< .001**
Cerebellar Grey Matter	8.91	**0.92 ± 0.024**	0.85 ± 0.071	**< .001**	**0.86 ± 0.72**	6.20 ± 9.98	**< .001**
Thalamus	1.32	**0.92 ± 0.019**	0.87 ± 0.118	**< .001**	**0.82 ± 0.76**	9.16 ± 15.00	**< .001**
Caudate	0.59	**0.90 ± 0.030**	0.83 ± 0.081	**< .001**	**1.57 ± 1.52**	7.66 ± 10.29	**< .001**
Putamen	0.88	**0.92 ± 0.017**	0.84 ± 0.054	**< .001**	**1.06 ± 0.92**	5.09 ± 6.98	**< .001**
Pallidum	0.32	**0.87 ± 0.032**	0.77 ± 0.085	**< .001**	**1.25 ± 1.07**	8.53 ± 10.65	**< .001**
3rd Ventricle	0.08	**0.82 ± 0.076**	0.60 ± 0.220	**< .001**	**3.11 ± 2.57**	40.30 ± 35.75	**< .001**
4th Ventricle	0.19	**0.86 ± 0.028**	0.73 ± 0.210	**< .001**	**3.45 ± 3.20**	28.08 ± 32.65	**< .001**
Brain Stem	1.73	**0.94 ± 0.014**	0.85 ± 0.141	**< .001**	**0.96 ± 0.81**	11.38 ± 17.47	**< .001**
Hippocampus	0.65	**0.89 ± 0.031**	0.81 ± 0.064	**< .001**	**1.44 ± 1.32**	7.10 ± 8.37	**< .001**
Amygdala	0.26	**0.87 ± 0.021**	0.76 ± 0.042	**< .001**	**1.46 ± 1.20**	6.65 ± 6.14	**< .001**
Accumbens Area	0.09	**0.78 ± 0.047**	0.63 ± 0.111	**< .001**	**2.18 ± 1.89**	13.75 ± 16.46	**< .001**
Ventral Diencephalon	0.72	**0.87 ± 0.041**	0.80 ± 0.120	**< .001**	**1.05 ± 0.90**	8.98 ± 14.55	**< .001**

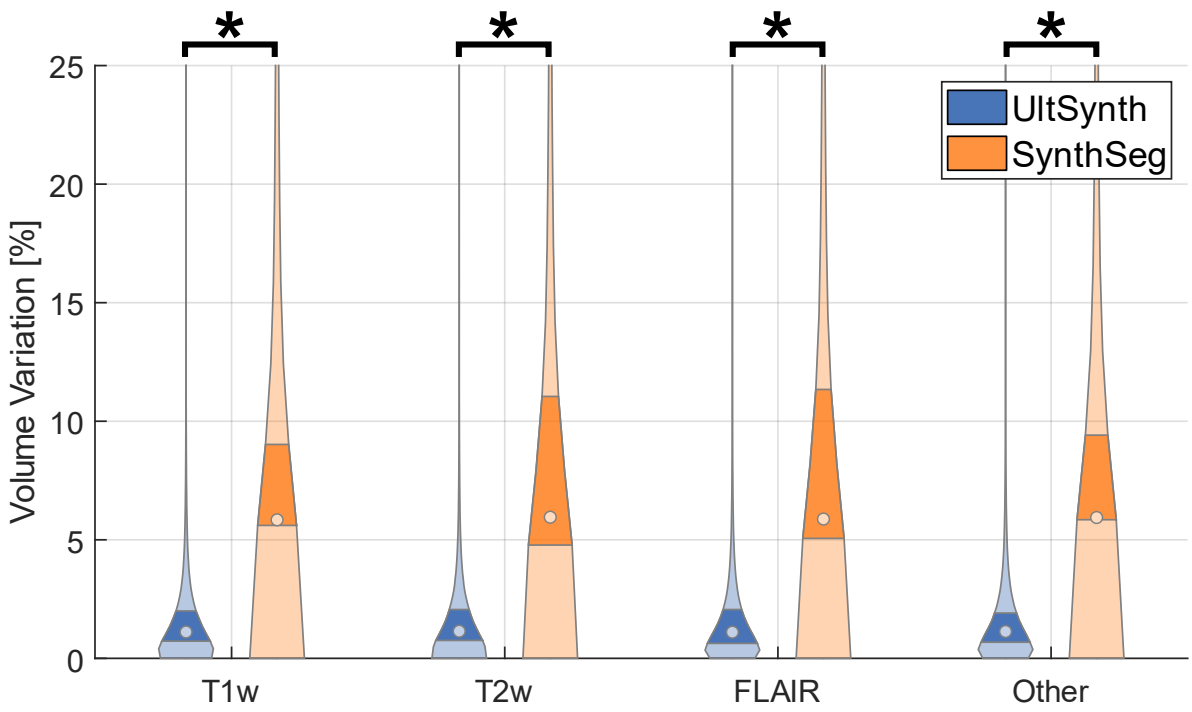

Fig. S2 | Brain segmentation performance for different MR scan types. We divided the 155,520 UltimateSynth image test set from Figure 5 into 48,280 T1-weighted images, 18,304 T2-weighted images, 48,936 FLAIR images, and 40,000 other contrast images, and re-examined the label volume variation (LVV) for different MR scan types. UltimateSynth never exceeds 1.92% mean variation.

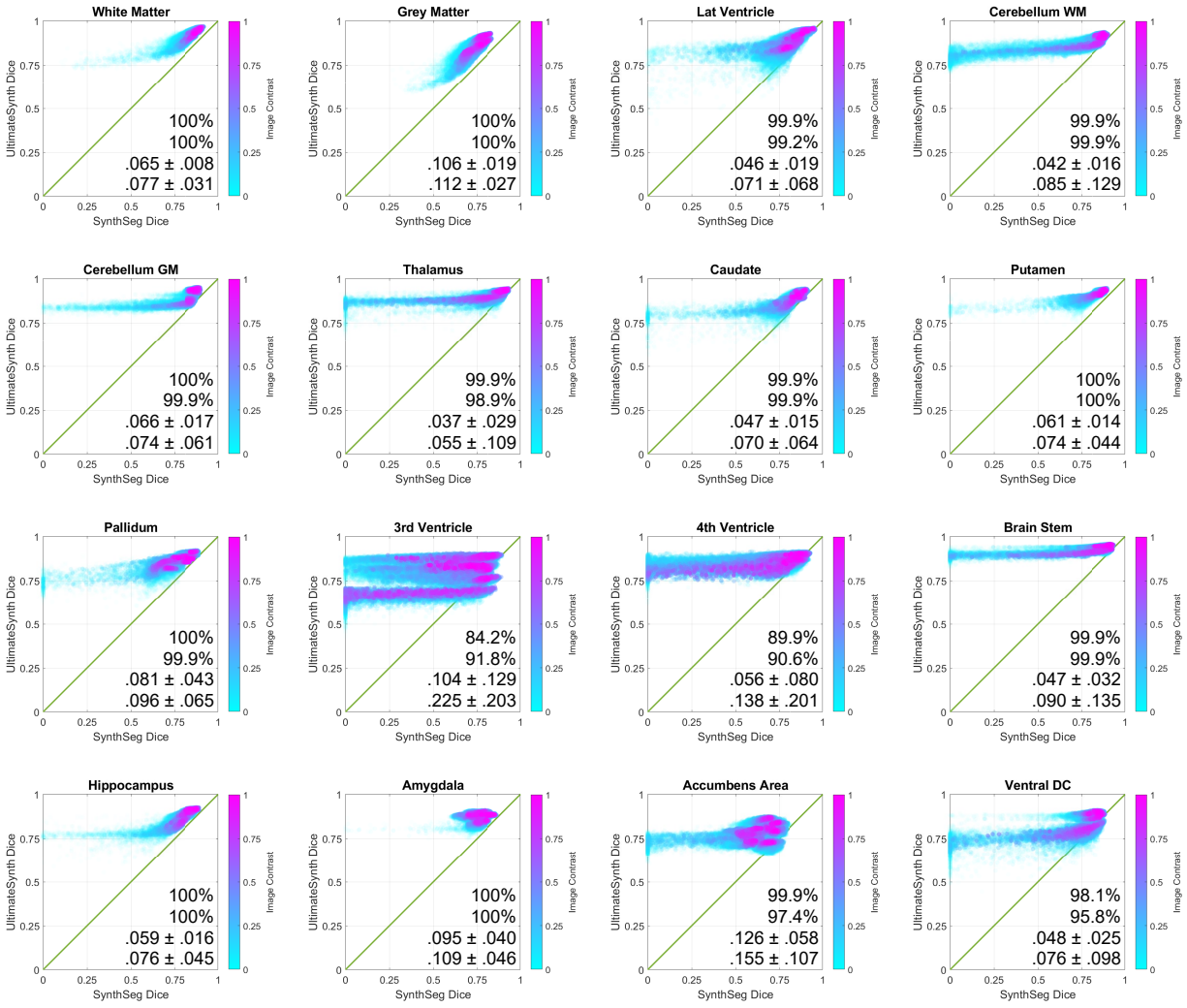

Fig. S3 | Image-wise comparison of Dice scores between UltimateSynth and SynthSeg on 155,520 synthetic pan-contrast images. UltimateSynth generally outperforms SynthSeg, especially for white matter and gray matter and small structures like the putamen, pallidum, and amygdala. For each label, the percent of images where UltimateSynth exceeds SynthSeg in Dice scores are reported for images with Image Contrast >0.5 (top) and <0.5 (bottom). Metrics are similarly reported for the mean and standard deviation of Dice score differences across images per label for each contrast class. Y = X plotted in green for visual reference.

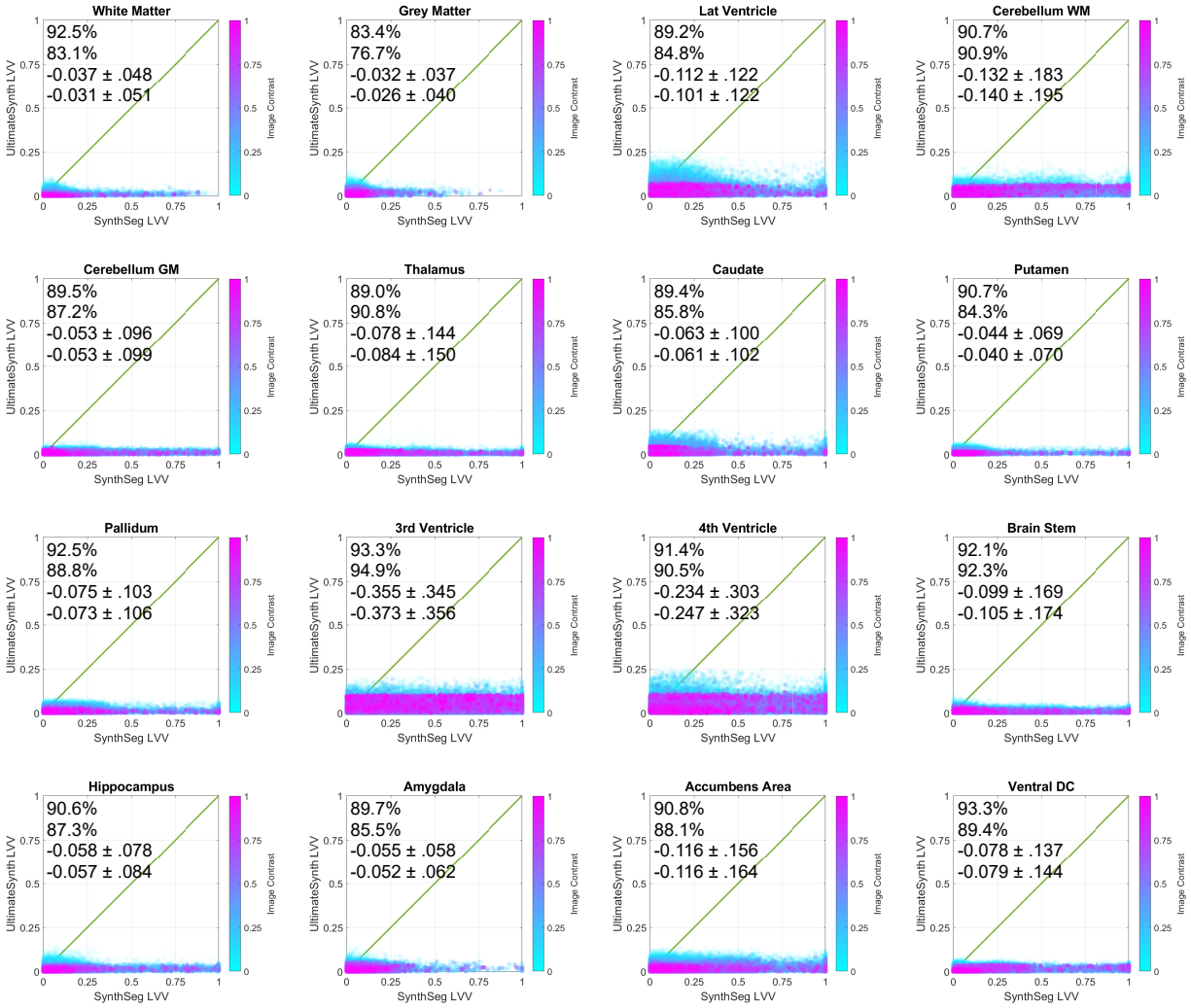

Fig. S4 | Image-wise comparison of LVV between UltimateSynth and SynthSeg on 155,520 synthetic pan-contrast images. Across all labels, UltimateSynth tends to have low LVV compared to SynthSeg. For each label, the percent of images where UltimateSynth LVV is less than SynthSeg LVV are reported for images with Image Contrast >0.5 (top) and <0.5 (bottom). Metrics are similarly reported for the mean and standard deviation of LVV across images per label for each contrast class. Y = X plotted in green for visual reference.

Tab. S2 | LVV specific to MR scan types in 155,520 UltimateSynth images from Figure S2. UltimateSynth performs significantly better than SynthSeg in all four MR scan types ($P < .001$).

MR Scan Type	UltSynth Volume Variation [%]	SynthSeg Volume Variation [%]	Volume Variation Difference *P*-Value
T1w	**1.68 ± 1.88**	11.79 ± 19.03	**< .001**
T2w	**1.71 ± 1.87**	11.81 ± 18.80	**< .001**
T2-FLAIR	**1.67 ± 1.86**	11.73 ± 18.82	**< .001**
Other	**1.70 ± 1.89**	12.02 ± 19.33	**< .001**

Tab. S3 | Structure-specific LVV across 3,456 UltimateSynth cerebellum images from Figure 6. UltimateSynth performs significantly better than SynthSeg in both primary cerebellar tissues ($P < .001$).

Structure Label	UltSynth Volume Variation [%]	SynthSeg Volume Variation [%]	Volume Variation Difference *P*-Value
Cerebellar White Matter	**0.45 ± 0.39**	8.62 ± 8.94	**< .001**
Cerebellar Gray Matter	**0.75 ± 0.59**	3.88 ± 3.13	**< .001**

Tab. S4 | Whole-brain weighted Dice scores across 160 T1-weighted and FLAIR scans averaged over 10 subjects from the ON-Harmony dataset in Figure 7b. UltimateSynth and SynthSeg have numerically comparable Dice, whereas PhysSeg lags significantly behind ($P < .001$).

Subject Number	UltSynth Dice	SynthSeg Dice	PhysSeg Dice	US-SS Dice Difference *P*-Value	US-PS Dice Difference *P*-Value	SS-PS Dice Difference *P*-Value
1	0.91 ± 0.02	0.89 ± 0.01	0.74 ± 0.01	.02	**< .001**	**< .001**
2	0.89 ± 0.02	0.88 ± 0.01	0.76 ± 0.01	.10	**< .001**	**< .001**
3	0.90 ± 0.02	0.90 ± 0.02	0.79 ± 0.01	0.52	**< .001**	**< .001**
4	0.91 ± 0.02	0.89 ± 0.01	0.78 ± 0.01	.02	**< .001**	**< .001**
5	**0.90 ± 0.01**	0.88 ± 0.01	0.75 ± 0.004	**< .001**	**< .001**	**< .001**
6	0.91 ± 0.02	0.90 ± 0.01	0.77 ± 0.004	.03	**< .001**	**< .001**
7	0.89 ± 0.02	0.88 ± 0.01	0.78 ± 0.01	.15	**< .001**	**< .001**
8	**0.90 ± 0.01**	0.88 ± 0.01	0.75 ± 0.004	**< .001**	**< .001**	**< .001**
9	0.90 ± 0.01	0.90 ± 0.01	0.78 ± 0.01	.11	**< .001**	**< .001**
10	0.90 ± 0.02	0.88 ± 0.01	0.76 ± 0.01	**.003**	**< .001**	**< .001**

Tab. S5 | Structure-specific LVV across 160 T1-weighted and FLAIR scans averaged over 10 subjects from the ON-Harmony dataset (Figure 7b). UltimateSynth demonstrates significantly reduced LVV over SynthSeg and PhysSeg in 10 out of 16 structure labels ($P < .001$).

Structure Label	Brain Volume [%]	UltSynth Volume Variation [%]	SynthSeg Volume Variation [%]	PhysSeg Volume Variation [%]	US-SS Volume Variation Difference *P*-Value	US-PS Volume Variation Difference *P*-Value	SS-PS Volume Variation Difference *P*-Value
White Matter	39.26	**1.76 ± 1.65**	2.40 ± 1.29	2.23 ± 2.33	**< .001**	.11	.55
Grey Matter	42.65	1.73 ± 1.01	2.23 ± 1.28	**1.28 ± 1.17**	**< .001**	**.004**	**< .001**
Lateral Ventricle	1.27	2.21 ± 1.76	2.40 ± 1.90	-	.37	-	-
Cerebellar White Matter	1.93	**2.35 ± 1.87**	9.26 ± 2.78	-	**< .001**	-	-
Cerebellar Grey Matter	6.89	2.18 ± 1.47	**1.41 ± 1.13**	-	**< .001**	-	-
Thalamus	1.41	**0.93 ± 0.72**	3.50 ± 2.35	-	**< .001**	-	-
Caudate	0.73	1.31 ± 1.27	1.41 ± 1.28	-	0.52	-	-
Putamen	1.02	**1.11 ± 0.91**	3.17 ± 2.54	-	**< .001**	-	-
Pallidum	0.40	**1.26 ± 1.08**	6.03 ± 4.42	-	**< .001**	-	-
3rd Ventricle	0.09	**3.45 ± 2.51**	9.80 ± 6.01	-	**< .001**	-	-
4th Ventricle	0.12	**4.84 ± 3.51**	15.64 ± 7.54	-	**< .001**	-	-
Brain Stem	1.79	1.45 ± 1.31	1.73 ± 1.22	-	.05	-	-
Hippocampus	0.73	**1.21 ± 1.05**	3.70 ± 2.32	-	**< .001**	-	-
Amygdala	0.35	**1.52 ± 1.29**	6.46 ± 4.03	-	**< .001**	-	-
Accumbens Area	0.12	**1.78 ± 1.38**	4.09 ± 3.03	-	**< .001**	-	-
Ventral Diencephalon	0.76	**1.30 ± 0.99**	3.30 ± 2.48	-	**< .001**	-	-

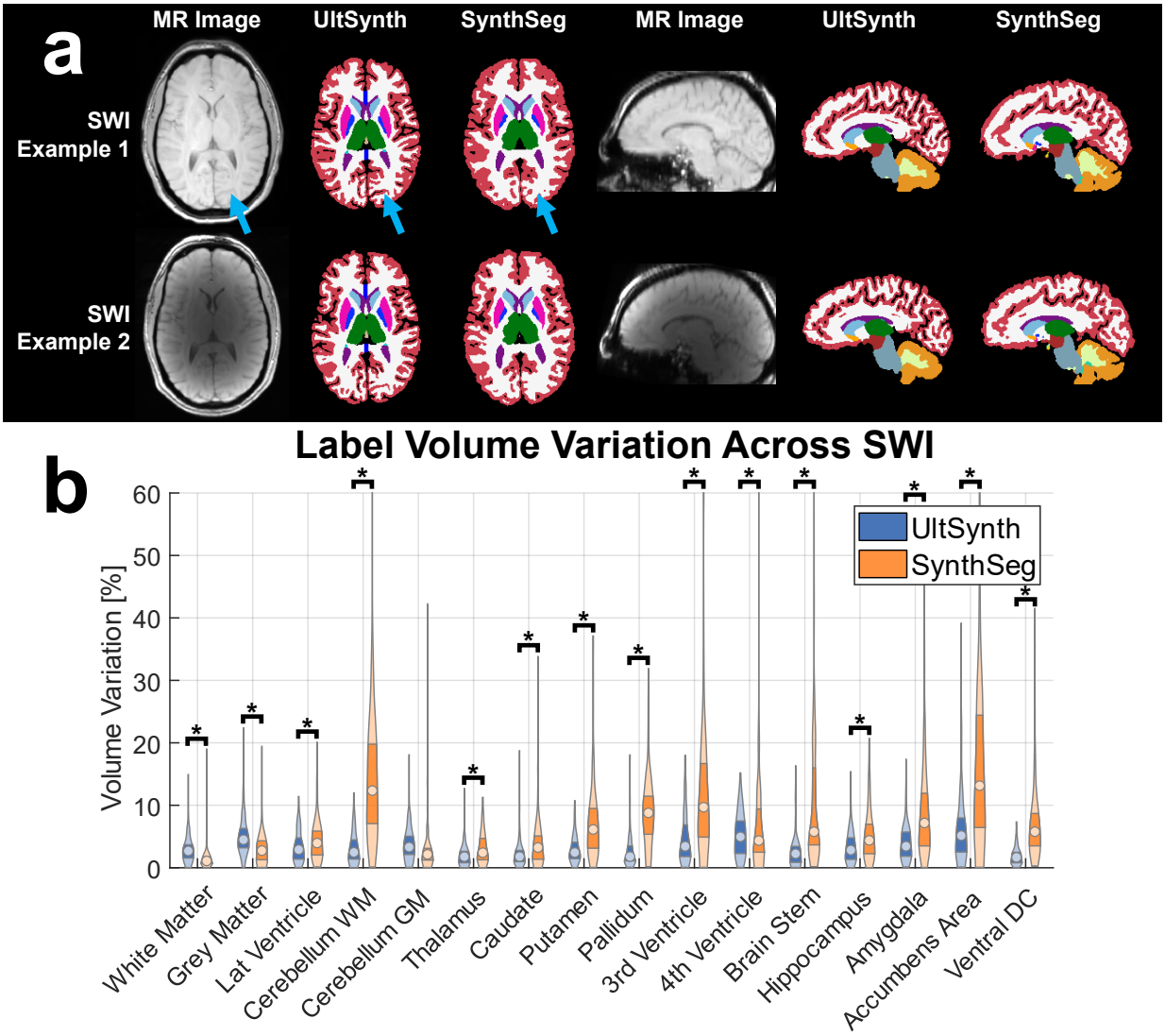

Fig. S5 | Susceptibility-weighted image segmentation examples and quantitative results across 160 ON-Harmony images. **a**, UltimateSynth retains cortical details despite the low-contrast nature of SWI scans (blue arrows). **b**, LVV for UltBrainNet and SynthSeg for the 160 SWI test images. UltimateSynth generally yields lower volume variation compared with SynthSeg.

Tab. S6 | Structure-specific LVV across 160 SWI images from 10 subjects in the ON-Harmony dataset, as presented in Figure S5. UltimateSynth has significantly lower LVV in 13 structure labels ($P < .05$) despite low image contrast and reduced through-plane resolution.

Structure Label	UltSynth Volume Variation [%]	SynthSeg Volume Variation [%]	US-SS Volume Variation Difference *P*-Value
White Matter	2.98 ± 2.24	**1.49 ± 2.13**	**< .001**
Grey Matter	5.09 ± 3.36	**2.99 ± 2.66**	**< .001**
Lateral Ventricle	**3.38 ± 2.53**	4.41 ± 3.33	**.002**
Cerebellar White Matter	**3.18 ± 2.65**	14.00 ± 10.08	**< .001**
Cerebellar Grey Matter	3.77 ± 2.86	3.07 ± 4.30	.09
Thalamus	**2.01 ± 1.84**	3.23 ± 2.66	**< .001**
Caudate	**2.27 ± 2.37**	4.01 ± 4.37	**< .001**
Putamen	**2.96 ± 2.20**	7.67 ± 6.78	**< .001**
Pallidum	**2.64 ± 2.68**	8.81 ± 5.54	**< .001**
3rd Ventricle	**4.79 ± 4.18**	13.43 ± 14.41	**< .001**
4th Ventricle	**5.34 ± 3.55**	6.78 ± 7.99	**.04**
Brain Stem	**2.57 ± 2.55**	10.98 ± 11.82	**< .001**
Hippocampus	**3.35 ± 2.64**	5.26 ± 4.26	**< .001**
Amygdala	**3.95 ± 2.82**	9.22 ± 8.06	**< .001**
Accumbens Area	**6.27 ± 5.67**	17.58 ± 16.89	**< .001**
Ventral Diencephalon	**1.94 ± 1.47**	7.16 ± 6.42	**< .001**

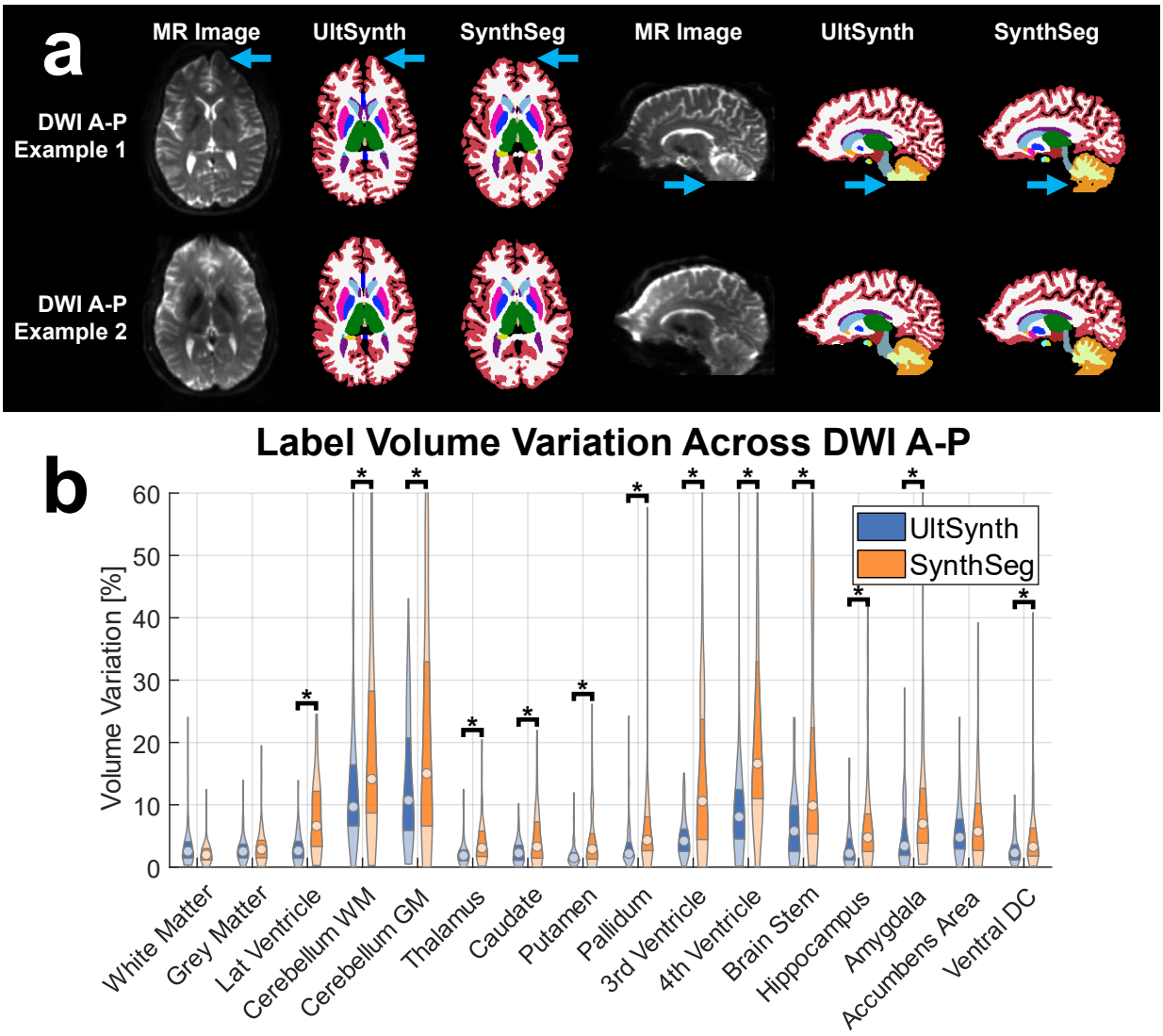

Fig. S6 | DWI A-P segmentation examples and quantitative results across 80 ON-Harmony images. **a**, UltimateSynth shows appreciable adaptability to frontal lobe distortions or cerebellum incompleteness compared to SynthSeg. **b**, LVV for UltBrainNet and SynthSeg for the 80 DWI test images. While geometric distortions associated with DWI mildly decrease UltimateSynth's segmentation performance for WM and GM in the cerebrum and cerebellum, numerical performance remains excellent.

Tab. S7 | Structure-specific LVV from 80 DWI A-P images spanning 10 subjects from the ON-Harmony dataset (Figure S6). Despite severe geometric distortions, UltimateSynth still yields significantly lower LVV than SynthSeg for 13 out of 16 labels ($P < .004$).

Structure Label	UltSynth Volume Variation [%]	SynthSeg Volume Variation [%]	US-SS Volume Variation Difference *P*-Value
White Matter	3.41 ± 3.51	2.69 ± 2.51	0.14
Grey Matter	2.85 ± 2.17	3.58 ± 3.66	0.13
Lateral Ventricle	**3.18 ± 2.65**	8.14 ± 6.00	**< .001**
Cerebellar White Matter	**13.92 ± 13.07**	23.35 ± 23.07	**.002**
Cerebellar Grey Matter	**13.72 ± 10.24**	21.50 ± 18.76	**.001**
Thalamus	**2.39 ± 2.40**	4.26 ± 4.11	**< .001**
Caudate	**2.51 ± 1.99**	4.95 ± 4.89	**< .001**
Putamen	**2.08 ± 2.08**	4.43 ± 4.80	**< .001**
Pallidum	**3.37 ± 3.53**	6.55 ± 8.54	**.003**
3rd Ventricle	**4.84 ± 3.41**	14.59 ± 14.70	**< .001**
4th Ventricle	**9.94 ± 10.75**	26.82 ± 26.00	**< .001**
Brain Stem	**6.84 ± 5.75**	18.09 ± 20.95	**< .001**
Hippocampus	**3.37 ± 3.66**	7.47 ± 8.57	**< .001**
Amygdala	**5.32 ± 5.48**	12.58 ± 15.26	**< .001**
Accumbens Area	5.73 ± 4.65	7.37 ± 7.35	.09
Ventral Diencephalon	**2.97 ± 2.49**	5.70 ± 6.97	**.001**

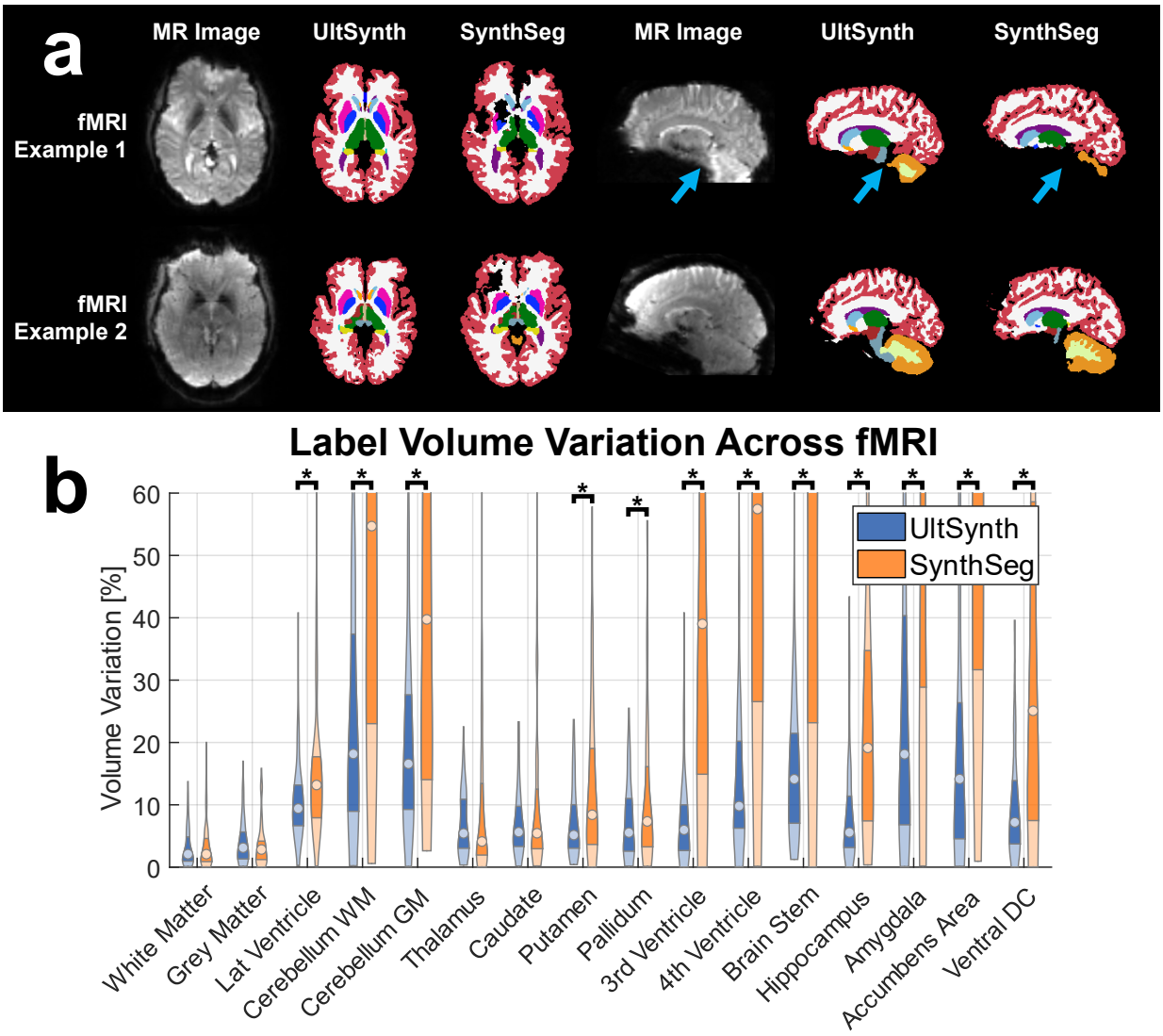

Fig. S7 | Resting fMRI segmentation examples and numerical results across 80 ON-Harmony images. **a**, UltimateSynth predicts plausible segmentations despite severe signal voids and geometric distortions (blue arrows). **b**, LVV for UltBrainNet and SynthSeg for the 80 fMRI test images. Due to poor image quality, both UltimateSynth and SynthSeg see increases in variation. SynthSeg's misprediction causes a great increase in variation, especially in small basal ganglia tissues.

Tab. S8 | Structure-specific LVV across 80 resting state fMRI images from 10 subjects in the ON-Harmony dataset, as seen in Figure S7. Low image resolution and signal dropout artifacts cause both UltimateSynth's and SOTA methods' LVV performance to mildly worsen. Despite this, the UltimateSynth creates labels that vary from the cross-image mean volume significantly less compared to SynthSeg's segmentations in 12 structure labels ($P < .03$).

Structure Label	UltSynth Volume Variation [%]	SynthSeg Volume Variation [%]	US-SS Volume Variation Difference *P*-Value
White Matter	3.23 ± 3.14	3.44 ± 3.83	0.70
Grey Matter	4.11 ± 3.66	3.62 ± 3.71	0.40
Lateral Ventricle	**10.68 ± 6.66**	15.36 ± 11.26	**.002**
Cerebellar White Matter	**24.71 ± 21.68**	57.92 ± 39.42	**< .001**
Cerebellar Grey Matter	**20.79 ± 17.34**	42.04 ± 28.99	**< .001**
Thalamus	6.94 ± 5.19	9.53 ± 12.16	0.08
Caudate	7.20 ± 5.62	10.08 ± 12.08	0.06
Putamen	**7.25 ± 5.86**	13.32 ± 12.72	**< .001**
Pallidum	**7.50 ± 6.16**	11.16 ± 12.26	**.02**
3rd Ventricle	**7.56 ± 7.25**	57.63 ± 62.60	**< .001**
4th Ventricle	**15.15 ± 15.99**	67.86 ± 54.58	**< .001**
Brain Stem	**15.49 ± 11.58**	63.77 ± 46.95	**< .001**
Hippocampus	**8.78 ± 8.94**	25.21 ± 21.65	**< .001**
Amygdala	**26.70 ± 28.19**	83.36 ± 95.96	**< .001**
Accumbens Area	**20.70 ± 22.13**	72.73 ± 62.67	**< .001**
Ventral Diencephalon	**9.56 ± 8.67**	36.91 ± 37.12	**< .001**

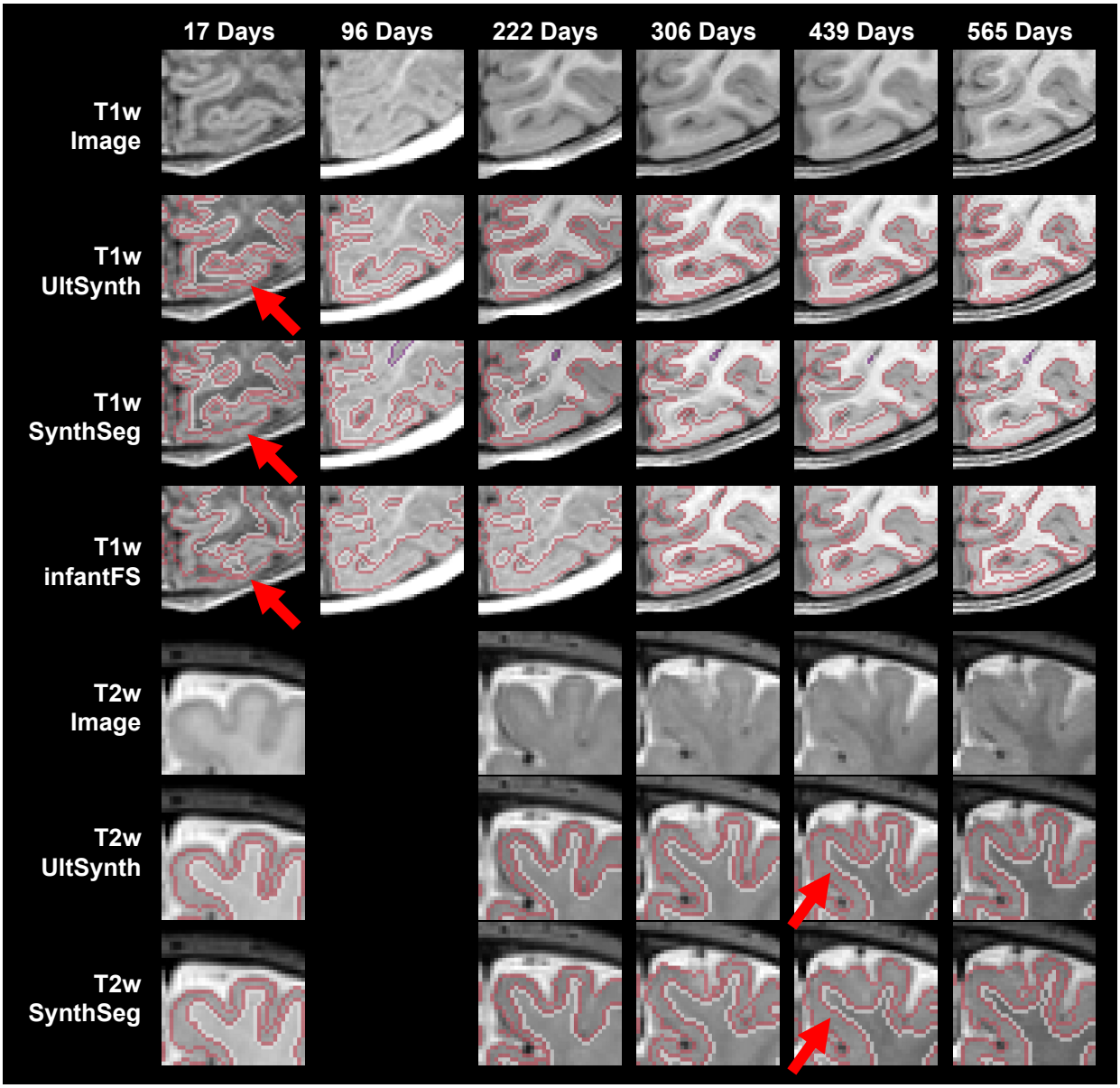

Fig. S8 | UltimateSynth segmentation with age-related contrast changes. UltimateSynth consistently segments highly convoluted cortical sulci (red arrows) despite poor and changing contrasts.

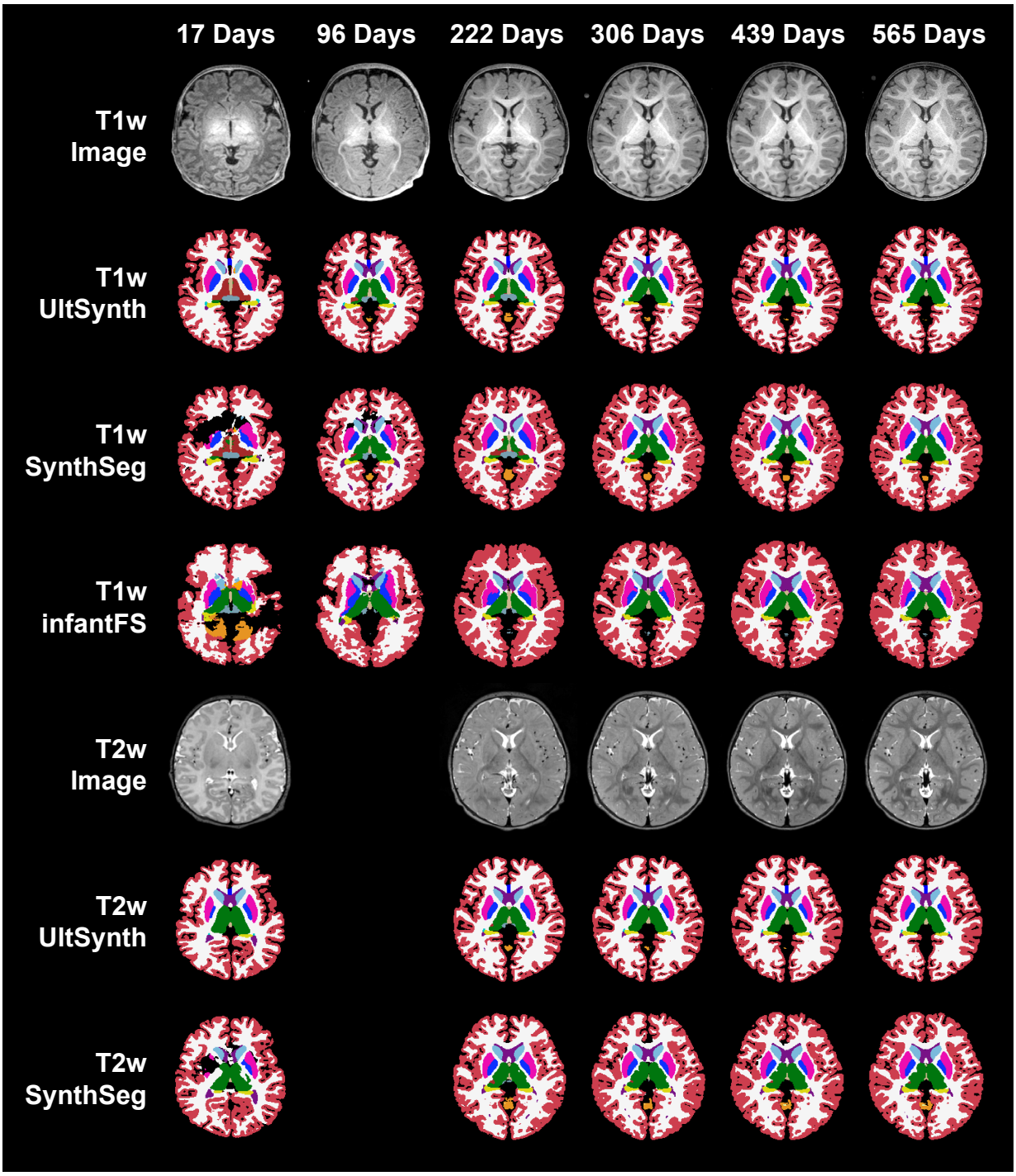

Fig. S9 | Six longitudinal timepoints and two MR image contrasts of the subject depicted in Figure 8. UltimateSynth demonstrates segmentation stability across time and contrasts. The 96-day-old T2-weighted image was not successfully acquired. Infant FreeSurfer accepts only T1-weighted images and is excluded from the T2-weighted comparison.

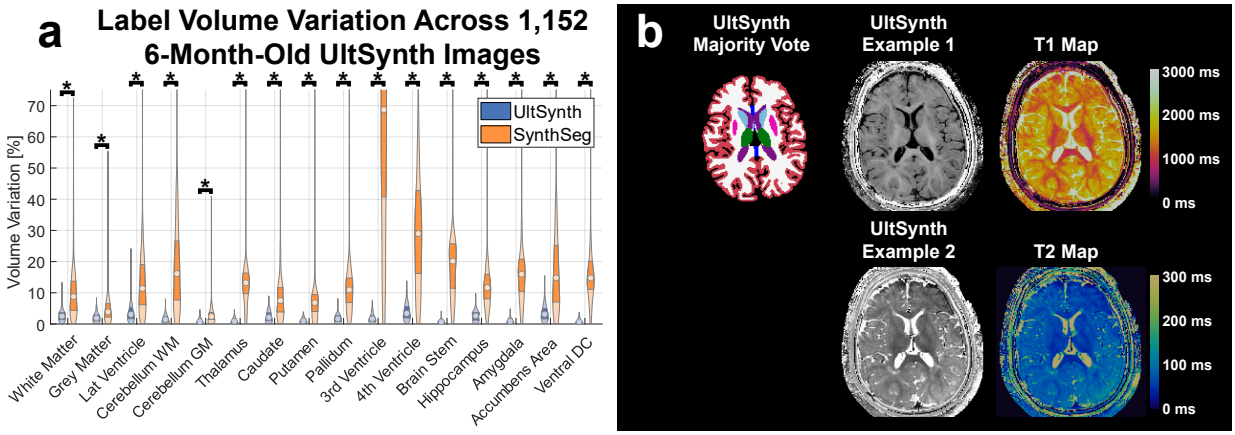

Fig. S10 | UltimateSynth segmentation of infant images. **a**, Percent LVV across 1,152 same-subject infant UltimateSynth images. UltimateSynth yields mean label variation $2.45 \pm 1.86\%$ compared to SynthSeg's $19.62 \pm 18.19\%$. **b**, Example UltimateSynth images used for segmentation and the original tissue parameter maps used for simulation. Compared to adult cases, infant brain segmentation is challenging due to the smaller brain size and on-going myelination.

Tab. S9 | Structure-specific LVV across 1,152 UltimateSynth images of a 6-month-old infant (Figure S10). Like the adult cases in Figure 5, UltimateSynth yields significantly lower LVV in all 16 structure labels shared with SynthSeg ($P < .001$).

Structure Label	UltSynth Volume Variation [%]	SynthSeg Volume Variation [%]	US-SS Volume Variation Difference *P*-Value
White Matter	**2.75 ± 1.99**	11.16 ± 11.19	**< .001**
Grey Matter	**2.07 ± 1.43**	5.76 ± 7.75	**< .001**
Lateral Ventricle	**4.31 ± 4.18**	15.42 ± 16.60	**< .001**
Cerebellar White Matter	**1.61 ± 1.17**	18.61 ± 14.06	**< .001**
Cerebellar Grey Matter	**1.00 ± 0.80**	3.04 ± 3.33	**< .001**
Thalamus	**1.06 ± 0.82**	18.83 ± 22.43	**< .001**
Caudate	**2.38 ± 1.66**	10.55 ± 13.56	**< .001**
Putamen	**1.16 ± 0.82**	8.99 ± 12.48	**< .001**
Pallidum	**1.89 ± 1.39**	13.93 ± 16.44	**< .001**
3rd Ventricle	**1.98 ± 1.48**	66.94 ± 35.39	**< .001**
4th Ventricle	**4.00 ± 2.75**	33.83 ± 25.30	**< .001**
Brain Stem	**0.86 ± 0.71**	25.98 ± 25.25	**< .001**
Hippocampus	**2.65 ± 1.73**	17.79 ± 22.07	**< .001**
Amygdala	**1.01 ± 0.75**	21.61 ± 22.51	**< .001**
Accumbens Area	**3.59 ± 2.59**	19.15 ± 17.90	**< .001**
Ventral Diencephalon	**0.91 ± 0.73**	22.48 ± 24.93	**< .001**

Tab. S10 | UltimateSynth performance as a function of training sample size (see Figure S12). Structure-specific Dice scores across 160 T1-weighted and FLAIR scans averaged over 10 subjects from the ON-Harmony dataset. Dice score improvement from US16 is only significant for only 9 labels ($P < .009$).

Structure Label	US16 Dice	US32 Dice	SS Dice	US16-US32 Dice Difference *P*-Value	US16-SS Dice Difference *P*-Value	US32-SS Dice Difference *P*-Value
White Matter	0.93 ± 0.02	0.93 ± 0.01	0.92 ± 0.01	**.007**	**< .001**	**< .001**
Grey Matter	0.87 ± 0.03	0.87 ± 0.02	0.86 ± 0.02	.97	**.008**	**.007**
Lateral Ventricle	0.86 ± 0.03	0.87 ± 0.03	0.88 ± 0.04	**< .001**	**< .001**	.06
Cerebellar White Matter	0.87 ± 0.02	0.86 ± 0.02	0.83 ± 0.02	.02	**< .001**	**< .001**
Cerebellar Grey Matter	0.90 ± 0.02	0.89 ± 0.02	0.91 ± 0.01	.02	**< .001**	**< .001**
Thalamus	0.91 ± 0.01	0.91 ± 0.01	0.89 ± 0.01	**< .001**	**< .001**	**< .001**
Caudate	0.87 ± 0.04	0.88 ± 0.02	0.88 ± 0.02	**.001**	.02	.05
Putamen	0.89 ± 0.02	0.90 ± 0.01	0.87 ± 0.02	.15	**< .001**	**< .001**
Pallidum	0.86 ± 0.02	0.86 ± 0.02	0.77 ± 0.04	**.008**	**< .001**	**< .001**
3rd Ventricle	0.80 ± 0.04	0.81 ± 0.03	0.79 ± 0.07	**< .001**	.60	**< .001**
4th Ventricle	0.82 ± 0.05	0.82 ± 0.05	0.81 ± 0.05	.77	.05	.07
Brain Stem	0.93 ± 0.01	0.92 ± 0.01	0.93 ± 0.01	**< .001**	.11	**< .001**
Hippocampus	0.86 ± 0.02	0.87 ± 0.01	0.86 ± 0.02	**< .001**	.46	**< .001**
Amygdala	0.82 ± 0.03	0.83 ± 0.02	0.84 ± 0.02	**< .001**	**< .001**	**< .001**
Accumbens Area	0.77 ± 0.02	0.77 ± 0.04	0.73 ± 0.03	.27	**< .001**	**< .001**
Ventral Diencephalon	0.85 ± 0.02	0.86 ± 0.02	0.83 ± 0.02	**< .001**	**< .001**	**< .001**

Tab. S11 | LVV analysis of UltimateSynth networks as a function of training sample size (see Figure S12). Structure-specific LVV across 160 T1-weighted and FLAIR scans from the ON-Harmony dataset shows few substantial differences between US16 and US32, with only 5 labels having significant differences in performance ($P < .008$).

Structure Label	US16 LVV	US32 LVV	SS LVV	US16-US32 LVV Difference *P*-Value	US16-SS LVV Difference *P*-Value	US32-SS LVV Difference *P*-Value
White Matter	1.49 ± 1.44	1.80 ± 1.62	2.50 ± 1.36	.07	**< .001**	**< .001**
Grey Matter	2.12 ± 1.29	1.54 ± 1.43	1.69 ± 1.50	**< .001**	**.007**	.35
Lateral Ventricle	3.26 ± 2.96	2.21 ± 1.76	2.40 ± 1.90	**< .001**	**.002**	.37
Cerebellar White Matter	1.64 ± 1.29	2.35 ± 1.87	9.26 ± 2.78	**< .001**	**< .001**	**< .001**
Cerebellar Grey Matter	2.39 ± 1.64	2.18 ± 1.52	1.72 ± 1.88	.23	**< .001**	**.016**
Thalamus	1.19 ± 0.91	0.94 ± 0.72	3.50 ± 2.35	**.007**	**< .001**	**< .001**
Caudate	2.27 ± 5.06	1.31 ± 1.27	1.41 ± 1.28	.02	.04	.52
Putamen	0.89 ± 0.74	1.11 ± 0.91	3.17 ± 2.54	.02	**< .001**	**< .001**
Pallidum	1.32 ± 1.09	1.26 ± 1.09	6.03 ± 4.42	.60	**< .001**	**< .001**
3rd Ventricle	3.00 ± 2.37	3.45 ± 2.50	9.76 ± 6.05	.10	**< .001**	**< .001**
4th Ventricle	3.99 ± 2.94	4.84 ± 3.51	15.64 ± 7.54	.02	**< .001**	**< .001**
Brain Stem	1.68 ± 1.46	1.45 ± 1.31	1.82 ± 1.27	.16	.36	**.01**
Hippocampus	1.11 ± 0.91	1.21 ± 1.05	3.69 ± 2.32	.36	**< .001**	**< .001**
Amygdala	1.81 ± 1.46	1.52 ± 1.29	6.46 ± 4.03	.07	**< .001**	**< .001**
Accumbens Area	2.52 ± 2.03	1.78 ± 1.38	4.09 ± 3.03	**< .001**	**< .001**	**< .001**
Ventral Diencephalon	1.40 ± 1.25	1.31 ± 0.99	3.29 ± 2.48	.48	**< .001**	**< .001**

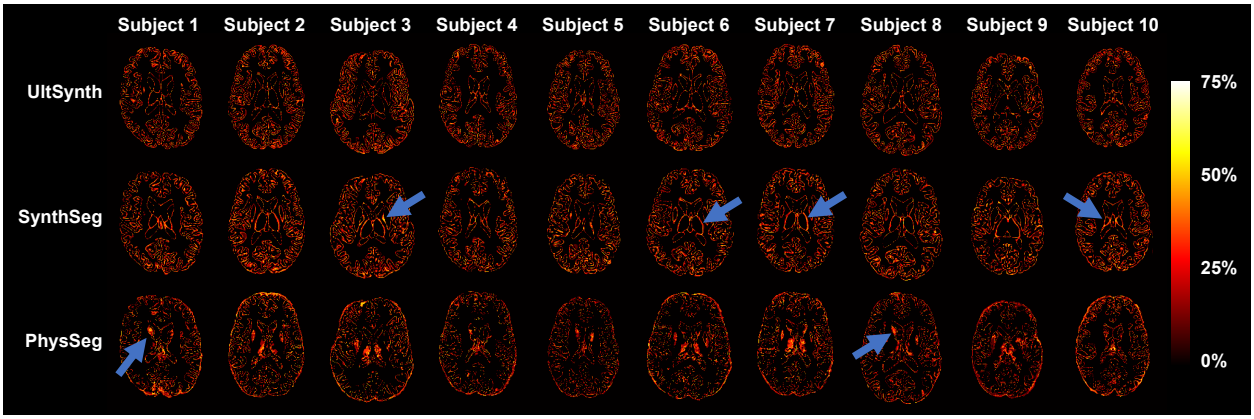

Fig. S11 | Voxel-wise percentage of disagreement with the majority-voted segmentation labels across ON-Harmony repeat images. UltimateSynth results demonstrate low uncertainty, particularly around the thalamus and putamen. SynthSeg and PhysSeg show consistent increases in disagreement around these structures (blue arrows).

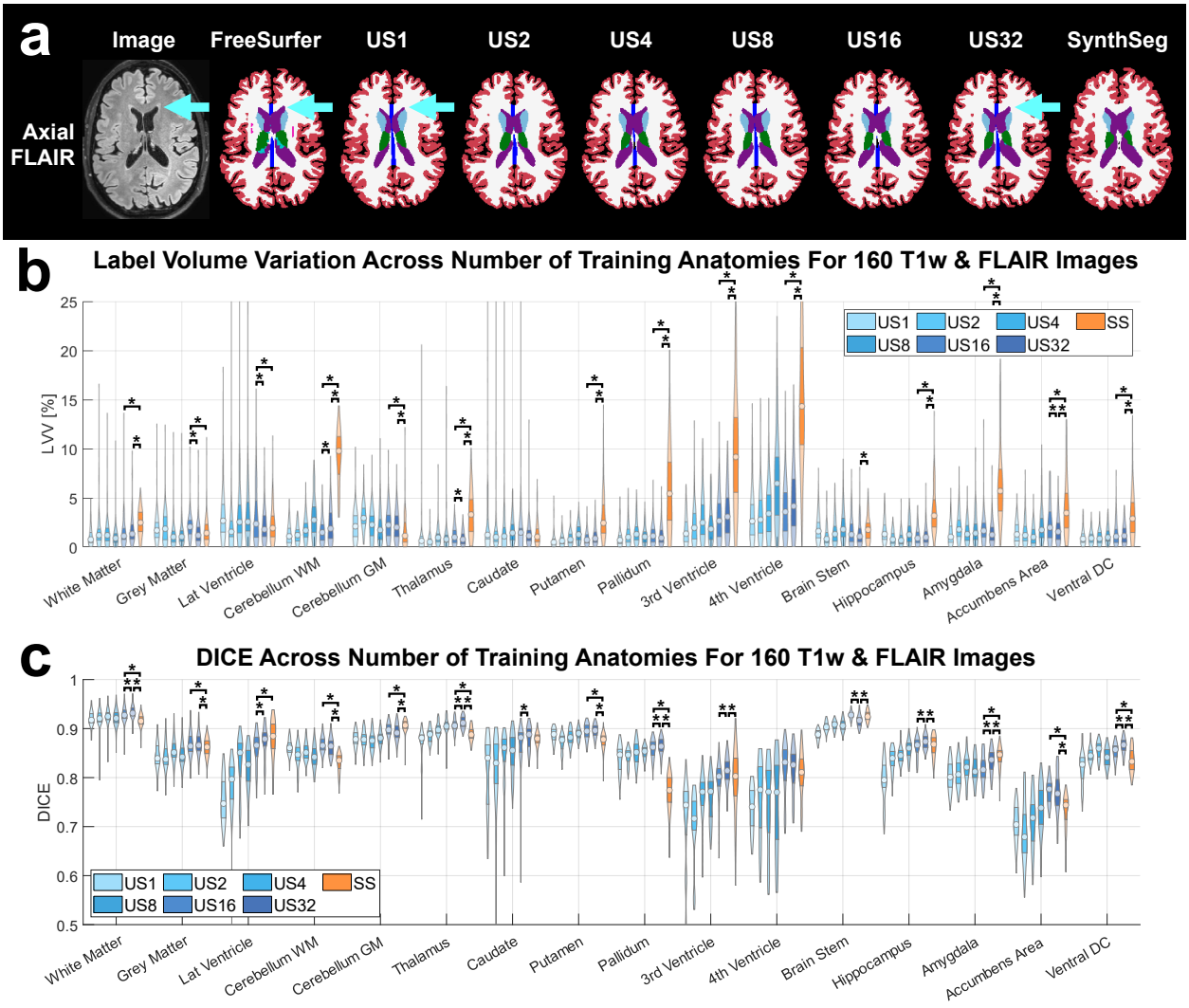

Fig. S12 | Performance of UltimateSynth-based networks as a function of unique training anatomies. **a**, Compared to FreeSurfer consensus labels, an UltimateSynth network trained with only one training anatomy (US1) sometimes mispredicts volumes of small structures such as the lateral ventricle (blue arrows), whereas a network trained with 32 diverse anatomical examples (US32) yields accurate segmentation. **b**, Structure-specific LVV with respect to FreeSurfer reference labels for SynthSeg and UltimateSynth networks trained with 1, 2, 4, 8, 16, and 32 unique training anatomies, based on 160 T1-weighted and FLAIR test images. All 6 UltimateSynth networks demonstrate low LVV across varying contrasts with no clear dependence on the training sample size. **c**, Dice scores for SynthSeg and a range of UltimateSynth networks with varying unique training anatomies compared to FreeSurfer reference labels across the same 160 test images as in **b**. UltimateSynth networks show a general trend of increasing Dice score performance with greater training sample sizes, with 9 of 16 tissue labels seeing statistically significant improvements in Dice scores between US16 and US32 ($P < .008$).

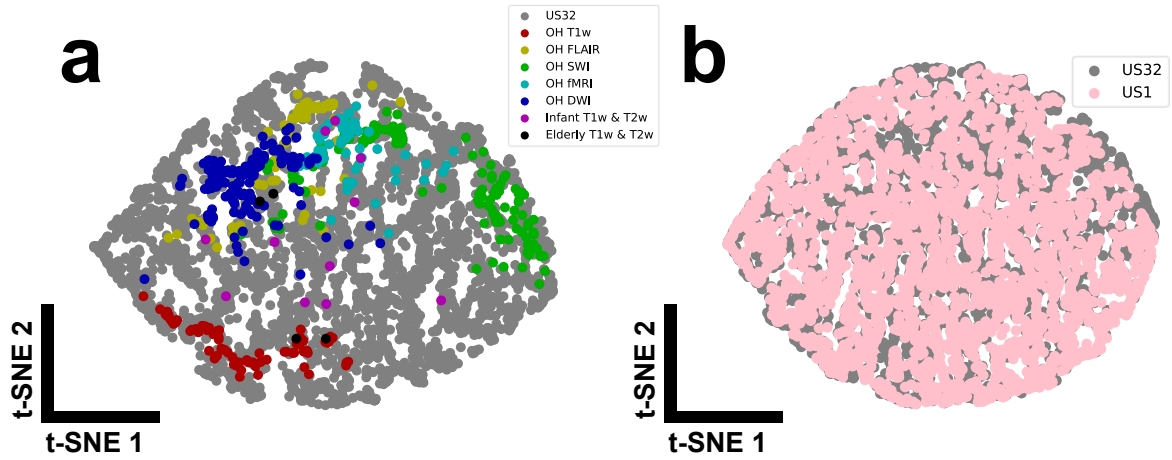

Fig. S13 | t-SNE visualization of the pan-contrast completeness of the US32 and US1 training datasets compared to an array of test images. For each label, 15 first-order radiomics features of image intensity were extracted, given a total of 270 features per image. One point in the 2D scatter plot represents one image. **a**, Comparison of the image contrasts in the US32 training set with 575 external test images. Points are generally grouped by feature similarity, emphasizing that the 3,200 UltimateSynth images in the US32 training dataset truly encompass all common and uncommon contrasts in the test images. Test-time T1-weighted, FLAIR, and other uncommon contrasts all fall within the scope of the gray point cloud. **b**, The image contrast scatters between US1 and US32 training images closely overlap, suggesting that pan-contrast training datasets can be produced independent of anatomical feature variation.

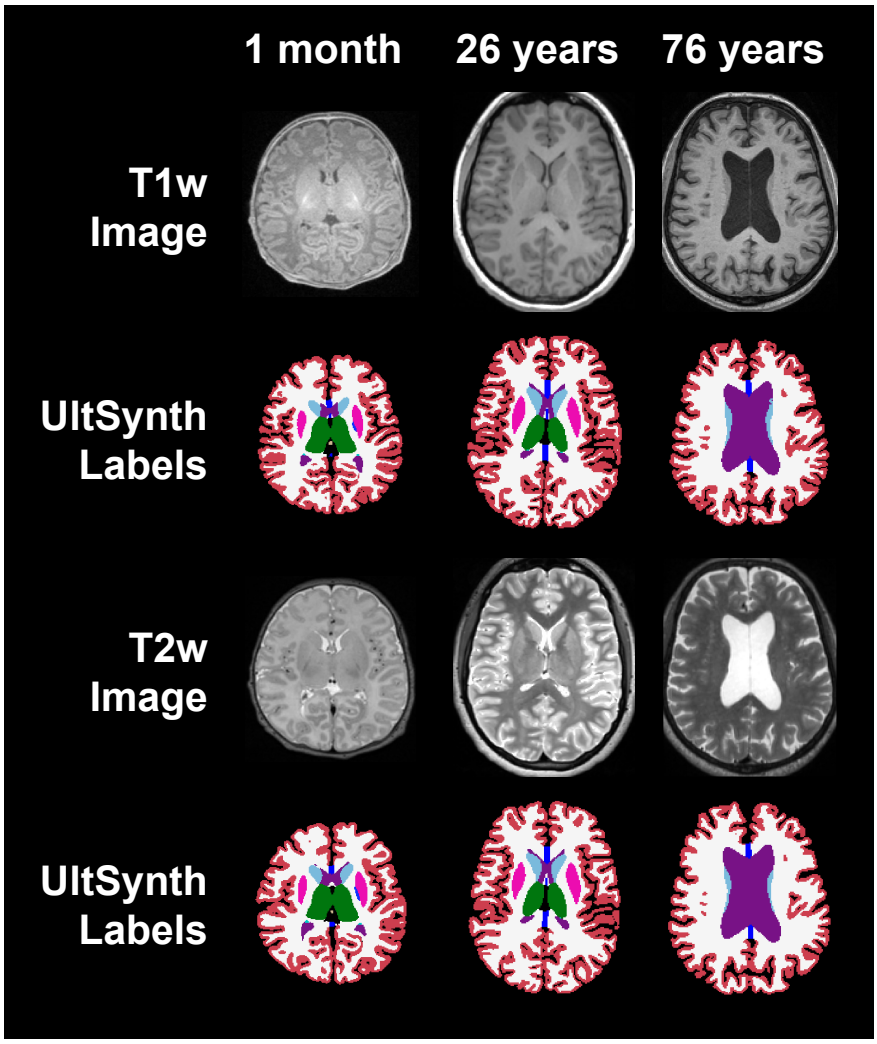

Fig. S14 | Segmentation across the human lifespan. UltimateSynth predicts reliable, accurate labels not only for adults, but also infants with small brain sizes and elderly individuals with enlarged ventricles.

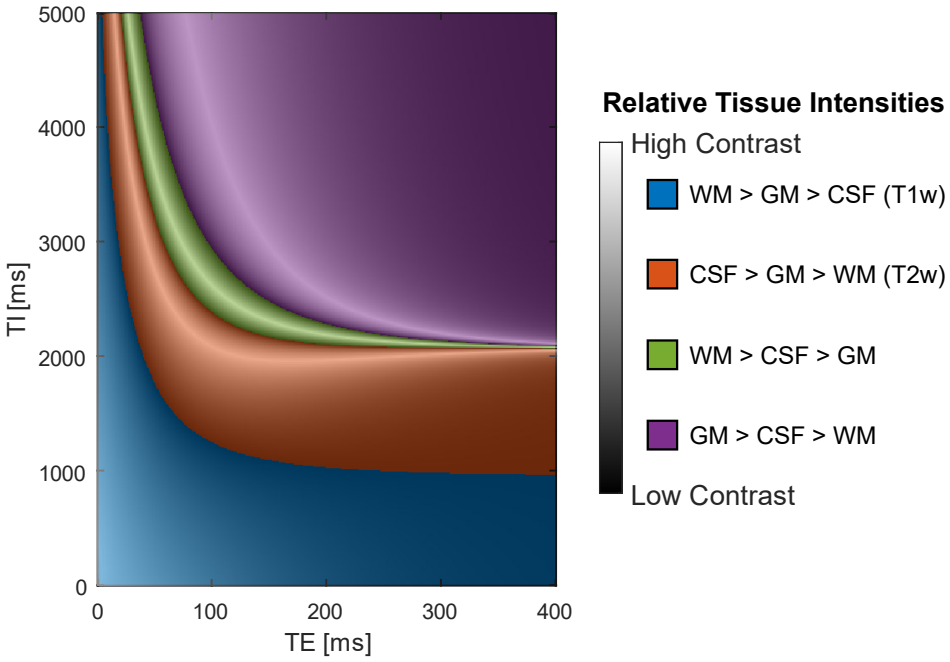

Fig. S15 | Distribution of inter-tissue contrasts between three tissues of interest (gray matter, white matter, and cerebrospinal fluid) in the UltimateSynth dictionary's scan parameters dimension (here plotting TI and TE as two separate axes). Each colored region corresponds to a different ordering of relative tissue intensities, which gives rise to different contrasts such as T1-weighted and T2-weighted images. Scan parameter combinations with high tissue contrast are depicted as the bright bands passing through colored like-contrast regions. As magnetization decays exponentially with time, these regions of desirable contrasts are not uniformly distributed throughout the possible scan parameter combinations.

Movie S1 `UltSynthImgs.gif` shows an example of 500 out of 19,440 images created for a single subject in Figure 5. Notice the wide range of contrasts covered as the SVD manifold coordinates vary.