TUMSyn: A Text-Guided Generalist Model for Customized Multimodal MR Image Synthesis

Yulin Wang[1,2†], Honglin Xiong[2†], Yi Xie[2], Jiameng Liu[2], Qian Wang[2,4], Qian Liu[1,5(✉)], Dinggang Shen[2,3,4(✉)]

[1] State Key Laboratory of Digital Medical Engineering, School of Biomedical Engineering, Hainan University, Haikou 570228, China
qliu@hainanu.edu.cn

[2] School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai 201210, China
dgshen@shanghaitech.edu.cn

[3] Shanghai United Imaging Intelligence Co., Ltd., Shanghai 200230, China

[4] Shanghai Clinical Research and Trial Center, Shanghai 201210, China

[5] Key Laboratory of Biomedical Engineering of Hainan Province, One Health Institute, Hainan University, Haikou 570228, China

Abstract. Multimodal magnetic resonance (MR) imaging has revolutionized our understanding of the human brain. However, various limitations in clinical scanning hinder the data acquisition process. Current medical image synthesis techniques, often designed for specific tasks or modalities, exhibit diminished performance when confronted with heterogeneous-source MRI data. Here we introduce a **T**ext-guided **U**niversal **M**R image **Syn**thesis (TUMSyn) generalist model to generate text-specified multimodal brain MRI sequences from any real-acquired sequences. By leveraging demographic data and imaging parameters as text prompts, TUMSyn achieves diverse cross-sequence synthesis tasks using a unified model. To enhance the efficacy of text features in steering synthesis, we pre-train a text encoder by using contrastive learning strategy to align and fuse image and text semantic information. Developed and evaluated on a multi-center dataset of over 20K brain MR image-text pairs with 7 structural MR contrasts, spanning almost entire age spectrum and various physical conditions, TUMSyn demonstrates comparable or exceeding performance compared to task-specific methods in both supervised and zero-shot settings, and the synthesized images exhibit accurate anatomical morphology suitable for various downstream clinical-related tasks. In summary, by incorporating text metadata into the image synthesis, the accuracy, versatility, and generalizability position TUMSyn as a powerful augmentative tool for conventional MRI systems, offering rapid and cost-effective acquisition of multi-sequence MR images for clinical and research applications.

Keywords: Foundation Model · Multimodal MRI · MRI Synthesis · Super-resolution.

† These authors contributed equally to this work.

1 Introduction

Magnetic resonance imaging (MRI) plays a pivotal role in neuroscience and clinical practice, enabling the exploration of the intricate structure and function of the human brain. However, clinical scanning constraints, including patient conditions and limited acquisition time, often lead to less sequences and suboptimal data quality. While several deep learning-based image synthesis methods [7,10,16–18] have been proposed to address these issues, their efficacy is often limited by task-specific and domain-specific training and inference paradigms. Consequently, these approaches exhibit suboptimal performance when applied to real-world clinical and research scenarios characterized by heterogeneous data.

Recent advances have sparked an increase in interest in exploring foundation models [11, 12, 20]. Built on large-scale and diverse datasets, these models exhibit flexibility in tackling multiple modalities, generating expressive outputs, and swiftly adapting to downstream tasks not explicitly defined by the training datasets by leveraging their mastered knowledge.

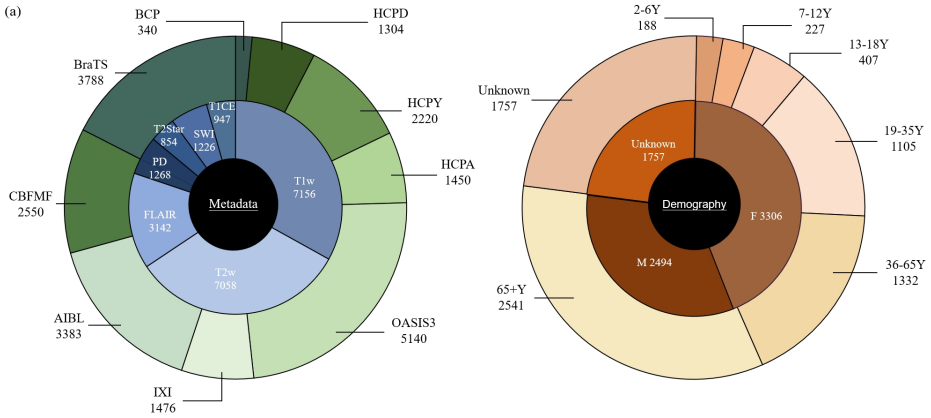
Drawing inspiration from the multimodal processing capabilities and robust generalizability of foundation models, here we present a **T**ext-guided **U**niversal **M**R image **Syn**thesis (TUMSyn) generalist model to synthesize neuroimages across nearly entire lifespan and multiple structural MRI modalities from any available MRI scan. By incorporating textual imaging parameters and demographic information into the image generation process, TUMSyn enables a single model to perform all cross-sequence synthesis tasks guided by text prompts, and achieves zero-shot generalization to novel data domains and tasks.

To achieve these capabilities, we embark on constructing a comprehensive brain MRI dataset, including over 20K 3D scans from public repositories and proprietary sources across multiple institutions. This extensive dataset incorporates seven common structural MRI modalities and represents a diverse population aged 2 to 100+, including healthy individuals and patients with various conditions. Furthermore, we pre-train a text encoder by leveraging the contrastive learning strategy [14, 19] to extract image-aligned text features to enhance the efficacy of text prompts in guiding image synthesis.

To assess the accuracy, versatility, and generalizability of our model, we conduct experiments on multi-center MR scans of controls and patients with brain tumors and Alzheimer's disease. The remarkable generated results and their effectiveness in downstream tasks demonstrate our model's proficiency in customized sequence synthesis and seamless adapting to various MR data domains prevalent across hospitals and institutions, with the promise of streamlining clinical workflows and reducing healthcare costs by augmenting acquired MR scan(s).

2 Method

The TUMSyn framework, illustrated in Fig.1, achieves general cross-sequence image synthesis through a two-stage process. In the first stage, we pre-train a



(b) **Age**:{34.1}; **Gender**:{M}; **Scanner**:{3.0 Siemens}; **Modality**:{T1w}; **Voxel size**:{(0.8, 0.8, 0.8)}; **Imaging parameter TR(ms), TE(ms), TI(ms), and FA(degree)**:{(2500.0, 2.2, 1000.0, 8.0)}

Fig. 1. Overview of the training set of brain MRI dataset. (a) The left plot illustrates the sample numbers of each dataset and MR modality, and the right one presents the sample numbers of each age group and gender. (b) A sample of the text prompt.

text encoder to learn the relationships between MR image representations and their corresponding demographic and imaging information. The second stage employs the frozen text encoder as the prompt producer to flexibly guide target images generation using any of the available sequences.

2.1 Data Collection

To ensure the versatility and generalizability of our model, our brain MRI multimodal dataset comprises over 20K 3D brain MRI scans from diverse global institutions, including OASIS [8], HCP [15], IXI [1], BCP [6], ADNI [13], AIBL [5], BraTS2021 [2], and a in-built dataset (CBMFM). And the HCP database contains three age groups, *i.e.*, HCP Young (HCPY), HCP Development (HCPD), and HCP Aging (HCPA). This comprehensive collection encompasses seven prevalent structural MRI modalities, *i.e.*, T1-weighted (T1w), T2-weighted (T2w), Fluid Attenuated Inversion Recovery (FLAIR), Susceptibility Weighted Imaging (SWI), T2 Star, Proton Density (PD), and Contrast-Enhanced T1w (T1CE), representing an entire lifespan cohort that includes healthy individuals and patients with neurodegenerative and neurodevelopment disorders and brain tumors. The detailed description of the training set is illustrated in Fig.1(a), and we reserved the ADNI dataset for external validation experiments to ensure a robust evaluation of our model's generalizability. To ensure the model can adapt to real clinically heterogeneous data as well as learn general MR image features, we simplify the image pre-processing steps as follows: (i) registering multiple modalities of each subject, and (ii) skull-stripping. Concurrently, the corresponding textual metadata for each scan is also an important part of our

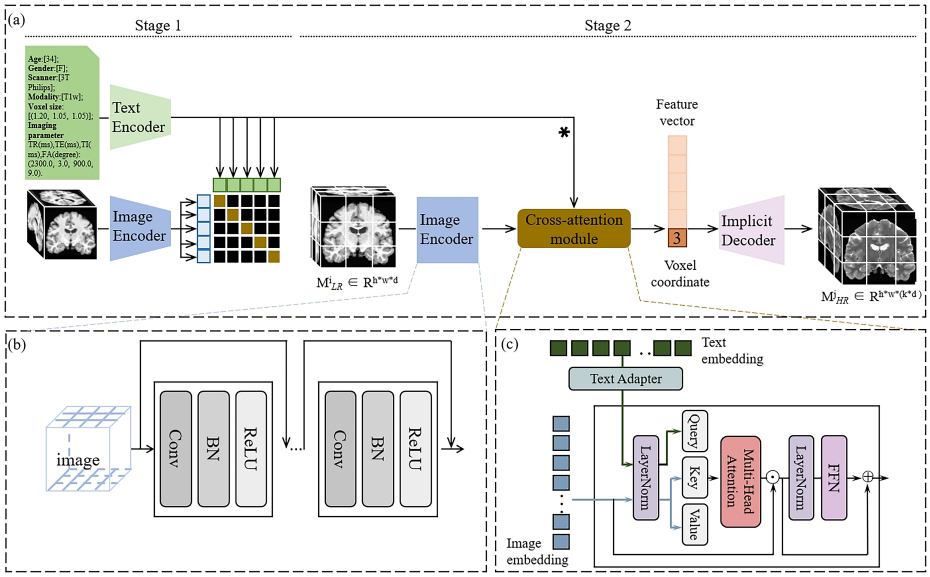


Fig. 2. Overview of TUMSyn. (a) The pipeline of prompt-guided MR image synthesis and super-resolution, which includes pre-training image-text alignment (Stage 1) and universal MR image SR and cross-sequence synthesis (Stage 2). i and j represent different modalities. (b) Model architecture of image encoder. (c) The model architecture of image-text cross-attention module.

dataset. A sample of our template is shown in Fig.1(b), where TR, TE, TI, and FA represent repetition time, echo time, inversion time, and flip angle, respectively. These parameters significantly influence the contrast and signal intensity of the scanned images. All imaging parameters and demographic information are derived from DICOM header files, official websites, and demographic statements. In cases of missing information, we consistently use 'None' placeholders. Note that information about the target modality is invariably present in all text prompts.

2.2 Pre-trained Text Encoder for Enhanced Feature Extraction

To enable precise guidance of downstream image synthesis tasks by textual prompts, we introduce a contrastive language-image pre-training (CLIP) model dedicated to brain MR images (BMLIP) in the first stage. Our goal is to equip the text encoder with the capability to extract metadata embeddings corresponding to images. The overall framework is depicted in Stage 1 of Fig.2(a). Specifically, given a batch of text-image pairs as inputs, BMLIP learns multimodal feature correspondences in latent space through joint training of image and text encoders. The training objective is to minimize the embedding distances between paired samples and maximize the distances between non-paired samples. It can be formulated as:

$$\mathcal{L} = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{\exp(\cos(\mathbf{v}_i,\mathbf{t}_i)/\tau)}{\sum j=1^{N}\exp(\cos(\mathbf{v}_i,\mathbf{t}_j)/\tau)} + \log\frac{\exp(\cos(\mathbf{t}_i,\mathbf{v}_i)/\tau)}{\sum j=1^{N}\exp(\cos(\mathbf{t}_i,\mathbf{v}_j)/\tau)}\right] \quad (1)$$

Where N denotes the batch size, and $\mathbf{v}_i$ and $\mathbf{t}_i$ represent the encoded image and text features for the i-th sample, respectively. The cosine similarity between two features is calculated using the $\cos(\cdot,\cdot)$ operator, and τ is a temperature parameter controlling the distribution concentration.

Details of the model structure and workflow of BMLIP are given below. For the text encoder, templated textual metadata (Fig.1(b)) corresponding to each scan are fed into a tokenizer, where each character is converted into a numerical identifier. The numerical sequences are then passed into a standard transformer-based text encoder to obtain text representations. Due to the richer content in our text compared to the prompts used by standard CLIP [14], we set the encoded length to 90 tokens. For the image encoder, recognizing MRI scans are typically 3D and the critical importance of inter-slice structural information for diagnostic and downstream image analysis tasks, we modify the ViT-B/16 architecture used in standard CLIP designed for 2D natural images to accommodate 3D MR images. To manage the computational demands of 3D image processing, we first downscale the entire input image to half of its original size and then crop all images to 96×96×96 resolutions. This approach preserves semantic information in the images, ensuring effective semantic matching between text and images during training. Upon completion of BMLIP training, we only apply the text encoder to the subsequent stage.

2.3 Tailored MRI Synthesis and Super-Resolution Guided by Text Prompt

The cross-sequence image synthesis model, illustrated in Stage 2 of Fig.2(a), aims to generate tailored target MRI scans from available images of varying MR modalities, resolutions, and orientations, guided by text prompts. This process begins with encoding input pairs of text and image patches (60×60×60) using a convolutional neural network (CNN)-based image encoder and a frozen pre-trained text encoder to extract multi-modality features, respectively. To accommodate super-resolution requirements, we employ a modified ResNet-34 without downsampling as our image encoder (Fig.2(b)). Textual MR knowledge is incorporated into the synthesis process through a cross-attention module (Fig.2(c)). The cross-modality integration process begins with a text adapter that updates text embeddings to align with our specific task requirements. Subsequently, multi-head attention mechanisms fuse these adapted text embeddings with image features, generating multi-modality representations enriched by textual information. To ensure the versatility of our model, we implement an image decoder capable of arbitrary resolution upsampling. This is achieved through the Local Implicit Image Function (LIIF) [4], which adds high-resolution image coordinates to the feature vector around the low-resolution coordinates as input and predicts the intensity value at a given continuous coordinate using

an implicit decoder. Finally, we obtain an image with tailored demographic and imaging parameter information specified in the text prompt. In this stage, we employ pixel loss as supervision to synthesize images.

3 Experiment

3.1 Implementation Details

TUMSyn is implemented using PyTorch 1.12.1 and trained on a server equipped with a Nvidia A100 GPU. We adopt the Adam optimizer for both model training stages. The learning rate (LR) in the first stage increases from 0 to 0.0005 for model warm-up, and then gradually decreases to 0, while the LR in the second stage is set to 0.0001 initially and decays by half every 100 epochs. The training epochs for models in both stages are 100 and 300, respectively.

3.2 Comparative evaluation of accuracy and versatility in image synthesis

In this experiment, we aim to demonstrate the model's capability to accurately and uniformly generate target images under the guidance of text prompts. First, we compare TUMSyn with several state-of-the-art methods and perform in-depth quantitative assessments across seven typical tasks, as outlined in Table1. SC-GAN [9] is a task-specific image mapping method. The one-hot model and BiomedCLIP model share the identical synthesis model architecture, but one-hot model uses numerical labels to replace the text prompts and BiomedCLIP model uses Biomedclip [19] to produce text embeddings. The PSNR and SSIM scores reveal that TUMSyn consistently exhibits superior performance across all the tasks. Compared to SC-GAN, our findings highlight the effectiveness of generic feature representations learning from abundant data. When compared to BiomedCLIP, the results underscore the necessity of specifically pre-trained foundational models in the medical domain. The comparison results with the one-hot method emphasize the accuracy of leveraging text prompts to steer the image translation process. We also perform a comparative analysis of the qualitative outcomes between TUMSyn and the aforementioned competitors on multiple tasks (Fig.3). SynthSR [7] is further enrolled, which is a general MRI image synthesis algorithm that specializes in translating diverse MRI sequences into T1w sequences with isotropic voxel sizes.

Notably, TUMSyn adeptly achieves the best results in all cases. Rows 1 to 4 illustrate the capacity of TUMSyn to generate modalities tailored to the specified resolution, MR contrast, and voxel intensity as prompted by the text, irrespective of the imaging parameters of the input sequences. Subsequently, the final two rows demonstrate the remarkable ability of BMLIP to capture and retain brain anatomical structures and generate brain tissues, and the segmentation results obtained from Synthseg+ [3] reaffirm the compatibility of its generated outputs with existing medical image analysis tools.

Table 1. Universality and accuracy comparison with different methods on the test set

		AIBL PD->FLAIR	HCPD T1->T2	BRATS FLAIR->T1CE	CBMFM T2->FLAIR	IXI T1->PD	OASIS T1->SWI	OASIS T2->T2S
SC-GAN [9]	PSNR	25.05±0.98	28.24±0.43	24.11±1.52	28.27±1.82	30.33±0.87	25.48±1.02	28.48±0.99
	SSIM	0.903±0.012	0.940±0.005	0.871±0.015	0.958±0.020	0.961±0.003	0.851±0.019	0.939±0.011
BiomedCLIP [19]	PSNR	27.26±1.22	28.76±0.61	26.37±1.48	30.91±2.26	31.29±1.10	25.63±1.15	28.48±1.14
	SSIM	0.937±0.012	0.943±0.007	0.913±0.013	0.975±0.019	0.967±0.007	0.855±0.028	0.938±0.016
One-hot	PSNR	25.58±0.97	27.27±0.67	24.57±1.23	28.63±1.82	28.57±0.51	21.87±0.92	25.12±0.88
	SSIM	0.919±0.010	0.930±0.008	0.905±0.014	0.966±0.019	0.635±0.017	0.849±0.030	0.911±0.015
TUMSyn	PSNR	**27.43±1.17**	**29.02±0.68**	**26.78±1.67**	**31.13±2.43**	**31.78±0.88**	**25.63±1.02**	**28.49±1.17**
	SSIM	**0.938±0.010**	**0.945±0.007**	**0.915±0.013**	**0.976±0.019**	**0.969±0.007**	**0.856±0.028**	**0.939±0.016**

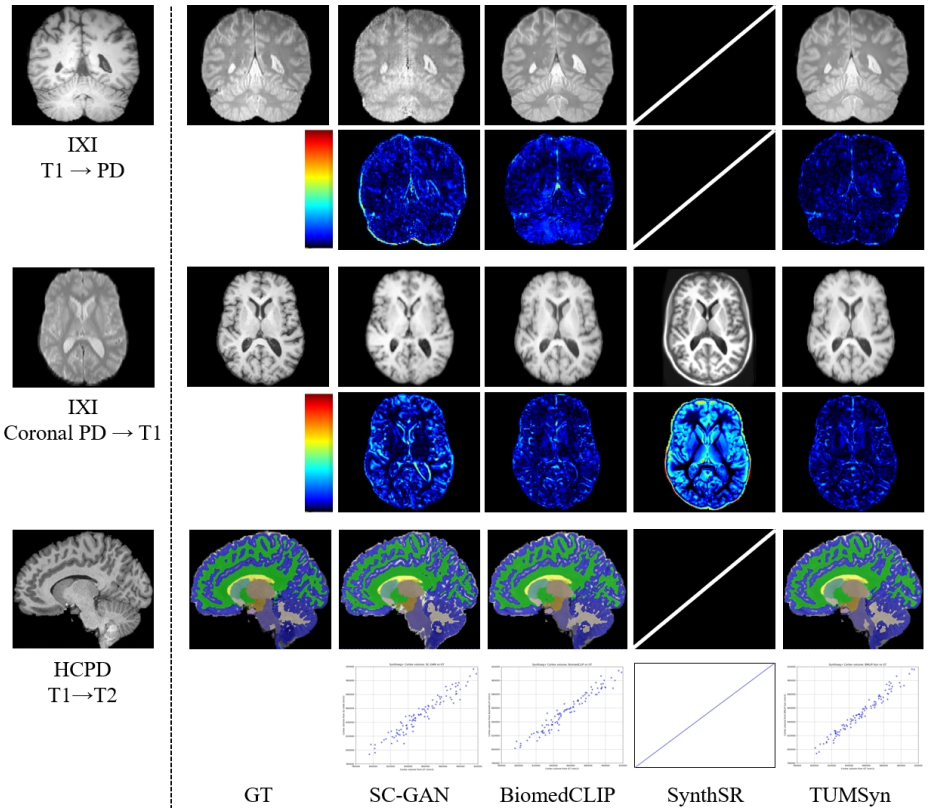


Fig. 3. Visualization of synthesized images from heterogeneous input data. Row 1 and 2 show synthetic PD images from real-acquired T1w and corresponding error maps compared to the ground truth in the IXI dataset, respectively. Row 3 and 4 present T1w synthetic results from simulated coronal PD sequence with 3.2mm slice spacing and corresponding error maps compared to ground truth, respectively, and we present these results in the orthogonal view to illustrate the performance of joint image super-resolution and cross-sequence synthesis. Row 5 and 6 present whole-brain segmentation results obtained using SynthSeg+ and scatter plots of cortex volumes derived from real-acquired T1w and synthetic T2w images, respectively.

Table 2. Hippocampal volumetry of AD, MCI, and NC. The rows indicate the average (Avg.) volume of ROI for different cognitive stages.

	Real T1w	SC-GAN	BiomedCLIP	SynthSR	TUMSyn
Avg. AD (mm^3)	3280.11	3326.58	3560.73	3419.84	3391.97
Avg. MCI (mm^3)	3614.33	3664.55	3957.30	3817.15	3770.19
Avg. NC (mm^3)	4030.67	4037.89	4237.52	4061.15	4040.37

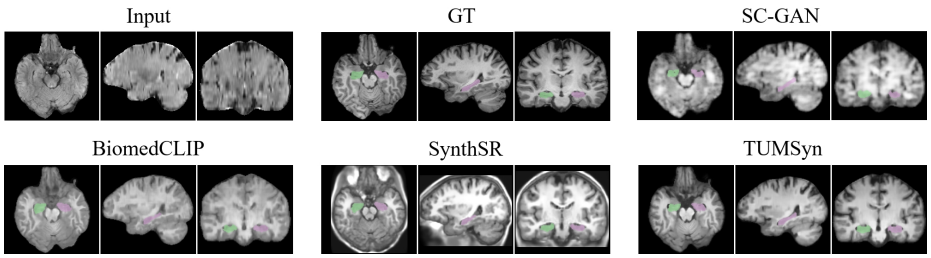


Fig. 4. Joint image synthesis and SR results of a MCI subject from ADNI datasets. The hippocampal segmentation results obtained by Synthseg+ are overlaid on the generated T1w images from 5mm axial FLAIR.

3.3 Zero-shot detecting AD-induced brain region atrophy

The full potential of a foundation model cannot be realized without sufficient generalization. In this experiment, we evaluate TUMSyn's zero-shot performance using an out-of-distribution dataset ADNI, assessing its ability to generate lesion images that reflect the effects of cognitive impairment on hippocampal shape and volume. The hippocampus is a key biomarker of neurodegenerative diseases, and typically exhibits progressive volume reduction as the disease advances. Table2 presents the average hippocampal volumes for Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), and Normal Control (NC), as estimated by each method. Volumes derived from real T1w images serve as the ground truth for each disease stage.

In comparison to other methods, TUMSyn demonstrates robust lesion volume delineation capabilities and strong discriminative power across the three disease stages, approaching the performance of real T1w images and SC-GAN trained on this dataset. Fig.4 illustrates the T1w sequence outputs from various methods and the resulting hippocampal segmentation for a representative MCI patient scan. Qualitatively, TUMSyn exhibits the superior recovery of high-frequency details, facilitating accurate hippocampal segmentation when used with Synthseg+.

4 Conclusion

In this study, we present TUMSyn, a text-guided universal model for the unified synthesis of tailored structural MRI modalities, using any available MRI sequences. Developed and evaluated on a diverse-source dataset comprising over 20K 3D MRI scans, TUMSyn demonstrates promising synthesis accuracy in producing diagnostically relevant sequences and remarkable adaptability to novel data domains. Altogether, TUMSyn holds significant potential for optimizing clinical workflows and reducing expenses in both healthcare and research settings.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China (grant numbers 62131015, U23A20295, 62250710165), the STI 2030-Major Projects (No. 2022ZD0209000), Shanghai Municipal Central Guided Local Science and Technology Development Fund (grant number YDZX20233100001001), The Key R&D Program of Guangdong Province, China (grant numbers 2023B0303040001, 2021B0101420006), and Science and Technology special fund of Hainan Province (grant number KJRC2023B06).

Disclosure of Interests. The authors declare no competing interests.

References

1. IXI dataset. `https://brain-development.org/ixi-dataset/`
2. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
3. Billot, B., Greve, D.N., Puonti, O., Thielscher, A., Van Leemput, K., Fischl, B., Dalca, A.V., Iglesias, J.E., et al.: Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining. Medical image analysis **86**, 102789 (2023)
4. Chen, Y., Liu, S., Wang, X.: Learning continuous image representation with local implicit image function. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8628–8638 (2021)
5. Ellis, K.A., Bush, A.I., Darby, D., De Fazio, D., Foster, J., Hudson, P., Lautenschlager, N.T., Lenzo, N., Martins, R.N., Maruff, P., et al.: The australian imaging, biomarkers and lifestyle (aibl) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of alzheimer's disease. International psychogeriatrics **21**(4), 672–687 (2009)
6. Howell, B.R., Styner, M.A., Gao, W., Yap, P.T., Wang, L., Baluyot, K., Yacoub, E., Chen, G., Potts, T., Salzwedel, A., et al.: The unc/umn baby connectome project (bcp): An overview of the study design and protocol development. NeuroImage **185**, 891–905 (2019)
7. Iglesias, J.E., Billot, B., Balbastre, Y., Magdamo, C., Arnold, S.E., Das, S., Edlow, B.L., Alexander, D.C., Golland, P., Fischl, B.: Synthsr: A public ai tool to turn heterogeneous clinical brain scans into high-resolution t1-weighted images for 3d morphometry. Science advances **9**(5), eadd3607 (2023)

8. LaMontagne, P.J., Benzinger, T.L., Morris, J.C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A.G., et al.: Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. MedRxiv pp. 2019–12 (2019)
9. Lan, H., Initiative, A.D.N., Toga, A.W., Sepehrband, F.: Three-dimensional self-attention conditional gan with spectral normalization for multimodal neuroimaging synthesis. Magnetic resonance in medicine **86**(3), 1718–1733 (2021)
10. Liu, J., Pasumarthi, S., Duffy, B., Gong, E., Datta, K., Zaharchuk, G.: One model to synthesize them all: Multi-contrast multi-scale transformer for missing data imputation. IEEE Transactions on Medical Imaging (2023)
11. Ma, J., Wang, B.: Towards foundation models of biological image segmentation. Nature Methods **20**(7), 953–955 (2023)
12. Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P.: Foundation models for generalist medical artificial intelligence. Nature **616**(7956), 259–265 (2023)
13. Petersen, R.C., Aisen, P.S., Beckett, L.A., Donohue, M.C., Gamst, A.C., Harvey, D.J., Jack, C.R., Jagust, W.J., Shaw, L.M., Toga, A.W., et al.: Alzheimer's disease neuroimaging initiative (adni): clinical characterization. Neurology **74**(3), 201–209 (2010)
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
15. Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., Consortium, W.M.H., et al.: The wu-minn human connectome project: an overview. Neuroimage **80**, 62–79 (2013)
16. Wang, Y., Hu, H., Yu, S., Yang, Y., Guo, Y., Song, X., Chen, F., Liu, Q.: A unified hybrid transformer for joint mri sequences super-resolution and missing data imputation. Physics in Medicine and Biology (2023)
17. Wang, Y., Wu, W., Yang, Y., Hu, H., Yu, S., Dong, X., Chen, F., Liu, Q.: Deep learning-based 3d mri contrast-enhanced synthesis from a 2d noncontrast t2flair sequence. Medical Physics **49**(7), 4478–4493 (2022)
18. Wu, Q., Li, Y., Sun, Y., Zhou, Y., Wei, H., Yu, J., Zhang, Y.: An arbitrary scale super-resolution approach for 3d mr images via implicit neural representation. IEEE Journal of Biomedical and Health Informatics **27**(2), 1004–1015 (2022)
19. Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al.: Large-scale domain-specific pretraining for biomedical vision-language processing. arXiv preprint arXiv:2303.00915 **2**(3), 6 (2023)
20. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., et al.: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (2023)