Appendix

A Datasets and Metrics

Datasets We test and compare our method over various datasets including modalities of MR and CT, the MR images further contain T1-weighted , T2-weighted and FLAIR (fluid-attenuated inversion recovery) images.

- ADNI [30]: we use T1-weighted (2045 cases) MRI scans from the Alzheimer's Disease Neuroimaging Initiative (ADNI). All scans are acquired at 1 *mm* isotropic resolution from a wide array of scanners and protocols. The dataset contains aging subjects, some diagnosed with mild cognitive impairment (MCI) or Alzheimer's Disease (AD). Many subjects present strong atrophy patterns and white matter lesions.
- HCP [19]: we use T1-weighted (897 cases) and T2-weighted (897 cases) MRI scans of young subjects from the Human Connectome Project, acquired at 0.7 mm resolution.
- ADNI3 [60]: we use T1-weighted (331 cases) and FLAIR (331 cases) MRI scans from ADNI3, which continues the previously funded ADNI1, ADNI-GO, and ADNI2 studies to determine the relationships between the clinical, cognitive, imaging, genetic and biochemical biomarker characteristics of the entire spectrum of sporadic late onset AD.
- ADHD200 [9]: we use T1-weighted (961 cases) MRI scans from ADH200 Sample, which is a grassroots initiative dedicated to the understanding of the neural basis of Attention Deficit Hyperactivity Disorder (ADHD).
- AIBL [23]: we use T1-weighted (668 cases), T2-weighted (302 cases) and FLAIR (336 cases) MRI scans from The Australian Imaging, Biomarkers and Lifestyle (AIBL) Study, which is a study of cognitive impairment (MCI) and Alzheimer's disease dementia.
- OASIS3 [35]: we use CT (885 cases) scans from OASIS3, which is a longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and AD. For our experiments, we use CT and T1-weighted MRI pair with the earliest date, from each subject.

Data for Synthetic Generator Brain-ID's synthetic generator uses (1) brain segmentation labels, for random-contrast input images generation (Sec. 3.1), and (2) MP-RAGE, the target ground truth for anatomy-guided supervision (Sec. 3.2). In this work, we use the segmentation maps of training images from ADNI [30], as well as their corresponding MP-RAGE images. Note that we do not use any type of real images from ADNI as input for Brain-ID's pre-training.

Data Preprocessing For all datasets, we skull-strip all the images using SynthStrip [26], and resample them to 1 *mm* isotropic resolution. For all the images, except T1-weighted MRI, in each dataset, we use NiftyReg [43] rigid registration to register all images to their same-subject T1-weighted MRI counterparts. The gold-standard brain segmentation maps are obtained by performing SynthSeg [4] on the T1-weighted MR images of all the subjects.

Metrics We resort to various metrics for evaluating individual tasks across multiple aspects:

- `L1`: the average $L1$ distance, is used for intra/inter-subject feature distance evaluation (Sec. 4.2), and the overall prediction correctness of anatomy reconstruction (Sec. 4.3), super-resolution and bias-field estimation (Sec. 4.3).
- `normL2`: the normalized $L2$ distance for bias field [16] is defined as:

$$\texttt{normL2} = \sqrt{\frac{\sum_x \left(wB_{\text{est}}(x) - B_{\text{true}}(x)\right)^2}{\sum_x B^2_{\text{true}}(x)}}\,, \tag{A.1}$$

  where w is the normalization coefficient obtained by:

$$w = \frac{\sum_x B_{\text{true}}(x) B_{\text{est}}(x)}{\sum_x B^2_{\text{est}}(x)}\,, \quad x \in \Omega\,, \tag{A.2}$$

  Ω refers to the brain domain, B_{est} and B_{true} are the estimated and ground truth bias fields, respectively. Normalization is necessary for the evaluation of bias field estimation (Sec. 4.3) because nonuniformity correction may result in arbitrary scaling of the bias field.
- `PSNR`: the peak signal-to-noise ratio (PSNR) that indicates the fidelity of predictions. It is used in anatomy reconstruction (Sec. 4.3) and super-resolution (Sec. 4.3).
- `(MS-)SSIM`: the structural similarity scores between the generated and real images. `MS-SSIM` is a variant of `SSIM` focusing on multiple scales of the images that are shown to correlate well with human perception [41, 58, 59]. They are used in intra/inter-subject feature distance evaluation (Sec. 4.2), reconstruction (Sec. 4.3), and super-resolution (Sec. 4.3),
- `Dice`: the similarity score between predicted and ground truth segmentations, and it is used in brain segmentation evaluation (Sec. 4.3).

B Implementation Details

Model Architecture As mentioned in Sec. 4, `Brain-ID` can use any backbone to extract brain features. We use the five-level 3D UNet [47] as `Brain-ID`'s backbone for feature extraction, with 64 feature channels in the last layer.

- During the feature pre-training stage (Sec. 3.2), a linear regression layer is added following the feature outputs for anatomy supervision (Eq. (3)).
- During downstream tasks adaptions, the regression layer for anatomy supervision is abandoned, instead, a task-specific activation layer is added following the feature outputs (Sec. 3.3). Specifically, a linear regression layer is added for the tasks, anatomy reconstruction/contrast synthesis, image super-resolution, and bias field estimation. An additional softmax activation is added for segmentation probability outputs.

Table B.1: `Brain-ID` synthetic generator setups: mild, medium and severe levels. p denotes probability, μ and σ refer to the mean and variance of the Gaussian distributions, respectively.

Category	Param	Corruption Level		
		Mild	Medium	Severe
Deformation	affine-rotation$_{max}$	15	=	=
	affine-shearing$_{max}$	0.2	=	=
	affine-scaling$_{max}$	0.2	=	=
	nonlinear-scale μ_{min}	0.03	=	=
	nonlinear-scale μ_{max}	0.06	=	=
	nonlinear-scale σ_{max}	4	=	=
Resolution	$p_{\text{low-field}}$	0.1	0.3	0.5
	$p_{\text{anisotropic}}$	0	0.1	0.25
Bias Field	μ_{min}	0.01	0.02	0.02
	μ_{max}	0.02	0.03	0.04
	σ_{min}	0.01	0.05	0.1
	σ_{max}	0.05	0.3	0.6
Noises	σ_{min}	0.01	0.5	5
	σ_{max}	1	5	15

Feature Backbone Pre-training We pre-train `Brain-ID` on the synthetic data from our generator (Sec. 3.1) for 300,000 iterations, with a patch size of 128^3 and a mini-batch size (i.e., number of intra-subject augmented samples) of 4. We use the synchronized AdamW optimization, with a base learning rate of 10^{-4} and a linear warm-up in the first 2,000 iterations followed by a multi-step learning schedule (learning rate drops at 160,000 and 240,000 iterations) with a multiplier of 0.1.

Synthetic Data Generator As shown in Fig. 3, `Brain-ID` simulates its training samples of increasing corruption levels, from mild to severe. Tab. 3 also explores the effects of different levels of sample corruption on feature robustness and downstream performance. In Tab. B.1, we list the generator parameters for mild, medium, and severe data corruption levels, respectively. Note that (1) for each level, the setup parameters only control the corruption value ranges, since the simulation is randomized, there could still be mildly corrupted samples generated under the "severe" settings; (2) The random deformation fields are independent of data corruption levels.

C Feature Robustness Evaluation

For the evaluation of feature robustness, we use T1-weighted MRI of 100 randomly selected subjects from ADNI [30]. To challenge the model's robustness against data corruptions, and meanwhile obtain comparable and reproducible results, we use our data generator (Sec. 3.1) to pre-augment all the input images. Note that in this section, there is no contrast simulation step within data augmentation, and only the random deformation and data corruptions are applied.

For each selected subject, we generate intra-augmented samples. All samples are generated with "medium" corruption settings as listed in Tab. B.1.

D Downstream Task Comparisons

For all the downstream tasks, the model architecture and date generation strategy used for `Brain-ID`, `SCRATCH` and `CIFL` are the same. The only difference between the three compared models lies in their initial weights. `Brain-ID` and `CIFL` are initialized by their pre-trained weights from training on synthetic data as described in Secs. 3.1 and 3.2.

For the state-of-the-art comparisons, we consider `FastSurfer` [24] and `SAMSEG` [13,46] for brain segmentation, and `SynthSR` [27] for anatomy reconstruction/contrast synthesis and super-resolution for T1-weighted images.

- `FastSurfer` [24] is a state-of-the-art brain segmentation model, which is designed for segmentation on T1-weighted images. Therefore, we only report the segmentation performance of `FastSurfer` on T1-weighted MRI. In addition, since `FastSurfer` does not predict cerebrospinal fluid (CSF), we remove the CSF label during the `Dice` score computation.
- `SAMSEG` [13,46] is a state-of-the-art, multi-modal brain segmentation model, which works on both MR and CT images. Similar to `FastSurfer`, `SAMSEG` does not predict the CSF label either, the CSF label is therefore removed during the `Dice` score computation of `SAMSEG`.
- `SynthSR` [27] is a state-of-the-art, contrast-agnostic model for anatomy reconstruction/contrast synthesis. For input MRI images with any contrast and resolution, `SynthSR` generates their corresponding high-resolution, 1 mm isotropic T1-weighted MRI. In our comparisons, we apply `SynthSR` on our anatomy reconstruction/contrast synthesis task, as well as the image super-resolution task of T1-weighted MRI.

E Additional Experimental Results

Feature Representation Learning As discussed in Sec. 4.5, in `Brain-ID` we adopt the high-resolution MP-RAGE scan as the anatomy guidance for brain feature representation. Experimental comparisons in Tab. 3 illustrate that incorporating segmentation as the target for anatomical supervision in learning brain feature representation leads to reduced high-frequency texture compared to the use of MP-RAGE alone. The visual comparisons presented in Fig. E.1 reveal that the features with segmentation guidance indeed encompass anatomical structures, however, they exhibit notably smoother texture *within* each structural region defined by the brain segmentation labels. In contrast, employing MP-RAGE as the target for anatomical supervision inherently entails the tasks of anatomy reconstruction and super-resolution simultaneously. The resulting features from `Brain-ID` are shown to carry richer information content, as evidenced by the more pronounced high-frequency textures they manifest.

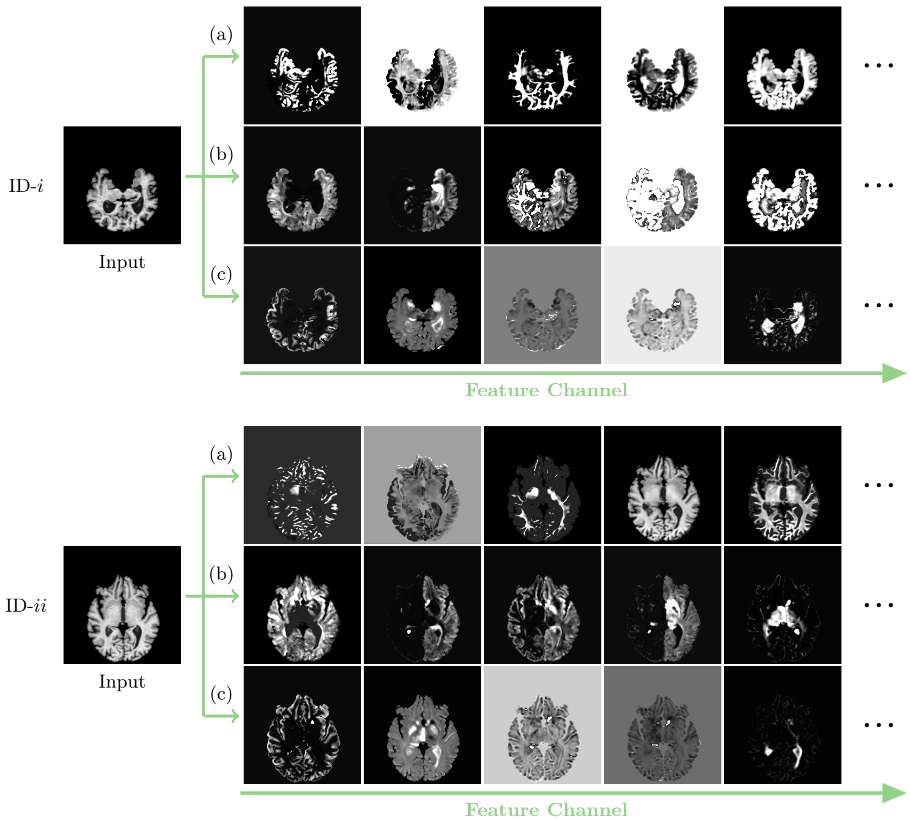

Fig. E.1: Visualizations of features from (a) `Brain-ID` (MP-RAGE guided) with its variants: (b) segmentation guided and (c) segmentation + MP-RAGE guided feature representation models. Note that although the two testing subjects here for the three models are the same, their respective selected feature channels are different, for the purpose of better showing different frequency levels of features from each model.

Downstream Task Evaluation In addition to the qualitative results in Figs. 1 and 5, we provide more visualization comparisons of the downstream tasks in Figs. E.2 to E.4.

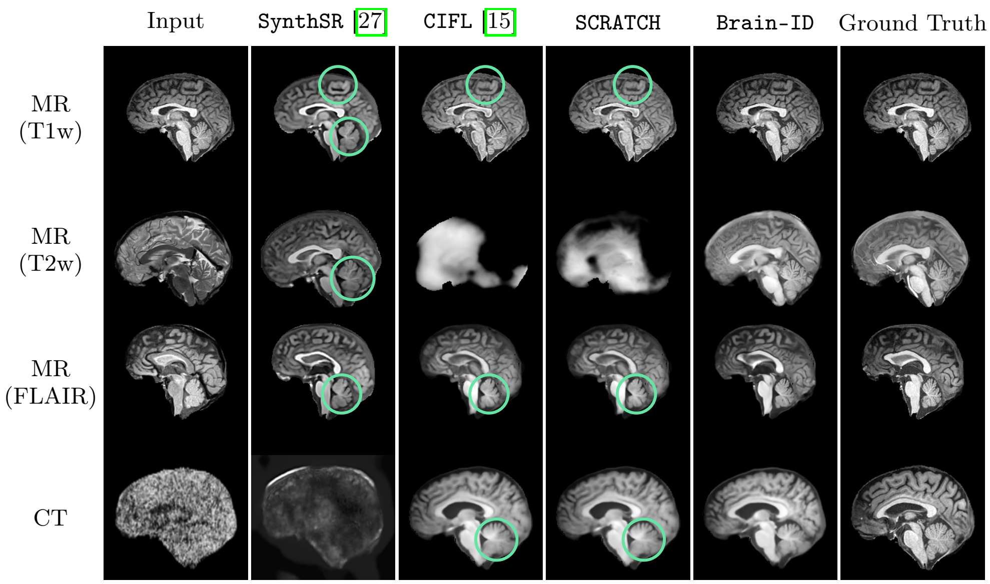

Fig. E.2: Qualitative comparisons on the downstream task of anatomy reconstruction/contrast synthesis, between `Brain-ID`, the baseline `SCRATCH`, and the state-of-the-art methods `CIFL` [15], `SynthSR` [27]. Each row presents the comparison results of inputs with their respective modality/contrast, as indicated in the listing. The mint circles highlight some less noticeable details.

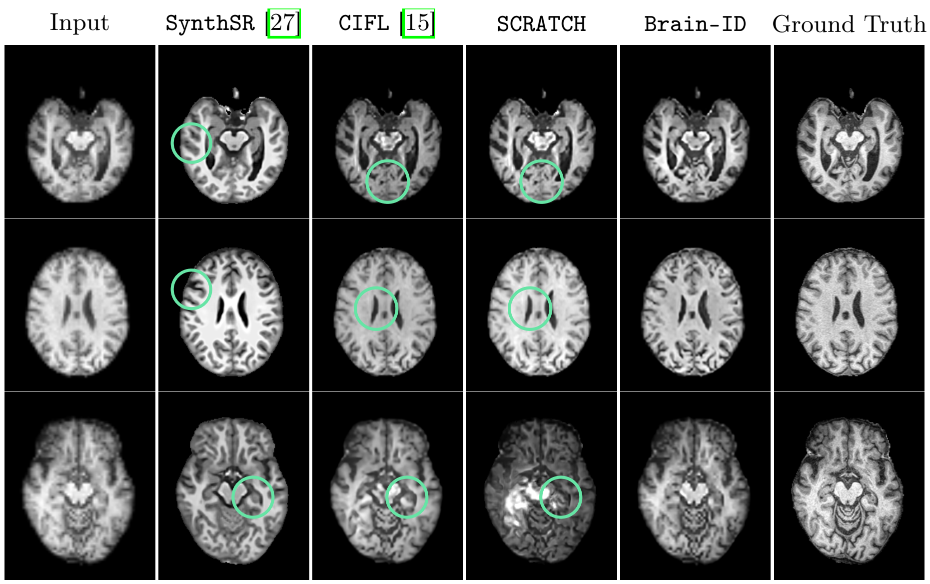

Fig. E.3: Qualitative comparisons on the downstream task of image super-resolution, between Brain-ID, the baseline SCRATCH, and the state-of-the-art methods CIFL [15], SynthSR [27]. Each row corresponds to a different testing subject. The mint circles highlight some less noticeable details.

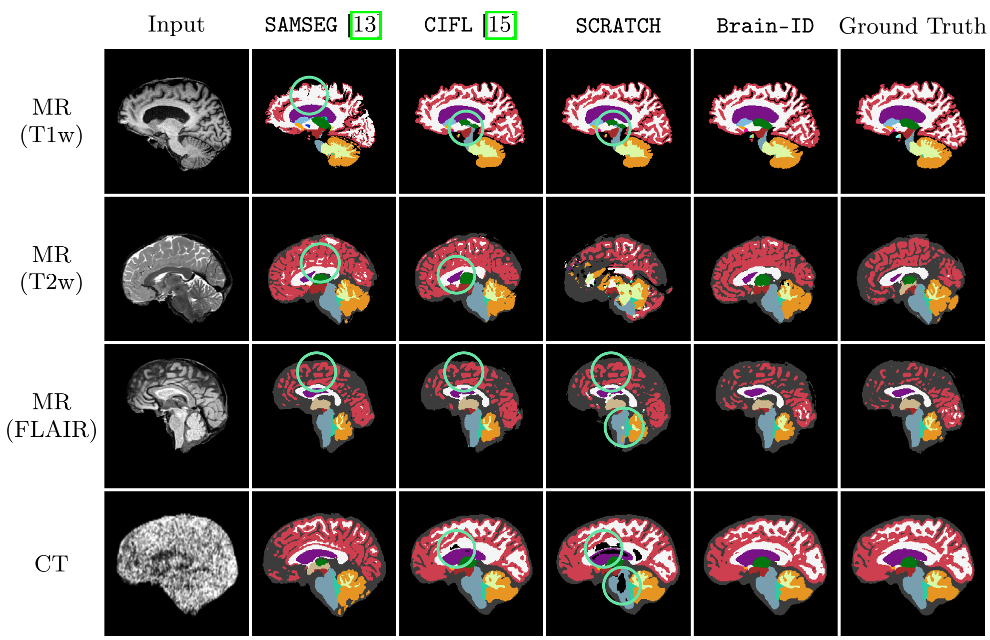

Fig. E.4: Qualitative comparisons on the downstream task of brain segmentation, between Brain-ID, the baseline SCRATCH, and the state-of-the-art methods CIFL [15], SAMSEG [13]. Each row presents the comparison results of inputs with their respective modality/contrast, as indicated in the listing. The mint circles highlight some less noticeable details.