Equivariance Allows Handling Multiple Nuisance Variables When Analyzing Pooled Datasets

SUPPLEMENTARY MATERIAL

1. Proofs of theoretical results

In this section, we will provide the proofs of Lemma 4 and Lemma 5 discussed in the main paper.

Lemma. *Given two latent space representations* $\boldsymbol{\ell}_i, \boldsymbol{\ell}_j \in \mathbf{S}^{n-1}$*, and the corresponding cosets* $g_iH = \tau(\boldsymbol{\ell}_i)$ *and* $g_jH = \tau(\boldsymbol{\ell}_j)$*,* $\exists! g_{ij} = g_j g_i^{-1} \in G$ *such that* $\boldsymbol{\ell}_j = g_{ij} \cdot \boldsymbol{\ell}_i$*.*

Proof. Given $g_iH = \tau(\boldsymbol{\ell}_i)$ and $g_jH = \tau(\boldsymbol{\ell}_j)$, we use $g_{ij} = g_i g_j^{-1} \in G$ such that, $g_jH = g_{ij}g_iH$.

Now using the equivariance fact (3) , we get,

$$\begin{aligned} &g_jH = g_{ij}g_iH \\ &\implies \tau(\boldsymbol{\ell}_j) = g_{ij}\tau(\boldsymbol{\ell}_i) \\ &\implies \tau(\boldsymbol{\ell}_j) = \tau(g_{ij} \cdot \boldsymbol{\ell}_i) \end{aligned}$$

Now as τ is an identification, i.e., a diffeomorphism, we get $\boldsymbol{\ell}_j = g_{ij}\boldsymbol{\ell}_i$. Note that $\mathbf{S}^{n-1}$ is a Riemannian homogeneous space and the group G acts transitively on $\mathbf{S}^{n-1}$, i.e., given $\mathbf{x}, \mathbf{y} \in \mathbf{S}^{n-1}$, $\exists g \in G$ such that, $\mathbf{y} = g \cdot \mathbf{x}$. Hence from $\boldsymbol{\ell}_j = g_{ij}\boldsymbol{\ell}_i$ and the transitivity property we can conclude that g_{ij} is unique. □

Lemma. *For a* $\tau : \mathcal{L} \to G/H$ *as defined above, and a mapping* $b : \mathcal{L} \to \mathcal{Z}$*, the function* $\Phi : \mathcal{L} \to \mathcal{Z}$ *defined by*

$$\Phi(\boldsymbol{\ell}) = \tau(\boldsymbol{\ell}) \cdot b\left(\tau(\boldsymbol{\ell})^{-1} \cdot \boldsymbol{\ell}\right) \tag{1}$$

is G-equivariant, i.e., $\Phi(g \cdot \boldsymbol{\ell}) = g\Phi(\boldsymbol{\ell})$*.*

Proof. Let $\boldsymbol{\ell} \in \mathcal{L}$. Consider the Φ mapping of $g \cdot \boldsymbol{\ell}$, that is $\Phi(g \cdot \boldsymbol{\ell}) = \tau(g \cdot \boldsymbol{\ell}) \cdot b\left(\tau(g \cdot \boldsymbol{\ell})^{-1} \cdot \boldsymbol{\ell}\right)$.

Using the fact (3) from the main paper, we have $\tau(g \cdot \boldsymbol{\ell}) = g\tau(\boldsymbol{\ell})$ and $\tau(g \cdot \boldsymbol{\ell})^{-1} = \tau(\boldsymbol{\ell})^{-1}g^{-1}$. Substituting these in $\Phi(g \cdot \boldsymbol{\ell})$, we get

$$\begin{aligned} \Phi(g \cdot \boldsymbol{\ell}) &= g\tau(\boldsymbol{\ell}) \cdot b\left(\tau(\boldsymbol{\ell})^{-1}g^{-1}g \cdot \boldsymbol{\ell}\right) \\ &= g\tau(\boldsymbol{\ell})b\left(\tau(\boldsymbol{\ell})^{-1} \cdot \boldsymbol{\ell}\right) \end{aligned}$$

Thus, $\Phi(g \cdot \boldsymbol{\ell}) = g\Phi(\boldsymbol{\ell})$

□

2. Details on Evaluation Metrics

Recall from Section 4 of the paper, our discussion on three metrics – $\boldsymbol{\Delta_{Eq}}$, $\boldsymbol{Adv}$ and $\boldsymbol{\mathcal{M}}$. While $\boldsymbol{\Delta_{Eq}}$ and $\boldsymbol{\mathcal{M}}$ are variants of distance measure on the latent space, $\boldsymbol{Adv}$ assesses the ability to predict the nuisance attributes from the latent representation (and is therefore probabilistic in nature). Observe that $\boldsymbol{\Delta_{Eq}}$ and $\boldsymbol{\mathcal{M}}$ are (euclidean) distance measures and could be very different depending on the normalization of the vectors. For our purposes of evaluating these latent vectors/features in downstream tasks, we perform a simple feature normalization in order to obtain $0-1$ latent vectors given by,

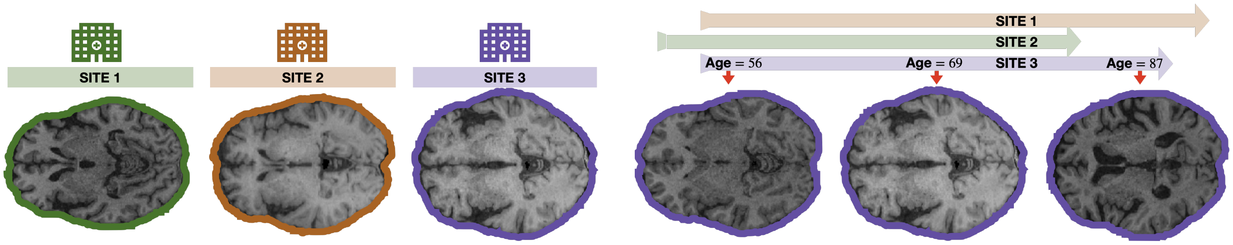
$$\tilde{z_i} = \frac{z_i - \min(z_i)}{\max(z_i) - \min(z_i)}. \tag{2}$$

Our feature normalization is composed of two steps: (i) centering – the numerator in (2) ensures that the mean of z (along its coordinates) is 0; and (ii) scale – the denominator projects the features z on the sphere at origin with radius $\|z_i\|_\infty^{\gtrless} = \max(z_i) - \min(z_i) \geq 0$. Note that our scaling step can be thought of as the usual projection in a special case: when z_i is guaranteed to be nonnegative (for example, when z_i represent activations), then $\|z_i\|_\infty^{\gtrless}$ simply corresponds to a lower bound of the usual infinity norm, $\|z\|_\infty$ (hence projection on a scaled ℓ_∞ ball). We adopt this normalization only to compute $\boldsymbol{\Delta_{Eq}}$ and $\boldsymbol{\mathcal{M}}$ measures, and not for model training.

For computing the $\boldsymbol{Adv}$ measure, we follow [7] to train an adversarial neural network predicting the nuisance attributes. We use a three-layered fully connected network with batch normalization and train for 150 epochs. [6] uses similar architecture for the adversaries with different hidden layers of $0, 1, 2, 3$. We found that a three-layer adversary is powerful enough to predict the nuisance attributes and hence we use it to report the $\boldsymbol{Adv}$ measure.

3. Understanding ADNI dataset

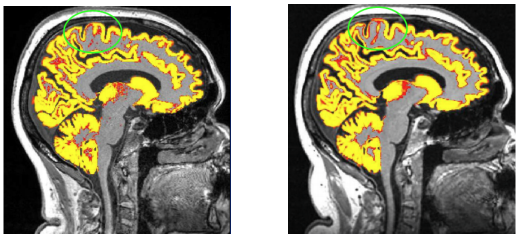
Dataset. The data was downloaded from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investi-



(a) Variation due to scanner for particular age group.

(b) Variation due to covariates (age) in scanner 3.

Figure 1. **Sample Images from ADCP dataset.** **(a)** MRI images on control subjects from the ADCP dataset for different **sites** in the age group 70-80. **(b)** Images obtained from Site 3 for three extreme **age groups**. The gantt chart on top of the image indicates the respective age range in the other sites.



(a) GE

(b) Siemens

Figure 2. **Scanner effects on images.** Two imaging protocols are shown: (a) Siemens, (b) GE. The yellow region is the cortical ribbon segmentation, and the green circle shows that the imaging protocol from different manufacturers have an effect on the scan. Image borrowed from [1].

gator Michael W. Weiner, MD. ADNI was set up with an objective to measure the progression of mild cognitive impairment (MCI) and early Alzheimers disease (AD) using serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers. We have three imaging protocol (scanner) types in the dataset, namely, GE, Siemens and Phillips. The count of samples AD/CN in each of these imaging protocols are provided in Table 1. An example illustration (borrowed from [1]) of using different scanner on the images is shown in Figure 2.

Preprocessing. All images were first normalized and skull-stripped using Freesurfer [3]. A linear (affine) registration was performed to register each image to MNI template space.

4. Understanding ADCP dataset

Participants. The data for ADCP was collected through an NIH-sponsored Alzheimer's Disease Connectome Project (ADCP) U01 AG051216. The study inclusion criteria for AD (Alzheimer's disease) / MCI (Mild Cognitive Impairment) patients consisted of age between 55-90 years, willing and able to undergo all procedures, retains decisional capacity at initial visit, meets criteria for probable AD or meets criteria for MCI.

Scanners. MRI images were acquired at three distinct sites on GE scanners. T1-weighted structural images were acquired using a 3D gradient-echo pulse sequence (repetition time (TR) = 604 ms, echo time (TE) = 2.516 ms, inversion time = 1060 ms, flip angle = 8^o, field of view (FOV) = 25.6 cm, 0.8 mm isotropic). T2-weighted structural images were acquired using a 3D fast spin-echo sequence (TR = 2500 ms, TE = 94.398 ms, flip angle = 90^o, FOV = 25.6 cm, 0.8 mm isotropic).

Preprocessing. The Human Connectome Project (HCP) minimal preprocessing pipeline version 3.4.0 [4] was followed for data processing. This pipeline is based on FMRIB Software Library [5]. Next, the T1w and T2w images are aligned, a B1 (bias field) correction is performed, and the subject's image in native structural volume space is registered to MNI space using FSL's FNIRT [2]. Only T1w images in the MNI space were used for further analysis and experiments.

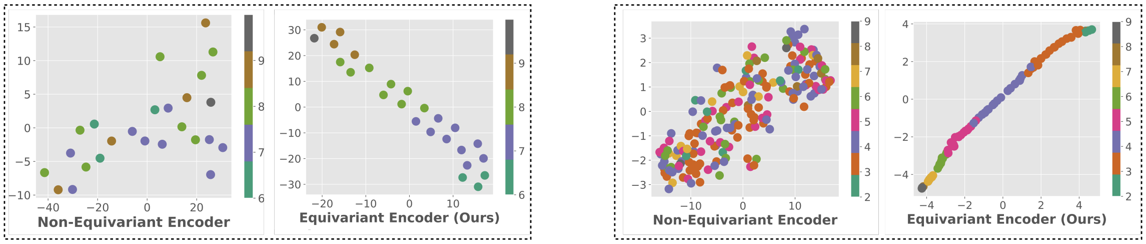
Data Statistics. We plot the distributions of several attributes in this dataset conditioned on the site. In Figure 4, we show that the values of age and cognitive scores differ across the three sites in this dataset. Cognitive scores are computed based on an test assigned to the patients. Higher scores indicate higher cognitive operation in the patient. Table 2 shows the sample counts for target variable of prediction AD (Alzheimer's disease) and Control group.

Table 1. Sample counts for ADNI dataset

Imaging Protocol	AD	CN
Manufacturer=GE Medical Systems	44	78
Manufacturer=Philips Medical Systems	32	50
Manufacturer=Siemens	83	162

Table 2. Sample counts for ADCP dataset

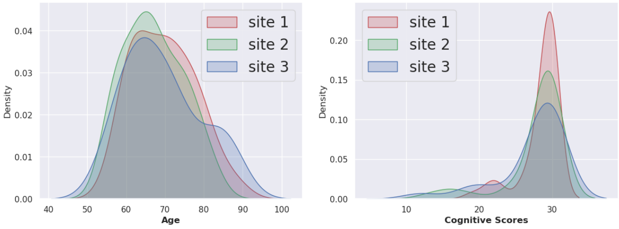
	AD	Control	Female	Male
site 1	10	39	29	20
site 2	10	33	30	13
site 3	5	19	14	10



(a) ADCP Dataset

(b) German Dataset

Figure 3. **t-SNE plots of latent representations of $\tau(\boldsymbol{\ell})$** . For both ADCP **(left)** and German **(right)**, the the latent vectors of the equivariant encoder are evenly distributed with respect to the age covariate value. The non-equivariant space is generated from the naïve pooling model. Different colors denote the discretized set of **age** covariate value present in the data.



(a) Age

(b) Cognitive Score

Figure 4. **Distribution of attributes in the ADCP dataset.** On the **left** we observe the distribution of age for the three different sites present in the ADCP dataset. On the **right**, we see the distribution of the cognitive scores. The cognitive scores are computed based on an ADCP test that assesses executive function. Higher scores indicate higher level of cognitive flexibility. Both age and cognitive scores are observed to vary across the sites.

5. Visualizing the latent space

In the paper Figure 4, we have seen the latent space $\tau(\boldsymbol{\ell})$ for the samples in the ADNI and the Adult datasets. Here, we will see similar qualitative results for the German and the ADCP dataset in Figure 3 of the supplement. In the plots, the latent representations for a non-equivariant encoder are stretched thoughout the latent space. In contrast, the representations of an equivariant encoder, for a discretized value of **Age**, are localized to specific regions. Further, these representations have a monotonic behaviour with respect to the values of **Age**.

6. Hyper-parameters and NN Architectures

For tabular datasets such as German and Adult, our encoders and decoders comprise of fully connected networks and a hidden layer of 64 nodes. The dimension of the quotient latent space $\tau(\boldsymbol{\ell}_i)$ is 30. Adam is used as a default optimizer and the learning rate is adjusted based on the validation set.

Imaging datasets like ADNI and ADCP require 3D convolutions and a ResNet architecture as the backbone. The last layer is used to describe the quotient space $\tau(\boldsymbol{\ell}_i)$. We present the residual and the fully connected block below. Detailed architectures can be viewed in the code.

Listing 1. Residual Block

```
    BatchNorm3d
    Swish
    Conv3d
    BatchNorm3d
    Swish
    Conv3d
```

Listing 2. Fully Connected Block

```
    AdaptiveAvgPool3d
    Flatten
    Dropout
    Linear
    BatchNorm1d
    Swish
    Dropout
    Linear
```

7. A note on multi-objective scaling factors

Recall from the Algorithm 1 of the main paper that our loss function for each stage comprises of reconstruction and prediction losses in addition to the objectives concerning equivariance and invariance. These multi-objective loss functions require scaling factors that upweight one objective over the other. These scaling factors group up as hyperparameters for the Algorithm. In our experiments, it was observed that the results were robust to a range of scaling factor choices. For the results reported in Table 1 of the paper, they were identified through cross-validation. Here we provide an example for the scaling factors used for the Adult dataset, please refer to the bash scripts available in the code for the scaling factors of other datasets.

- **Stage one: Equivariance to Covariates**
  - Equivariance Loss L_{stage1}
    Scaling Factor : 1.0
  - Reconstruction Loss $\sum_i \|X_i - \mathfrak{D}(\mathfrak{E}(X_i))\|$
    Scaling Factor : 0.02
- **Stage two: Invariance to Site**

- Invariance Loss $\mathcal{MMD}$
  Scaling Factor : 0.1
- Prediction Loss $\|Y - h(\Phi(\boldsymbol{\ell}))\|^2$
  Scaling Factor : 1.0
- Reconstruction Loss $\|\boldsymbol{\ell} - \Psi(\Phi(\boldsymbol{\ell}))\|^2$
  Scaling Factor : 0.1

We refer the reader to Algorithm 1 and Section 3 of the main paper for the details on the notations used above.

References

[1] Paul S Aisen, Jeffrey Cummings, Clifford R Jack, John C Morris, Reisa Sperling, Lutz Frölich, Roy W Jones, Sherie A Dowsett, Brandy R Matthews, Joel Raskin, et al. On the path to 2025: understanding the alzheimer's disease continuum. *Alzheimer's research & therapy*, 9(1):1–10, 2017. 2

[2] Jesper LR Andersson, Mark Jenkinson, Stephen Smith, et al. Non-linear registration, aka spatial normalisation fmrib technical report tr07ja2. *FMRIB Analysis Group of the University of Oxford*, 2(1):e21, 2007. 2

[3] Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012. 2

[4] Matthew F Glasser, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R Polimeni, et al. The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124, 2013. 2

[5] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012. 2

[6] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 1

[7] Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 1