Deep Hyperalignment

Supplementary Materials

Muhammad Yousefnezhad & Daoqiang Zhang
College of Computer Science and Technology
Nanjing University of Aeronautics and Astronautics
{myousefnezhad,dqzhang}@nuaa.edu.cn

S.1 Notations

Table S1 illustrates all variables and functions that are employed in this paper.

S.2 Visualizing DHA Common Space ($\mathbf{G}$)

This section presents a visualized example of the DHA common space ($\mathbf{G}$). Here, we have firstly applied DHA to DS105 dataset and generated the common space by considering $V = V_{new}$ same as SRM [5] and CAE [6] methods. As depicted in Fig. S1, the average of neural activities in each voxel is calculated and visualized across categories of visual stimuli.

S.3 Technical Information

Figure S2 illustrates the graphical abstract of DHA, where the hyperalignment applied to a problem with 3-dimensional space for mapping the neural activities to a common space with 2-dimensions. In addition, Table S2 demonstrates the technical information of the employed datasets in this paper.

S.4 Overview of Hyperalignment methods

Preprocessed fMRI time series collected for S subjects can be defined by $\mathbf{X}^{(i)} = \left\{x_{mn}^{(i)}\right\} \in \mathbb{R}^{T \times V}$, $i = 1{:}S, m = 1{:}T, n = 1{:}V$, where T denotes the number of time points in unites of TRs (Time of Repetition), V is the number of voxels, and $x_{mn}^{(i)} \in \mathbb{R}$ denotes the functional activity for the $i\text{-}th$ subject in the $m\text{-}th$ time point and the $n\text{-}th$ voxel. For simplicity assume that all data points are normalized by zero-mean and unit-variance, $\mathbf{X}^{(i)} \sim \mathcal{N}(0, 1)$, $i = 1{:}S$. If the original data points are not normalized, we can consider this assumption as a preprocessing step. Since there are more voxels than TRs in most of the fMRI studies, $\mathbf{X}^{(i)}$ and the voxel correlation map $(\mathbf{X}^{(i)})^\top \mathbf{X}^{(j)}$ may not be full rank [1–4]. In training-set, time synchronized stimulus ensures temporal alignment, i.e. the $m\text{-}th$ time point for all of the subjects represents the same simulation [2, 3]. Indeed, the main goal of HA methods is aligning the columns of $\mathbf{X}^{(i)}$ across subjects [2]. In previous studies, Inter-Subject Correlation (ISC) was defined for functional alignment between two distinct subjects as follows, where $tr()$ is the trace function: [1–4]

$$\text{ISC}(\mathbf{X}^{(i)}, \mathbf{X}^{(j)}) = ({}^{1}\!/\!{}_{V})\text{tr}\Big((\mathbf{X}^{(i)})^\top \mathbf{X}^{(j)}\Big). \tag{S.1}$$

By considering $\mathbf{X}^{(\ell)} \sim \mathcal{N}(0, 1)$, $\ell = 1{:}S$ as column-wise standardized, ISC lies in $[-1, +1]$. Here, the large values illustrate better alignment [2, 4]. Based on (S.1), Hyperalignment (HA) problem can

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

be defined as follows: [1–3]

$$\max_{\mathbf{R}^{(i)},\mathbf{R}^{(j)}} \sum_{i=1}^{S} \sum_{j=i+1}^{S} \text{ISC}(\mathbf{X}^{(i)}\mathbf{R}^{(i)}, \mathbf{X}^{(j)}\mathbf{R}^{(j)}) \qquad \text{s.t.} \quad \big(\mathbf{R}^{(\ell)}\big)^{\top}\widetilde{\mathbf{\Phi}}^{(\ell)}\mathbf{R}^{(\ell)} = \mathbf{I}, \ell = 1{:}S, \tag{S.2}$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{R}^{(\ell)} = \left\{r_{mn}^{(\ell)}\right\} \in \mathbb{R}^{V \times V}$ denotes the solution for ℓ-th subject, and the matrices $\widetilde{\mathbf{\Phi}}^{(\ell)} \in \mathbb{R}^{V \times V}, \ell = 1{:}S$ are symmetric and positive definite. By considering $\widetilde{\mathbf{\Phi}}^{(\ell)} = \mathbf{I}$, (S.2) is equivalent to a multi-set orthogonal Procrustes problem, which is commonly used in share analysis. Furthermore, if $\widetilde{\mathbf{\Phi}}^{(\ell)} = (\mathbf{X}^{(\ell)})^{\top}\mathbf{X}^{(\ell)}$, then (S.2) denotes a form of multi-set Canonical Correlation Analysis (CCA) [3, 4] as follows:

$$\max_{\mathbf{R}^{(i)},\mathbf{R}^{(j)}} \sum_{i=1}^{S} \sum_{j=i+1}^{S} \text{tr}\Big((\mathbf{X}^{(i)}\mathbf{R}^{(i)})^{\top}\mathbf{X}^{(j)}\mathbf{R}^{(j)}\Big) \quad \text{s.t.} \quad \big(\mathbf{X}^{(\ell)}\mathbf{R}^{(\ell)}\big)^{\top}\mathbf{X}^{(\ell)}\mathbf{R}^{(\ell)} = \mathbf{I}, \quad \ell = 1{:}S. \tag{S.3}$$

Lemma S.1. (S.3) *maximizes the Pearson correlation ($cov(A,B)/\sigma_A\sigma_B$) between each pair of mapped features ($\mathbf{X}^{(\ell)}\mathbf{R}^{(\ell)}$) across subjects.*

Proof. By considering $\widetilde{\mathbf{\Phi}}^{(\ell)} \in \mathbb{R}^{V \times V} = \mathbb{E}\Big[(\mathbf{X}^{(\ell)})^{\top}\mathbf{X}^{(\ell)}\Big] = (\mathbf{X}^{(\ell)})^{\top}\mathbf{X}^{(\ell)}$, the Pearson correlation on (S.3) can be defined as follows:

$$\max_{\mathbf{R}^{(i)},\mathbf{R}^{(j)}} \sum_{i=1}^{S} \sum_{j=i+1}^{S} \text{tr}\Bigg(\frac{(\mathbf{X}^{(i)}\mathbf{R}^{(i)})^{\top}\mathbf{X}^{(i)}\mathbf{R}^{(j)}}{\sqrt{((\mathbf{R}^{(i)})^{\top}\widetilde{\mathbf{\Phi}}^{(i)}\mathbf{R}^{(i)})}\sqrt{((\mathbf{R}^{(j)})^{\top}\widetilde{\mathbf{\Phi}}^{(j)}\mathbf{R}^{(j)})}}\Bigg). \tag{S.4}$$

Since $\big(\mathbf{R}^{(\ell)}\big)^{\top}\widetilde{\mathbf{\Phi}}^{(\ell)}\mathbf{R}^{(\ell)} = \mathbf{I}$, we approach from (S.4) to (S.3).

In fact, the constrains must be imposed in $\mathbf{R}^{(\ell)}$ to avoid overfitting [2]. In order to seek an optimum solution, solving (S.3) may not be the best approach because there is no scale to evaluate the distance between current result and the optimum (fully maximized) solution [2, 4, 7].

Lemma S.2. (S.3) *can be rewritten as following minimization problem:*

$$\min_{\mathbf{R}^{(i)},\mathbf{R}^{(j)}} \sum_{i=1}^{S} \sum_{j=i+1}^{S} \Big\|\mathbf{X}^{(i)}\mathbf{R}^{(i)} - \mathbf{X}^{(j)}\mathbf{R}^{(j)}\Big\|_F^2, \quad \text{s.t.} \quad \Big(\mathbf{X}^{(\ell)}\mathbf{R}^{(\ell)}\Big)^{\top}\mathbf{X}^{(\ell)}\mathbf{R}^{(\ell)} = \mathbf{I}, \quad \ell = 1{:}S, \tag{S.5}$$

where (S.5) *approaches zero for an optimum result.*

Proof.

$$\min_{\mathbf{R}^{(i)},\mathbf{R}^{(j)}} \sum_{i=1}^{S} \sum_{j=i+1}^{S} \Big\|\mathbf{X}^{(i)}\mathbf{R}^{(i)} - \mathbf{X}^{(j)}\mathbf{R}^{(j)}\Big\|_F^2 = \min_{\mathbf{R}^{(i)},\mathbf{R}^{(j)}} \sum_{i=1}^{S} \sum_{j=i+1}^{S} \Bigg(\text{tr}\Big(\big(\mathbf{X}^{(i)}\mathbf{R}^{(i)}\big)^{\top}\mathbf{X}^{(i)}\mathbf{R}^{(i)}\Big) +$$

$$\text{tr}\Big(\big(\mathbf{X}^{(j)}\mathbf{R}^{(j)}\big)^{\top}\mathbf{X}^{(j)}\mathbf{R}^{(j)}\Big) - 2\text{tr}\Big(\big(\mathbf{X}^{(i)}\mathbf{R}^{(i)}\big)^{\top}\mathbf{X}^{(j)}\mathbf{R}^{(j)}\Big)\Bigg).$$

Since $\Big(\mathbf{X}^{(\ell)}\mathbf{R}^{(\ell)}\Big)^{\top}\mathbf{X}^{(\ell)}\mathbf{R}^{(\ell)} = \mathbf{I}$, so we have:

$$\min_{\mathbf{R}^{(i)},\mathbf{R}^{(j)}} \sum_{i=1}^{S} \sum_{j=i+1}^{S} \Bigg(2V_{new} - 2\text{tr}\Big(\big(\mathbf{X}^{(i)}\mathbf{R}^{(i)}\big)^{\top}\mathbf{X}^{(j)}\mathbf{R}^{(j)}\Big)\Bigg) \equiv \max_{\mathbf{R}^{(i)},\mathbf{R}^{(j)}} \sum_{i=1}^{S} \sum_{j=i+1}^{S} \text{tr}\Big((\mathbf{X}^{(i)}\mathbf{R}^{(i)})^{\top}\mathbf{X}^{(j)}\mathbf{R}^{(j)}\Big).$$

Indeed, the main assumption in the original HA is that the $\mathbf{R}^{(\ell)}, \ell = 1{:}S$ are noisy 'rotations' of a common template [1, 9]. In order to extend HA into the real-world problems, (S.5) can be reformulated by using an embedded space as follows:

$$\min_{\mathbf{R}^{(i)},\mathbf{R}^{(j)}} \sum_{i=1}^{S} \sum_{j=i+1}^{S} \Big\|f(\mathbf{X}^{(i)})\mathbf{R}^{(i)} - f(\mathbf{X}^{(j)})\mathbf{R}^{(j)}\Big\|_F^2, \text{s.t.} \ \Big(f(\mathbf{X}^{(\ell)})\mathbf{R}^{(\ell)}\Big)^{\top}f(\mathbf{X}^{(\ell)})\mathbf{R}^{(\ell)} = \mathbf{I}, \ell = 1{:}S. \tag{S.6}$$

Here, if $f\big(\mathbf{x}\big) = \mathbf{x}$, then we recover the original HA. Further, if $f\big(\mathbf{x}\big)$ denotes any classical fixed kernel function (e.g. Gaussian), then (S.6) is equivalent to Kernel Hyperalignment (KHA) [3]. In addition, if $f\big(\mathbf{x}\big)$ illustrates an SVD-based feature selection, then (S.6) is called SVD Hyperalignment (SVDHA) [4]. Finally, this paper proposes Deep Hyperalignment (DHA), where $f_\ell\big(\mathbf{X}^{(\ell)};\theta^{(\ell)}\big)$ is defined (rather than $f\big(\mathbf{x}\big)$) as a parametric deep neural network across subjects, i.e. multiple stacked layers of nonlinear transformation for each subject. As another alternative, (S.6) can be optimized by using different approaches. For instance, Regularized Hyperalignment (RHA) used the Generalized CCA $\Big(\min\limits_{\mathbf{G},\mathbf{R}^{(i)}} \sum_{i=1}^{S} \Big\|\mathbf{G} - \mathbf{X}^{(i)}\mathbf{R}^{(i)}\Big\|_F^2\Big)$ [2], Shared Response Model (SRM) employed Probabilistic CCA [5]. This paper also uses rank-m SVD [9] and Stochastic Gradient Descent (SGD) [12] for optimization.

S.5 Equations

This section presents the equations of the original paper that we need them in order to proof Lemmas in the next section.

1. Objective function of DHA is defined as follows:

$$\begin{aligned}
&\min_{\substack{\theta^{(i)},\mathbf{R}^{(i)}\\ \theta^{(j)},\mathbf{R}^{(j)}}} \sum_{i=1}^{S}\sum_{j=i+1}^{S} \Big\| f_i\big(\mathbf{X}^{(i)};\theta^{(i)}\big)\mathbf{R}^{(i)} - f_j\big(\mathbf{X}^{(j)};\theta^{(j)}\big)\mathbf{R}^{(j)} \Big\|_F^2 \\
&\text{s.t.} \quad \Big(\mathbf{R}^{(\ell)}\Big)^\top \Big(\Big(f_\ell\big(\mathbf{X}^{(\ell)};\theta^{(\ell)}\big)\Big)^\top f_\ell\big(\mathbf{X}^{(\ell)};\theta^{(\ell)}\big) + \epsilon\mathbf{I}\Big)\mathbf{R}^{(\ell)} = \mathbf{I}, \quad \ell = 1{:}S.
\end{aligned} \tag{3}$$

2. Projection matrix for ℓ-th subject can be generated as follows: [10]

$$\begin{aligned}
\mathbf{P}^{(\ell)} &= f_\ell\big(\mathbf{X}^{(\ell)};\theta^{(\ell)}\big)\Big(\Big(f_\ell\big(\mathbf{X}^{(\ell)};\theta^{(\ell)}\big)\Big)^\top f_\ell\big(\mathbf{X}^{(\ell)};\theta^{(\ell)}\big) + \epsilon\mathbf{I}\Big)^{-1}\Big(f_\ell\big(\mathbf{X}^{(\ell)};\theta^{(\ell)}\big)\Big)^\top \\
&= \mathbf{\Omega}^{(\ell)}\big(\mathbf{\Sigma}^{(\ell)}\big)^\top\Big(\mathbf{\Sigma}^{(\ell)}\big(\mathbf{\Sigma}^{(\ell)}\big)^\top + \epsilon\mathbf{I}\Big)^{-1}\mathbf{\Sigma}^{(\ell)}\big(\mathbf{\Omega}^{(\ell)}\big)^\top = \mathbf{\Omega}^{(\ell)}\mathbf{D}^{(\ell)}\Big(\mathbf{\Omega}^{(\ell)}\mathbf{D}^{(\ell)}\Big)^\top,
\end{aligned} \tag{8}$$

where diagonal matrix $\mathbf{D}^{(\ell)} \in \mathbb{R}^{m\times m}$ is

$$\mathbf{D}^{(\ell)}\big(\mathbf{D}^{(\ell)}\big)^\top = \big(\mathbf{\Sigma}^{(\ell)}\big)^\top\Big(\mathbf{\Sigma}^{(\ell)}\big(\mathbf{\Sigma}^{(\ell)}\big)^\top + \epsilon\mathbf{I}\Big)^{-1}\mathbf{\Sigma}^{(\ell)}. \tag{9}$$

3. The sum of projection can be defined as follows:

$$\mathbf{A} = \sum_{i=1}^{S}\mathbf{P}^{(i)} = \widetilde{\mathbf{A}}\big(\widetilde{\mathbf{A}}\big)^\top, \quad \text{where} \quad \widetilde{\mathbf{A}} \in \mathbb{R}^{T\times mS} = \big[\mathbf{\Omega}^{(1)}\mathbf{D}^{(1)}\ldots\mathbf{\Omega}^{(S)}\mathbf{D}^{(S)}\big]. \tag{10}$$

4. DHA mapping for ℓ-th subject is denoted as follows:

$$\mathbf{R}^{(\ell)} = \Big(\Big(f_\ell\big(\mathbf{X}^{(\ell)};\theta^{(\ell)}\big)\Big)^\top f_\ell\big(\mathbf{X}^{(\ell)};\theta^{(\ell)}\big) + \epsilon\mathbf{I}\Big)^{-1}\Big(f_\ell\big(\mathbf{X}^{(\ell)};\theta^{(\ell)}\big)\Big)^\top\mathbf{G}. \tag{12}$$

S.6 Proofs

Lemma 1. *The equation* (3) *can be reformulated as follows where* $\mathbf{G} \in \mathbb{R}^{T\times V_{new}}$ *is the HA template:*

$$\min_{\mathbf{G},\mathbf{R}^{(i)},\theta^{(i)}} \sum_{i=1}^{S}\Big\|\mathbf{G} - f_i\big(\mathbf{X}^{(i)};\theta^{(i)}\big)\mathbf{R}^{(i)}\Big\|_F^2 \quad \text{s.t.} \quad \mathbf{G}^\top\mathbf{G} = \mathbf{I}, \text{where } \mathbf{G} = \frac{1}{S}\sum_{j=1}^{S} f_j\big(\mathbf{X}^{(j)};\theta^{(j)}\big)\mathbf{R}^{(j)}. \tag{6}$$

Proof. For simplicity, we consider that $f_\ell\big(\mathbf{X}^{(\ell)};\theta^{(\ell)}\big) = \mathbf{Y}^{(\ell)}$:

$$\min_{\substack{\theta^{(i)},\mathbf{R}^{(i)}\\ \theta^{(j)},\mathbf{R}^{(j)}}} \sum_{i=1}^{S}\sum_{j=i+1}^{S} \Big\| f_i\big(\mathbf{X}^{(i)};\theta^{(i)}\big)\mathbf{R}^{(i)} - f_j\big(\mathbf{X}^{(j)};\theta^{(j)}\big)\mathbf{R}^{(j)}\Big\|_F^2 = \min_{\substack{\theta^{(i)},\mathbf{R}^{(i)}\\ \theta^{(j)},\mathbf{R}^{(j)}}} \sum_{i=1}^{S}\sum_{j=i+1}^{S} \Big\|\mathbf{Y}^{(i)}\mathbf{R}^{(i)} - \mathbf{Y}^{(j)}\mathbf{R}^{(j)}\Big\|_F^2$$

$$= \min_{\substack{\theta^{(i)},\mathbf{R}^{(i)}\\ \theta^{(j)},\mathbf{R}^{(j)}}} \sum_{i=1}^{S}\sum_{j=i+1}^{S} \bigg(\mathrm{tr}\Big(\big(\mathbf{Y}^{(i)}\mathbf{R}^{(i)}\big)^\top\mathbf{Y}^{(i)}\mathbf{R}^{(i)}\Big) + \mathrm{tr}\Big(\big(\mathbf{Y}^{(j)}\mathbf{R}^{(j)}\big)^\top\mathbf{Y}^{(j)}\mathbf{R}^{(j)}\Big) - 2\mathrm{tr}\Big(\big(\mathbf{Y}^{(i)}\mathbf{R}^{(i)}\big)^\top\mathbf{Y}^{(j)}\mathbf{R}^{(j)}\Big)\bigg)$$

$$\equiv \min_{\substack{\theta^{(i)},\mathbf{R}^{(i)}\\ \theta^{(j)},\mathbf{R}^{(j)}\\ \mathbf{G}}} \frac{1}{2}\sum_{i=1}^{S} \bigg(S\mathrm{tr}\Big(\big(\mathbf{Y}^{(i)}\mathbf{R}^{(i)}\big)^\top\mathbf{Y}^{(i)}\mathbf{R}^{(i)}\Big) - 2S\mathrm{tr}\Big(\big(\mathbf{Y}^{(i)}\mathbf{R}^{(i)}\big)^\top\mathbf{G}\Big) + \sum_{j=1}^{S}\Big(\mathrm{tr}\Big(\big(\mathbf{Y}^{(j)}\mathbf{R}^{(j)}\big)^\top\mathbf{Y}^{(j)}\mathbf{R}^{(j)}\Big)\Big)\bigg)$$

$$= \min_{\substack{\theta^{(i)},\mathbf{R}^{(i)}\\ \theta^{(j)},\mathbf{R}^{(j)}\\ \mathbf{G}}} \frac{1}{2}\bigg(\bigg(S\sum_{i=1}^{S}\mathrm{tr}\Big(\big(\mathbf{Y}^{(i)}\mathbf{R}^{(i)}\big)^\top\mathbf{Y}^{(i)}\mathbf{R}^{(i)}\Big)\bigg) - \Big(2S^2\mathrm{tr}\big(\mathbf{G}^\top\mathbf{G}\big)\Big) + \bigg(S\sum_{j=1}^{S}\mathrm{tr}\Big(\big(\mathbf{Y}^{(j)}\mathbf{R}^{(j)}\big)^\top\mathbf{Y}^{(j)}\mathbf{R}^{(j)}\Big)\bigg)\bigg)$$

$$= \min_{\substack{\theta^{(i)},\mathbf{R}^{(i)}\\ \mathbf{G}}} \frac{1}{2}\bigg(\bigg(2S\sum_{i=1}^{S}\mathrm{tr}\Big(\big(\mathbf{Y}^{(i)}\mathbf{R}^{(i)}\big)^\top\mathbf{Y}^{(i)}\mathbf{R}^{(i)}\Big) - \Big(2S^2\mathrm{tr}\big(\mathbf{G}^\top\mathbf{G}\big)\Big)\bigg)$$

$$= \min_{\substack{\theta^{(i)},\mathbf{R}^{(i)}\\ \mathbf{G}}} \bigg(\bigg(S\sum_{i=1}^{S}\mathrm{tr}\Big(\big(\mathbf{Y}^{(i)}\mathbf{R}^{(i)}\big)^\top\mathbf{Y}^{(i)}\mathbf{R}^{(i)}\Big)\bigg) - \Big(S^2\mathrm{tr}\big(\mathbf{G}^\top\mathbf{G}\big)\Big)\bigg)$$

$$= \min_{\substack{\theta^{(i)},\mathbf{R}^{(i)}\\ \mathbf{G}}} \bigg(\bigg(S\sum_{i=1}^{S}\mathrm{tr}\Big(\big(\mathbf{Y}^{(i)}\mathbf{R}^{(i)}\big)^\top\mathbf{Y}^{(i)}\mathbf{R}^{(i)}\Big)\bigg) - 2\Big(S^2\mathrm{tr}\big(\mathbf{G}^\top\mathbf{G}\big)\Big) + \Big(S^2\mathrm{tr}\big(\mathbf{G}^\top\mathbf{G}\big)\Big)\bigg)$$

$$= \min_{\substack{\theta^{(i)},\mathbf{R}^{(i)}\\ \mathbf{G}}} S\sum_{i=1}^{S}\bigg(\mathrm{tr}\Big(\big(\mathbf{Y}^{(i)}\mathbf{R}^{(i)}\big)^\top\mathbf{Y}^{(i)}\mathbf{R}^{(i)}\Big) + \mathrm{tr}\big(\mathbf{G}^\top\mathbf{G}\big) - 2\mathrm{tr}\Big(\big(\mathbf{Y}^{(i)}\mathbf{R}^{(i)}\big)^\top\mathbf{G}\Big)\bigg)$$

$$\equiv \min_{\substack{\theta^{(i)},\mathbf{R}^{(i)}\\ \mathbf{G}}} \sum_{i=1}^{S}\|\mathbf{G} - \mathbf{Y}^{(i)}\mathbf{R}^{(i)}\|_F^2 = \min_{\mathbf{G},\mathbf{R}^{(i)},\theta^{(i)}} \sum_{i=1}^{S}\Big\|\mathbf{G} - f_i\big(\mathbf{X}^{(i)};\theta^{(i)}\big)\mathbf{R}^{(i)}\Big\|_F^2 \blacksquare$$

Lemma 2. *Based on* (10)*, the objective function of DHA* (6) *can be rewritten as follows:*

$$\min_{\mathbf{G},\mathbf{R}^{(i)},\theta^{(i)}} \sum_{i=1}^{S}\Big\|\mathbf{G} - f_i\big(\mathbf{X}^{(i)};\theta^{(i)}\big)\mathbf{R}^{(i)}\Big\| \equiv \max_{\mathbf{G}}\Big(\mathrm{tr}\big(\mathbf{G}^\top\mathbf{A}\mathbf{G}\big)\Big). \tag{10}$$

Proof. For simplicity, we consider that $f_\ell\big(\mathbf{X}^{(\ell)};\theta^{(\ell)}\big) = \mathbf{Y}^{(\ell)}$. By considering (12), we have:

$$\min_{\mathbf{G},\theta^{(i)},\mathbf{R}^{(i)}} \sum_{i=1}^{S}\Big\|\mathbf{G} - \mathbf{Y}^{(i)}\Big(\big(\mathbf{Y}^{(i)}\big)^\top\mathbf{Y}^{(i)} + \epsilon\mathbf{I}\Big)^{-1}\big(\mathbf{Y}^{(i)}\big)^\top\mathbf{G}\Big\|_F^2$$

Based on (8), we have:

$$\min_{\mathbf{G}}\sum_{i=1}^{S}\|\mathbf{G} - \mathbf{P}^{(i)}\mathbf{G}\|_F^2 = \min_{\mathbf{G}}\sum_{i=1}^{S}\|\big(\mathbf{I} - \mathbf{P}^{(i)}\big)\mathbf{G}\|_F^2 = \min_{\mathbf{G}}\sum_{i=1}^{S}\mathrm{tr}\bigg(\Big(\big(\mathbf{I} - \mathbf{P}^{(i)}\big)\mathbf{G}\Big)^\top\big(\mathbf{I} - \mathbf{P}^{(i)}\big)\mathbf{G}\bigg)$$

$$= \min_{\mathbf{G}}\sum_{i=1}^{S}\mathrm{tr}\Big(\mathbf{G}^\top\big(\mathbf{I} - \mathbf{P}^{(i)}\big)^\top\big(\mathbf{I} - \mathbf{P}^{(i)}\big)\mathbf{G}\Big) = \min_{\mathbf{G}}\sum_{i=1}^{S}\mathrm{tr}\Big(\mathbf{G}^\top\big(\mathbf{I} - \mathbf{P}^{(i)}\big)^2\mathbf{G}\Big)$$

Since $\mathbf{P}^{(i)}$ is idempotent ($\big(\mathbf{P}^{(i)}\big)^2 = \mathbf{P}^{(i)}$) [10, 11], we have:

$$\begin{aligned}\min_{\mathbf{G}} \sum_{i=1}^{S} \operatorname{tr}\Big(\mathbf{G}^\top \big(\mathbf{I} - \mathbf{P}^{(i)}\big)^2 \mathbf{G}\Big) &= \min_{\mathbf{G}} \sum_{i=1}^{S} \operatorname{tr}\Big(\mathbf{G}^\top \Big(\mathbf{I}^2 + \big(\mathbf{P}^{(i)}\big)^2 - 2\mathbf{I}\mathbf{P}^{(i)}\Big)\mathbf{G}\Big) \\ = \min_{\mathbf{G}} \sum_{i=1}^{S} \operatorname{tr}\Big(\mathbf{G}^\top \Big(\mathbf{I}^2 + \mathbf{P}^{(i)} - 2\mathbf{P}^{(i)}\Big)\mathbf{G}\Big) &= \min_{\mathbf{G}} \sum_{i=1}^{S} \operatorname{tr}\Big(\mathbf{G}^\top \big(\mathbf{I} - \mathbf{P}^{(i)}\big)\mathbf{G}\Big) \\ = \min_{\mathbf{G}} \sum_{i=1}^{S} \operatorname{tr}\Big(\mathbf{G}^\top\mathbf{I}\mathbf{G} - \mathbf{G}^\top \mathbf{P}^{(i)}\mathbf{G}\Big) &= \min_{\mathbf{G}} \sum_{i=1}^{S} \Big(\operatorname{tr}\big(\mathbf{I}\big) - \operatorname{tr}\big(\mathbf{G}^\top \mathbf{P}^{(i)}\mathbf{G}\big)\Big) \\ &= \min_{\mathbf{G}} \Big(SV - \operatorname{tr}\Big(\mathbf{G}^\top \big(\sum_{i=1}^{S}\mathbf{P}^{(i)}\big)\mathbf{G}\Big)\Big)\end{aligned}$$

Based on (10), we have:

$$= \min_{\mathbf{G}} \Big(SV - \operatorname{tr}\Big(\mathbf{G}^\top \mathbf{A}\mathbf{G}\Big)\Big) \equiv \max_{\mathbf{G}} \Big(\operatorname{tr}\big(\mathbf{G}^\top \mathbf{A}\mathbf{G}\big)\Big) \quad \blacksquare$$

Lemma 3. *In order to update network parameters as the second step, the derivative of $\mathbf{Z} = \sum_{\ell=1}^{T} \lambda_\ell$, which is the sum of eigenvalues of $\mathbf{A}$, over the mapped neural activities of ℓ-th subject is defined as follows:*

$$\frac{\partial \mathbf{Z}}{\partial f_\ell\big(\mathbf{X}^{(\ell)};\theta^{(\ell)}\big)} = 2\mathbf{R}^{(\ell)}\mathbf{G}^\top - 2\mathbf{R}^{(\ell)}\big(\mathbf{R}^{(\ell)}\big)^\top\Big(f_\ell\big(\mathbf{X}^{(\ell)};\theta^{(\ell)}\big)\Big)^\top \tag{13}$$

Proof. Based on the parameters (regularization, template, mappings, etc.) in DHA, we present a modified version of the Appendix A proof proposed in [11]. For simplicity, we define $f_\ell\big(\mathbf{X}^{(\ell)};\theta^{(\ell)}\big) = \mathbf{Y}^{(\ell)}$ and the covariance matrix for ℓ-th subject as follows:

$$\widetilde{\mathbf{\Phi}}^{(\ell)} = \Big(\big(\mathbf{Y}^{(\ell)}\big)^\top \mathbf{Y}^{(\ell)} + \epsilon\mathbf{I}\Big) \tag{S.7}$$

Further, the inverse of the covariance matrix is denoted for ℓ-th subject as follows:

$$\mathbf{\Phi}^{(\ell)} = \big(\widetilde{\mathbf{\Phi}}^{(\ell)}\big)^{-1} = \Big(\big(\mathbf{Y}^{(\ell)}\big)^\top \mathbf{Y}^{(\ell)} + \epsilon\mathbf{I}\Big)^{-1} \tag{S.8}$$

where this matrix is symmetric ($\mathbf{\Phi}^{(\ell)} = \big(\mathbf{\Phi}^{(\ell)}\big)^\top$). Based on (S.8), the projection (8) can be rewritten as follows:

$$\mathbf{P}^{(\ell)} = \mathbf{Y}^{(\ell)}\mathbf{\Phi}^{(\ell)}\big(\mathbf{Y}^{(\ell)}\big)^\top \tag{S.9}$$

Further, the final mappings (12) can be reformulated as follows:

$$\mathbf{R}^{(\ell)} = \mathbf{\Phi}^{(\ell)}\big(\mathbf{Y}^{(\ell)}\big)^\top \mathbf{G} \tag{S.10}$$

Since $\partial\mathbf{Z}/\partial\mathbf{A} = \mathbf{G}\mathbf{G}^\top$ [14], we reformulate the left side of (13) based on the chain rule as follows:

$$\frac{\partial \mathbf{Z}}{\partial \mathbf{Y}^{(\ell)}_{\alpha\beta}} = \sum_{\mu,\tau=1}^{V_{new}} \frac{\partial \mathbf{Z}}{\partial \mathbf{A}_{\mu\tau}} \frac{\partial \mathbf{A}_{\mu\tau}}{\partial \mathbf{Y}^{(\ell)}_{\alpha\beta}} = \sum_{\mu,\tau=1}^{V_{new}} \big(\mathbf{G}\mathbf{G}^\top\big)_{\tau\mu} \frac{\partial \mathbf{A}_{\mu\tau}}{\partial \mathbf{Y}^{(\ell)}_{\alpha\beta}} \tag{S.11}$$

By considering (10), the product rule, and this key point that $\mathbf{P}^{(\ell)}$ is the only projection related to $\mathbf{Y}^{(\ell)}$, we have:

$$\begin{aligned}\frac{\partial \mathbf{A}_{\mu\tau}}{\partial \mathbf{Y}^{(\ell)}_{\alpha\beta}} = \frac{\partial \mathbf{P}^{(\ell)}_{\mu\tau}}{\partial \mathbf{Y}^{(\ell)}_{\alpha\beta}} &= \delta_{\mu\beta} \sum_{i=1}^{V_{new}} \mathbf{Y}^{(\ell)}_{\tau i}\mathbf{\Phi}^{(\ell)}_{\alpha i} + \delta_{\tau\beta} \sum_{j=1}^{V_{new}} \mathbf{Y}^{(\ell)}_{\mu j}\mathbf{\Phi}^{(\ell)}_{j\alpha} + \sum_{i,j=1}^{V_{new}} \mathbf{Y}^{(\ell)}_{\mu j}\mathbf{Y}^{(\ell)}_{\tau i}\frac{\partial \mathbf{\Phi}^{(\ell)}_{ji}}{\partial \mathbf{Y}^{(\ell)}_{\alpha\beta}} \\ &= \delta_{\mu\beta}\Big(\mathbf{\Phi}^{(\ell)}\big(\mathbf{Y}^{(\ell)}\big)^\top\Big)_{\alpha\tau} + \delta_{\tau\beta}\Big(\mathbf{\Phi}^{(\ell)}\big(\mathbf{Y}^{(\ell)}\big)^\top\Big)_{\alpha\mu} + \sum_{i,j=1}^{V_{new}} \mathbf{Y}^{(\ell)}_{\mu j}\mathbf{Y}^{(\ell)}_{\tau i}\frac{\partial \mathbf{\Phi}^{(\ell)}_{ji}}{\partial \mathbf{Y}^{(\ell)}_{\alpha\beta}}\end{aligned} \tag{S.12}$$

The last term also can be calculated by using the chain rule as follows:

$$\frac{\partial \boldsymbol{\Phi}^{(\ell)}_{ji}}{\partial \mathbf{Y}^{(\ell)}_{\alpha\beta}} = \sum_{m,n=1}^{T} \frac{\partial \boldsymbol{\Phi}^{(\ell)}_{ji}}{\widetilde{\boldsymbol{\Phi}}^{(\ell)}_{mn}} \frac{\widetilde{\boldsymbol{\Phi}}^{(\ell)}_{mn}}{\partial \mathbf{Y}^{(\ell)}_{\alpha\beta}} = -\sum_{m,n=1}^{T} \Big(\boldsymbol{\Phi}^{(\ell)}_{jm} \boldsymbol{\Phi}^{(\ell)}_{ni} \Big(\delta_{\alpha m} \mathbf{Y}^{(\ell)}_{\beta n} + \delta_{\alpha n} \mathbf{Y}^{(\ell)}_{\beta m} \Big) \Big)$$

$$= -\sum_{n=1}^{T} \boldsymbol{\Phi}^{(\ell)}_{j\alpha} \boldsymbol{\Phi}^{(\ell)}_{ni} \mathbf{Y}^{(\ell)}_{\beta n} - \sum_{m=1}^{T} \boldsymbol{\Phi}^{(\ell)}_{jm} \boldsymbol{\Phi}^{(\ell)}_{\alpha i} \mathbf{Y}^{(\ell)}_{\beta m} = -\boldsymbol{\Phi}^{(\ell)}_{j\alpha} \Big(\boldsymbol{\Phi}^{(\ell)} \big(\mathbf{Y}^{(\ell)}\big)^{\top} \Big)_{i\beta} - \boldsymbol{\Phi}^{(\ell)}_{\alpha i} \Big(\boldsymbol{\Phi}^{(\ell)} \big(\mathbf{Y}^{(\ell)}\big)^{\top} \Big)_{j\beta} \tag{S.13}$$

By applying (S.13) to (S.12), we have:

$$\frac{\partial \mathbf{P}^{(\ell)}_{\mu\tau}}{\partial \mathbf{Y}^{(\ell)}_{\alpha\beta}} = \delta_{\mu\beta} \Big(\boldsymbol{\Phi}^{(\ell)} \big(\mathbf{Y}^{(\ell)}\big)^{\top} \Big)_{\alpha\tau} + \delta_{\tau\beta} \Big(\boldsymbol{\Phi}^{(\ell)} \big(\mathbf{Y}^{(\ell)}\big)^{\top} \Big)_{\alpha\mu} - \Big(\boldsymbol{\Phi}^{(\ell)} \big(\mathbf{Y}^{(\ell)}\big)^{\top} \Big)_{\alpha\mu} \Big(\mathbf{Y}^{(\ell)} \boldsymbol{\Phi}^{(\ell)} \big(\mathbf{Y}^{(\ell)}\big)^{\top} \Big)_{\beta\tau}$$
$$- \Big(\boldsymbol{\Phi}^{(\ell)} \big(\mathbf{Y}^{(\ell)}\big)^{\top} \Big)_{\alpha\tau} \Big(\mathbf{Y}^{(\ell)} \boldsymbol{\Phi}^{(\ell)} \big(\mathbf{Y}^{(\ell)}\big)^{\top} \Big)_{\beta\mu} = \delta_{\mu\beta} \Big(\boldsymbol{\Phi}^{(\ell)} \big(\mathbf{Y}^{(\ell)}\big)^{\top} \Big)_{\alpha\tau} + \delta_{\tau\beta} \Big(\boldsymbol{\Phi}^{(\ell)} \big(\mathbf{Y}^{(\ell)}\big)^{\top} \Big)_{\alpha\mu} -$$
$$\Big(\boldsymbol{\Phi}^{(\ell)} \big(\mathbf{Y}^{(\ell)}\big)^{\top} \Big)_{\alpha\mu} \Big(\mathbf{P}^{(\ell)} \Big)_{\beta\tau} - \Big(\boldsymbol{\Phi}^{(\ell)} \big(\mathbf{Y}^{(\ell)}\big)^{\top} \Big)_{\alpha\tau} \Big(\mathbf{P}^{(\ell)} \Big)_{\beta\mu} =$$
$$\Big(\mathbf{I} - \mathbf{P}^{(\ell)} \Big)_{\beta\tau} \Big(\boldsymbol{\Phi}^{(\ell)} \big(\mathbf{Y}^{(\ell)}\big)^{\top} \Big)_{\alpha\mu} + \Big(\mathbf{I} - \mathbf{P}^{(\ell)} \Big)_{\beta\mu} \Big(\boldsymbol{\Phi}^{(\ell)} \big(\mathbf{Y}^{(\ell)}\big)^{\top} \Big)_{\alpha\tau} \tag{S.14}$$

Finally, we can calculate (13) as follows:

$$\frac{\partial \mathbf{Z}}{\partial \mathbf{Y}^{(\ell)}_{\alpha\beta}} = \sum_{\mu,\tau=1}^{T} \Big(\big(\mathbf{G}\mathbf{G}^{\top}\big)_{\mu\tau} \big(\mathbf{I} - \mathbf{P}^{(\ell)}\big)_{\mu\beta} \Big(\boldsymbol{\Phi}^{(\ell)} \big(\mathbf{Y}^{(\ell)}\big)^{\top} \Big)_{\alpha\tau} \Big)$$
$$+ \sum_{\mu,\tau=1}^{T} \Big(\big(\mathbf{G}\mathbf{G}^{\top}\big)_{\mu\tau} \big(\mathbf{I} - \mathbf{P}^{(\ell)}\big)_{\tau\beta} \Big(\boldsymbol{\Phi}^{(\ell)} \big(\mathbf{Y}^{(\ell)}\big)^{\top} \Big)_{\alpha\mu} \Big) = 2 \Big(\boldsymbol{\Phi}^{(\ell)} \big(\mathbf{Y}^{(\ell)}\big)^{\top} \mathbf{G}\mathbf{G}^{\top} \big(\mathbf{I} - \mathbf{P}^{(\ell)}\big) \Big)_{\alpha\beta} \tag{S.15}$$

Consequently, we have:

$$\frac{\partial \mathbf{Z}}{\partial \mathbf{Y}^{(\ell)}} = 2\boldsymbol{\Phi}^{(\ell)} \big(\mathbf{Y}^{(\ell)}\big)^{\top} \mathbf{G}\mathbf{G}^{\top} \big(\mathbf{I} - \mathbf{P}^{(\ell)}\big)$$

By considering (S.10), we have:

$$\frac{\partial \mathbf{Z}}{\partial \mathbf{Y}^{(\ell)}} = 2\mathbf{R}^{(\ell)} \mathbf{G}^{\top} \big(\mathbf{I} - \mathbf{P}^{(\ell)}\big) = 2\mathbf{R}^{(\ell)} \mathbf{G}^{\top} - 2\mathbf{R}^{(\ell)} \mathbf{G}^{\top} \mathbf{P}^{(\ell)}$$

By applying (S.9), we have:

$$= 2\mathbf{R}^{(\ell)} \mathbf{G}^{\top} - 2\mathbf{R}^{(\ell)} \mathbf{G}^{\top} \mathbf{Y}^{(\ell)} \boldsymbol{\Phi}^{(\ell)} \big(\mathbf{Y}^{(\ell)}\big)^{\top}$$

Since $\boldsymbol{\Phi}^{(\ell)}$ is symmetric:

$$= 2\mathbf{R}^{(\ell)} \mathbf{G}^{\top} - 2\mathbf{R}^{(\ell)} \mathbf{G}^{\top} \mathbf{Y}^{(\ell)} \big(\boldsymbol{\Phi}^{(\ell)}\big)^{\top} \big(\mathbf{Y}^{(\ell)}\big)^{\top}$$

and again using (S.10), we finally have:

$$\frac{\partial \mathbf{Z}}{\partial \mathbf{Y}^{(\ell)}} = 2\mathbf{R}^{(\ell)} \mathbf{G}^{\top} - 2\mathbf{R}^{(\ell)} \big(\mathbf{R}^{(\ell)}\big)^{\top} \big(\mathbf{Y}^{(\ell)}\big)^{\top}$$
$$\Longrightarrow \quad \frac{\partial \mathbf{Z}}{\partial f_{\ell}\big(\mathbf{X}^{(\ell)};\theta^{(\ell)}\big)} = 2\mathbf{R}^{(\ell)} \mathbf{G}^{\top} - 2\mathbf{R}^{(\ell)} \big(\mathbf{R}^{(\ell)}\big)^{\top} \Big(f_{\ell}\big(\mathbf{X}^{(\ell)};\theta^{(\ell)}\big) \Big)^{\top} \blacksquare$$

S.7 Experiment Schemes

This section presents experiment schemes. As the first step, each dataset is separately preprocessed by FSL 5.0.9 (`https://fsl.fmrib.ox.ac.uk`), i.e. slice timing, anatomical alignment, normalization, smoothing. Here, each session is preprocessed independently. Further, this paper separately applies the anatomical alignment to each session (i.e. whole-brain image) for visualizing the neural activities in the MNI space and calibrating the Regions of Interests (ROIs). Then, ROIs are denoted by employing the main reference of each dataset. As the next step, we have partitioned preprocessed data to the training set and testing set for each iteration by using one-subject-out cross-validation. Next, the training set is applied to different HA methods. Same as the previous studies [2-7], the HA parameters are selected based on the lowest alignment error among the training data (the error is estimated by using line 10 in Algorithm 1) for both the proposed method and other methods (RHA, KHA, SRM, and CAE). Here, the structure of the deep network in the proposed method is defined based on the error of functional alignment (line 10 in Algorithm 1), i.e. activation function, regularization term, learning rate, and numbers of layers and units. As the last step in the training phase, the aligned data are employed for learning a classification (cognitive) model (without changing the alignment parameters). In the testing phase, we just applied the test data to the generated model without changing the estimated parameters of the training phase. Indeed, we try to demonstrate the effect of functional alignment on the performance of the final classification model. In other words, we consider whole of the classification procedure is fixed except the alignment section, i.e. the same preprocessed dataset for each iteration, the same learning algorithm, the same parameters for learning, etc. We have to emphasize that the same preprocessed datasets are applied to all HA methods in each iteration of the cross-validation.

References

[1] Haxby, J.V. & Connolly, A.C. & Guntupalli, J.S. (2014) Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience.* 37:435–456,

[2] Xu, H. & Lorbert, A. & Ramadge, P.J. & Guntupalli, J.S. & Haxby, J.V. (2012) Regularized hyperalignment of multi-set fMRI data. *IEEE Statistical Signal Processing Workshop (SSP).* pp. 229–232, Aug/5–8, USA.

[3] Lorbert, A. & Ramadge, P.J. (2012) Kernel hyperalignment. *25th Advances in Neural Information Processing Systems (NIPS).* pp. 1790–179. Dec/3–8, Harveys.

[4] Chen, P.H. & Guntupalli, J.S. & Haxby, J.V. & Ramadge, P.J. (2014) Joint SVD-Hyperalignment for multi-subject FMRI data alignment. *24th IEEE International Workshop on Machine Learning for Signal Processing (MLSP).* pp. 1–6, Sep/21–24, France.

[5] Chen, P.H. & Chen, J. & Yeshurun, Y. & Hasson, U. & Haxby, J.V. & Ramadge, P.J. (2015) A reduced-dimension fMRI shared response model. *28th Advances in Neural Information Processing Systems (NIPS).* pp. 460–468, Dec/7–12, Canada.

[6] Chen, P.H. & Zhu, X. & Zhang, H. & Turek, J.S. & Chen, J. & Willke, T.L. & Hasson, U. & Ramadge, P.J. (2016) A convolutional autoencoder for multi-subject fMRI data aggregation. *29th Workshop of Representation Learning in Artificial and Biological Neural Networks.* NIPS, Dec/5–10, Barcelona.

[7] Yousefnezhad, M. & Zhang D. (2017) Local Discriminant Hyperalignment for multi-subject fMRI data alignment. *34th AAAI Conference on Artificial Intelligence.* pp. 59–61, Feb/4–9, San Francisco, USA.

[8] Guntupalli, J.S. & Hanke, M. & Halchenko, Y.O. & Connolly, A.C. & Ramadge, P.J. & Haxby, J.V. (2016) A model of representational spaces in human cortex. *Cerebral Cortex.* Oxford University Press.

[9] Arora, R. & Livescu, K. (2012) Kernel CCA for multi-view learning of acoustic features using articulatory measurements. *Machine Learning in Speech and Language Processing (MLSLP).* pp. 34–37, Sep/14, USA.

[10] Rastogi, P. & Van D.B. & Arora, R. (2015) Multiview LSA: Representation Learning via Generalized CCA. *14th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL).* pp. 556–566, May/31 to Jun/5, Denver, USA.

[11] Benton, A. & Khayrallah, H. & Gujral, B. & Reisinger, D. & Zhang, S. & Arora, R. (2017) Deep Generalized Canonical Correlation Analysis. *5th International Conference on Learning Representations (ICLR).* Apr/24–26.

[12] Wang, W. & Arora, R. & Livescu, K. & Srebro, N. Stochastic optimization for deep CCA via nonlinear orthogonal iterations. *53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton).* pp. 688–695, Oct/3–6, Urbana-Champaign, USA.

[13] Brand, M. (2002) Incremental singular value decomposition of uncertain data with missing values. *7th European Conference on Computer Vision (ECCV).* pp. 707–720, May/28–31, Copenhagen, Denmark.

[14] Petersen, K.B & Pedersen, M.S. (2012) The Matrix Cookbook. Version: November 15, *Technical University of Denmark.* Available: http://www.math.uwaterloo.ca/%7Ehwolkowi/matrixcookbook.pdf

Table S1: Variables or Functions

Variable or Function	Description
$\mathbb{R}$	The set of real numbers.
$\mathbf{I}$	The identity matrix.
$i, j, \ell, m, n, \alpha, \beta, \mu, \tau$	The indices.
K	The number of stimulus categories.
S	The number of Subjects.
T	The number of time points in unites of TRs (Time of Repetition).
V	The number of voxels in the original representation space.
V_{new}	The number of features after applying the Artificial Neural Network (ANN).
C	The number of layers in ANN.
$U^{(m)}$	The number of units in m-th intermediate layers of ANN.
ϵ	The regularized constant.
$\mathbf{X}^{(\ell)} = \left\{ x_{mn}^{(\ell)} \right\} \in \mathbb{R}^{T \times V}$	The original neural activities for ℓ-th subject.
$\mathbf{R}^{(\ell)}$	HA or DHA mapping for ℓ-th subject.
$\theta^{(\ell)}$=$\left\{\mathbf{W}_m^{(\ell)}, \mathbf{b}_m^{(\ell)}, m\text{=}1\text{:}C\right\}$	The parameters of ANN for ℓ-th subject.
$\mathbf{h}_m^{(\ell)} = \mathrm{g}(\dots)$	The result of activation function for ℓ-th subject and m-th layer.
$f_\ell\big(\mathbf{X}^{(\ell)};\theta^{(\ell)}\big) = \mathbf{Y}^{(\ell)}$	The cognitive features for ℓ-th subject after applying ANN.
$\mathbf{G} \in \mathbb{R}^{T \times V_{new}}$	The DHA template.
$\mathbf{Q} \in \mathbb{R}^{V_{new} \times V_{new}}$	An orthogonal matrix for mapping two different DHA template to each other.
$f_\ell\big(\mathbf{X}^{(\ell)};\theta^{(\ell)}\big) \overset{SVD}{=} \mathbf{\Omega}^{(\ell)}\mathbf{\Sigma}^{(\ell)}\big(\mathbf{\Psi}^{(\ell)}\big)^\top$	The SVD decomposition of mapped cognitive features for ℓ-th subject.
$\widetilde{\mathbf{\Phi}}^{(\ell)}$	The covariance matrix for ℓ-th subject (just used in this document)
$\mathbf{\Phi}^{(\ell)}$	The inverse of covariance matrix for ℓ-th subject (just used in this document)
$\mathbf{P}^{(\ell)}$	The projection of cognitive features for ℓ-th subject.
$\mathbf{D}^{(\ell)}$	The diagonal matrix used for decomposition of $\mathbf{P}^{(\ell)}$ for ℓ-th subject.
$\mathbf{A} = \widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\top$	Sum of all projections and its Cholesky decomposition.
$\Lambda = \big\{\lambda_1 \dots \lambda_T\big\}$	Eigenvalues of $\mathbf{A}$.
$\mathbf{Z} = \sum_{\ell=1}^{T} \lambda_\ell$	Sum of eigenvalues of $\mathbf{A}$.
η	Learning rate.
$\nabla\theta^{(\ell)}$	The gradient of ANN parameters for ℓ-th subject.
γ_m	Error of m-th iteration.
k_1, k_3, λ, ρ	Convolutional Autoencoder (CAE) parameters.
$tr()$	The trace function.
$backprop()$	The back-propagation function.
$\mathrm{g}(\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{x})}$	Sigmoid function.
$\mathrm{g}(\mathbf{x}) = \tanh(\mathbf{x}) = \frac{\exp(\mathbf{x})-\exp(-\mathbf{x})}{\exp(\mathbf{x})+\exp(-\mathbf{x})}$	Hyperbolic function.
$\mathrm{g}(\mathbf{x}) = \ln(1+\exp(\mathbf{x}))$	Rectified Linear Unit (ReLU).

Table S2: The datasets.

Title	ID	S	K	T	V	X	Y	Z	Scanner	TR	TE
Mixed-gambles task	DS005	48	2	240	450	53	63	52	S 3T	2	30
Visual Object Recognition	DS105	71	8	121	1963	79	95	79	G 3T	2.5	30
Word and Object Processing	DS107	98	4	164	932	53	63	52	S 3T	2	28
Auditory and Visual Oddball	DS116	102	2	170	2532	53	63	40	P 3T	2	25
Multi-subject, multi-modal	DS117	171	2	210	524	64	61	33	S 3T	2	30
Forrest Gump	DS113	20	10	451	2400	160	160	36	S 7T	2.3	22
Raiders of the Lost Ark	N/A	10	7	924	980	78	78	54	S 3T	3	30

S is the number of subject; K denotes the number of stimulus categories; T is the number of scans in unites of TRs (Time of Repetition); V denotes the number of voxels in ROI; X, Y, Z are the size of 3D images; Scanners include S=Siemens, G = General Electric, and P = Philips in 3 Tesla or 7 Tesla; TR is Time of Repetition in millisecond; TE denotes Echo Time in second; Please see *openfmri.org* for more information.

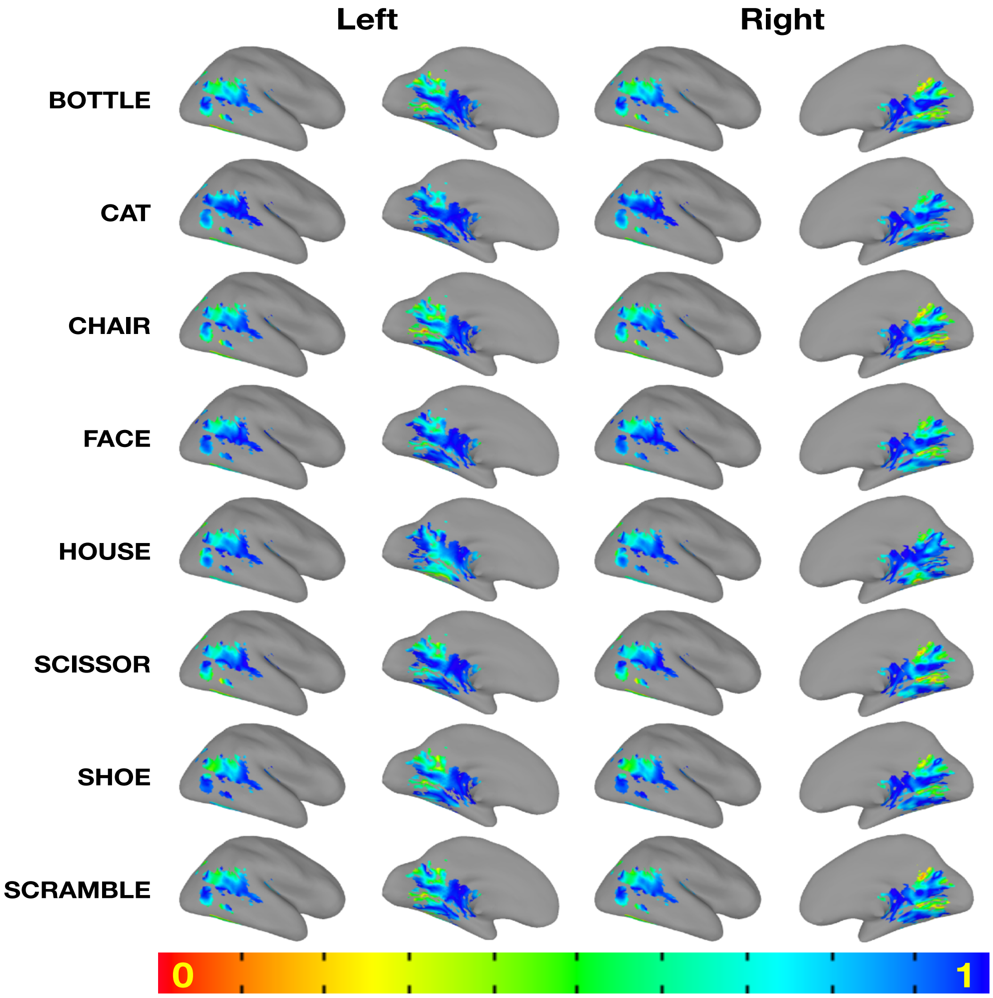

Figure S1: Visualizing average of active regions in common space across category of visual stimuli by using DS105 dataset

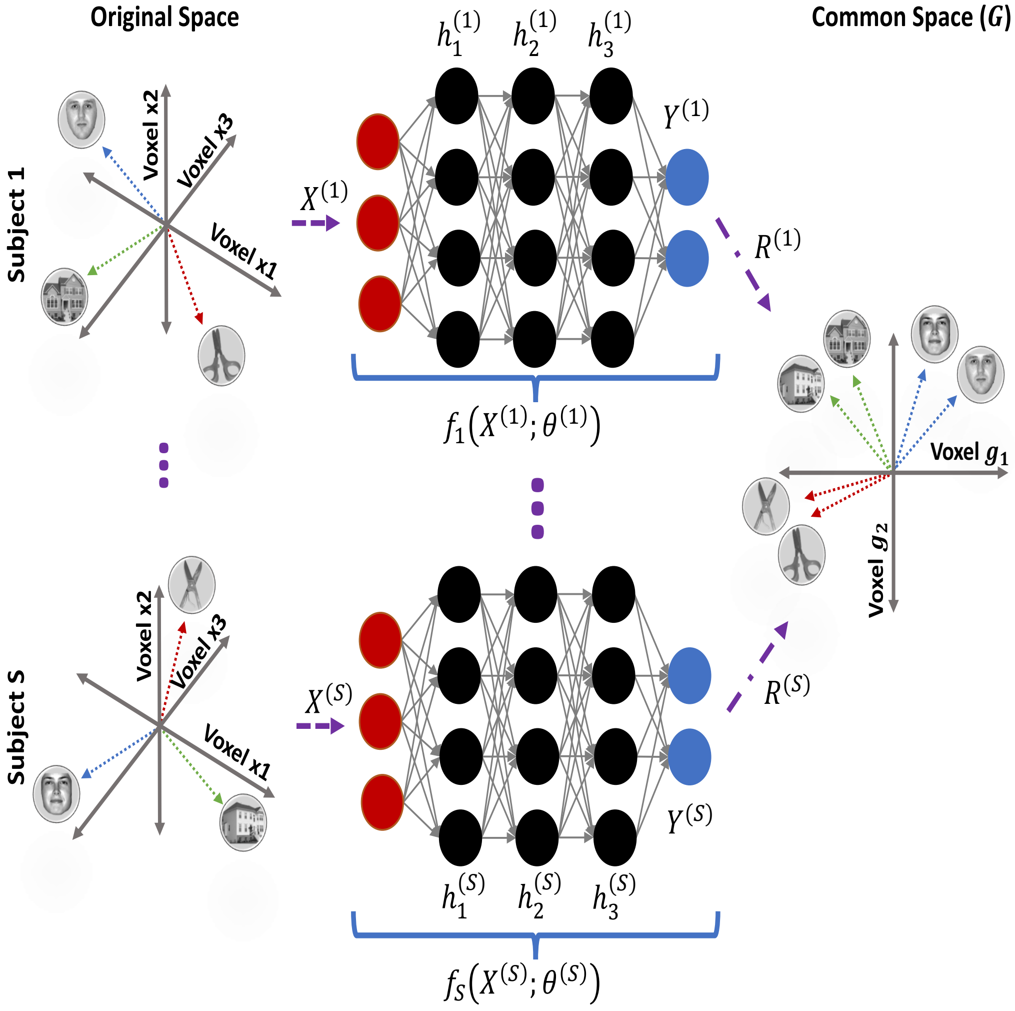

Figure S2: Graphical abstract of Deep Hyperalignment (DHA)