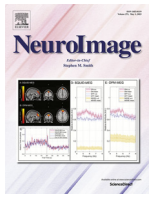
NeuroImage 271 (2023) 120007

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage



Inter-individual deep image reconstruction via hierarchical neural code conversion

Jun Kai Ho [a,¶,*], Tomoyasu Horikawa [b,¶], Kei Majima [a], Fan Cheng [a,b], Yukiyasu Kamitani [a,b,*]

[a] Graduate School of Informatics, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
[b] Department of Neuroinformatics, ATR Computational Neuroscience Laboratories, Hikaridai, Seika, Soraku, Kyoto, 619-0288, Japan

ARTICLE INFO

Keywords:
Visual image reconstruction
Decoding
Visual hierarchy
Functional alignment
fMRI

ABSTRACT

The sensory cortex is characterized by general organizational principles such as topography and hierarchy. However, measured brain activity given identical input exhibits substantially different patterns across individuals. Although anatomical and functional alignment methods have been proposed in functional magnetic resonance imaging (fMRI) studies, it remains unclear whether and how hierarchical and fine-grained representations can be converted between individuals while preserving the encoded perceptual content. In this study, we trained a method of functional alignment called neural code converter that predicts a target subject's brain activity pattern from a source subject given the same stimulus, and analyzed the converted patterns by decoding hierarchical visual features and reconstructing perceived images. The converters were trained on fMRI responses to identical sets of natural images presented to pairs of individuals, using the voxels on the visual cortex that covers from V1 through the ventral object areas without explicit labels of the visual areas. We decoded the converted brain activity patterns into the hierarchical visual features of a deep neural network using decoders pre-trained on the target subject and then reconstructed images via the decoded features. Without explicit information about the visual cortical hierarchy, the converters automatically learned the correspondence between visual areas of the same levels. Deep neural network feature decoding at each layer showed higher decoding accuracies from corresponding levels of visual areas, indicating that hierarchical representations were preserved after conversion. The visual images were reconstructed with recognizable silhouettes of objects even with relatively small numbers of data for converter training. The decoders trained on pooled data from multiple individuals through conversions led to a slight improvement over those trained on a single individual. These results demonstrate that the hierarchical and fine-grained representation can be converted by functional alignment, while preserving sufficient visual information to enable inter-individual visual image reconstruction.

1. Introduction

Sensory information is generally thought to be processed through a hierarchical pathway that detects topographically organized simple local features in the early stages and then progressively complex global features in the later stages, leading to holistic perception. In the ventral visual pathway, a stimulus is initially processed in the striate cortex (V1) to extract simple features, such as edges (Hubel and Wiesel, 1962), and is then further processed in the extrastriate cortices (V2–V4) and higher visual cortex (HVC) to detect more complex visual features, such as shape and face attributes, eventually identifying objects and scenes (Mishkin and Ungerleider, 1982). Whereas general principles such as topography and hierarchy appear to govern the organization of the visual cortex (VC), individual brains differ substantially in both macroscopic anatomy and the fine-grained organization of feature representations. These individual differences make it challenging to relate visual cortical activity and perceptual content by simple mapping rules common across individuals.

Recent advances in deep neural networks (DNNs) have enabled detailed analyses of hierarchical feature representations across different visual cortical areas (Yamins et al., 2014; Güçlü and van Gerven, 2015, 2017; Horikawa and Kamitani, 2017). Previous encoding and decoding studies have shown that DNNs pre-trained on natural images exhibit a correspondence between visual areas and DNN layers. These findings indicate that the visual cortex processes increasingly complex visual features along the ventral neural pathway, similar to how DNNs pro-

* Corresponding authors.
E-mail addresses: junkai125@gmail.com (J.K. Ho), kamitani@i.kyoto-u.ac.jp (Y. Kamitani).
¶ These authors contributed equally to this work.

https://doi.org/10.1016/j.neuroimage.2023.120007.
Received 29 December 2022; Received in revised form 26 February 2023; Accepted 7 March 2023
Available online 11 March 2023.

cess image features. Additionally, the use of DNN-based reconstruction algorithms has led to successful reconstruction of perceptual content encoded in brain responses as images (Shen et al., 2019a and 2019b). The deep image reconstruction (Shen et al., 2019b) first predicts the DNN features of an image from the brain activity given that image as a stimulus, and then an initial image is iteratively optimized such that its DNN features become close to the predicted DNN features. The use of DNN feature decoding enables comprehensive evaluations of hierarchical visual representations, and visual image reconstruction allows a holistic evaluation of how accurately perceptual content is encoded in the brain activity patterns. However, these predictive models require training data derived from hours of experiments that measure the brain responses to hundreds or thousands of images. Furthermore, a model trained on one subject does not generalize to other subjects because of individual differences in macroscopic brain structure and fine-grained neural representations.

Methods for the anatomical and functional alignment of different individuals' brains have been developed in decades of functional magnetic resonance imaging (fMRI) studies to account for individual differences. Human brain anatomy differs across individuals in terms of shape, size, and local anatomical landmarks. Functional brain area parcellation that clusters voxels/vertices with similar properties produces similar brain areas on the individual level but still exhibit distinct topological features (Blumensath et al., 2013; Laumann et al., 2015). The visual areas delineated by the retinotopy principle (Engel et al., 1994; Sereno et al., 1995) are often similar but not the same across individuals. Anatomical alignment can mitigate anatomical differences by matching anatomical features between brains (Fischl et al., 2008; van Essen, 2004, 2005), but it still cannot perfectly align the functional topography across individuals (Watson et al., 1993). Functional alignment adopts an anatomy-free approach by learning statistical relationships between subjects' brain activity patterns (Haxby et al., 2011; Yamada et al., 2011, 2015; Chen et al., 2015; Bilenko and Gallant, 2016; Guntupalli et al., 2016). Methodologies of functional alignment include pairwise alignments between two subjects, such as a neural code converter (Yamada et al., 2015), and template-based alignments, in which a shared template among subjects is constructed, such as hyperalignment (Haxby et al., 2011). Functional alignment methods have revealed common neural representations across individuals that are concealed under substantial individual variations in brain responses. However, previous investigations have often focused on a few specific features, such as object categories, image contrast, retinotopy, and semantics (Haxby et al., 2011; Yamada et al., 2015; Bilenko and Gallant, 2016; Van Uden et al., 2018), leaving it unclear whether distinct levels of fine-grained neural representations of hierarchical visual features can be converted across individuals such that an individual's perceptual experience can be reconstructed using other individuals models. Furthermore, previous studies have separately performed alignments on different brain areas using rough anatomical correspondences across individuals (Güçlü and van Gerven, 2015). It remains unknown whether data-driven methods trained on fMRI data can automatically detect hierarchical representations of distinct levels of visual features common across individuals.

Here, we aim to investigate the feasibility of converting fine-grained neural representations of hierarchical visual features between individuals while preserving the encoded perceptual content. To achieve this, we utilized a functional alignment method (neural code converter; Yamada et al., 2015) to convert brain activity, and then used the decoding of hierarchical DNN features (Horikawa and Kamitani, 2017) and reconstruction of perceived images (deep image reconstruction; Shen et al., 2019b) to analyze the converted brain activity. We also adopted other methods of pairwise alignment, including Procrustes transformation (Schönemann, 1966), optimal transport (Bazeille et al., 2019), and a template-based pairwise alignment via hyperalignment (Haxby et al., 2011). Our aim is not to exhaustively evaluate all available methods of pairwise alignment, but to show the robustness of the results across several methods. We restricted the work within the methods of pairwise alignment. We did not discuss shared templates due to difficulties in interpreting the correspondence of visual subareas between subjects in a shared template, and the question of how best a template can be estimated is distinct from the alignment methods (Bazeille et al., 2021). Instead, we used the template-based alignment only to construct a pairwise transformation via the template, which we call template-based pairwise alignment. Our approach involved constructing machine learning-based models that convert an fMRI pattern in the VC of one subject (the source) to the individual voxel responses of another subject (the target) given identical sequences of natural image stimuli (Fig. 1A). We also trained DNN feature decoders with measured fMRI responses of the target subject (Fig. 1A). Then, given the source subject's brain responses to novel stimuli, the converter transforms the brain activity into the target brain space (Fig. 1B). The converted brain activities are decoded by the DNN feature decoders pre-trained on the target subject, and then the decoded features are used in a reconstruction algorithm to create images (Fig. 1B).

In this study, we first show that machine learning-based converter models automatically learn the hierarchical correspondence of visual subareas between subjects, even without explicit information about cortical hierarchy during training. DNN feature decoding from the converted fMRI responses at each hierarchical DNN level shows greater accuracy in the corresponding levels of visual subareas than in other levels, indicating that fine-grained hierarchical feature representations are preserved. Visual image reconstruction using the decoded DNN features from the converted fMRI responses produces faithful reconstructions of the viewed images, even with small numbers of data for the converter training. We also demonstrate that the information about cortical hierarchy used in the training does not improve the performance of converters when given sufficient training data. Finally, by pooling data from multiple subjects through neural code conversions, we show that DNN feature decoders trained on the pooled data achieve a slight improvement in the inter-individual visual image reconstruction. These results demonstrate that the hierarchical correspondence can be automatically detected and the fine-grained representations of visual features can be preserved across individuals by the neural code converters, providing an efficient way to create visual image reconstructions for novel individuals.

2. Results

2.1. fMRI data

We analyzed fMRI data of the five subjects in the previously published studies (Shen et al., 2019b; Horikawa and Kamitani, 2022). For two of the five subjects, we collected additional data for this study (see Materials and Methods: "fMRI datasets"). The dataset consisted of fMRI data measured when subjects viewed images, each presented in an 8-s block (four fMRI volumes). To acquire training data, the presentation of 1200 natural images was repeated five times. For test data, the presentation of 50 natural images was repeated 24 times in the test natural image session, and the presentation of 40 artificial images (simple geometric shapes) was repeated 20 times in the test artificial image session. Artificial images were introduced to assess how well models trained on natural images generalize to a different type of images. The fMRI data were averaged in each 8-s stimulus block (four fMRI volumes shifted by 4 s to account for hemodynamic delays). Thus, 6,000 ($5 \times 1{,}200$) training samples, 1,200 (24×50) test samples with natural images, and 800 (20×40) test samples with artificial images were available. In decoding and reconstruction analyses, test samples were further averaged across repetitions (blocks) for each image. Notably, the fMRI data collection for 6,000 training samples required approximately 800 min of scan sessions per subject, conducted on different days. Although some of the training data and the test data were collected at different times, separated by more than several months or even a year, the trained model generalized well across the datasets, as demonstrated in Shen et al. (2019b).

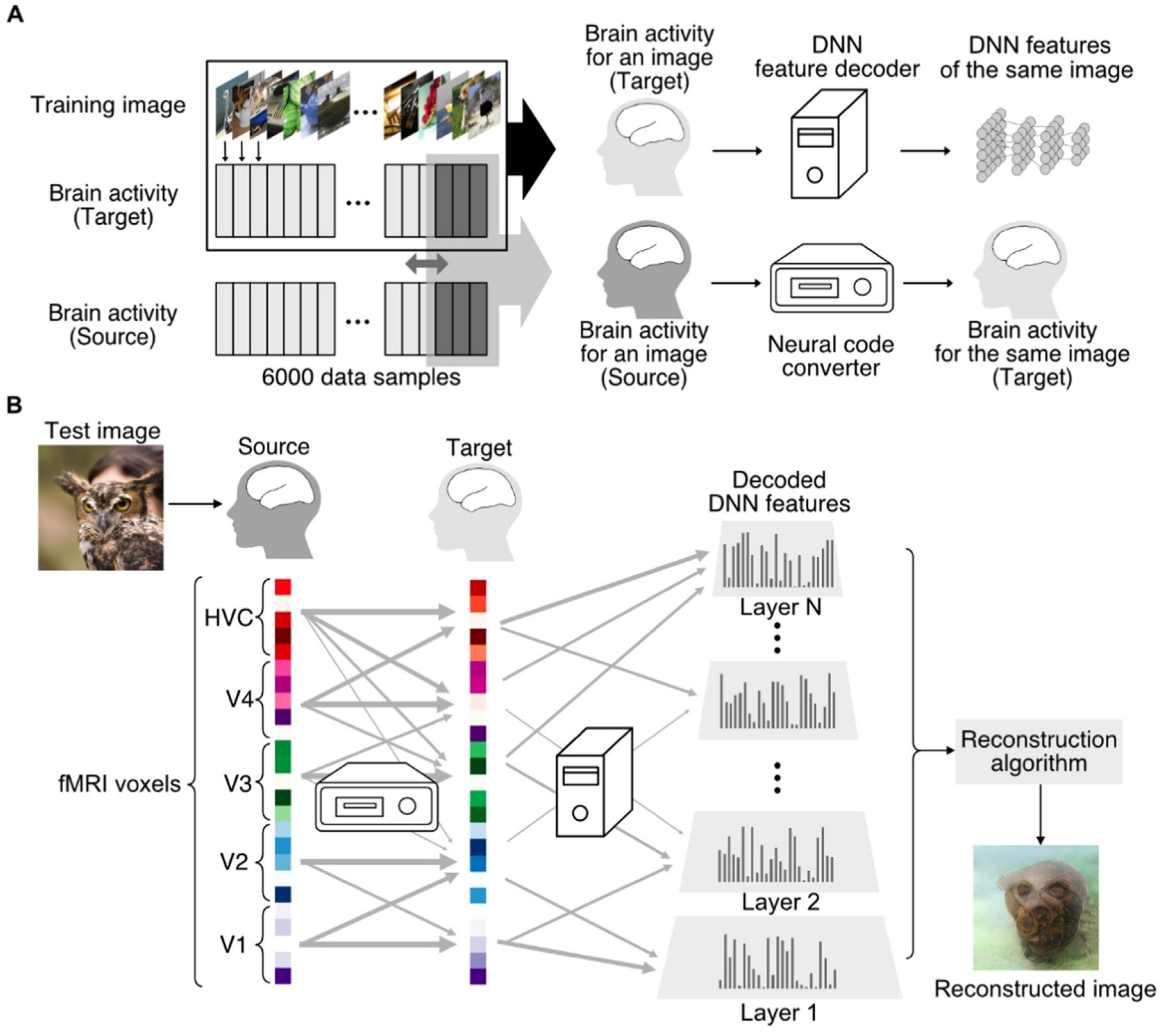


Fig. 1. Inter-individual deep image reconstruction.
(A) Training of the DNN feature decoders and a neural code converter. DNN feature decoding models were trained on the 6,000 samples of measured fMRI activities of the target subject and the corresponding DNN features. A converter model was trained on a subset of 6,000 samples of fMRI data responses to an identical stimulus sequence from both the source and target subject. No explicit information about cortical hierarchy is provided at the training stage.
(B) Inter-individual DNN feature decoding and visual image reconstruction. The converter model converts the source subject's stimulus-induced fMRI pattern into the target subject's brain space. The converted fMRI pattern is then decoded (or translated) into a DNN feature pattern using the feature decoders. Finally, the decoded features are fed into the reconstruction algorithm to reconstruct the stimulus image perceived by the source subject.

2.2. Neural code conversion

We first examined how the results of neural code conversion reflect cortical hierarchy using several evaluation methods. We constructed a neural code converter model between each pair of subjects, with one serving as the target subject and the other as the source subject, resulting in 20 individual pairs. A converter model comprises a set of regularized linear regression models (ridge regression), each trained to predict the activity of each voxel of the target subject's brain from the source subject's brain activity pattern in a broad region of interest (ROI) that covered the lower to higher visual cortex termed VC (see Materials and Methods: "Methods of functional alignment"). In the current study, neural code converter models were trained using a varying number of training samples (300, 600, 900, 1,200, 2,400, 3,600, 4,800, or 6,000 samples). Unless otherwise noted, we show the results obtained using 2,400 training samples (two repetitions of 1,200 images) as a representative case.

VC consists of V1–V4 and ventral object-responsive areas (see Materials and Methods: "Regions of interest"). We defined the continuous region covering the lateral occipital complex (LOC), fusiform face area (FFA), and parahippocampal place area (PPA) as the higher visual cortex (HVC). In the analyses of this section, all VC voxels were used as inputs to the converter without additional voxel selection (see Materials and Methods: "Neural code converter"). Conversion results were evaluated within individual ROIs (subareas) in the target subject's brain space.

Although we mainly present the results from the neural code converter analysis (Yamada et al., 2015), we also conducted similar pairwise alignment analyses using Procrustes transformation (Schönemann, 1966), optimal transport (Bazeille et al., 2019), and template-based pairwise alignment via hyperalignment (Fig. S1; see Materials and Methods: "Methods of functional alignment") to confirm the robustness of the results across different functional alignment methods (for an evaluation of different methods, see Bazeille et al., 2021). Compared to other methods, the neural code converter is simple and less computationally expensive.

We evaluated the models using two methods: (a) pattern correlation, which calculates the spatial Pearson correlation coefficient between the converted and measured voxel patterns for a test image, and (b) profile correlation, which is the Pearson correlation coefficient between the sequences of converted and measured individual voxel responses to

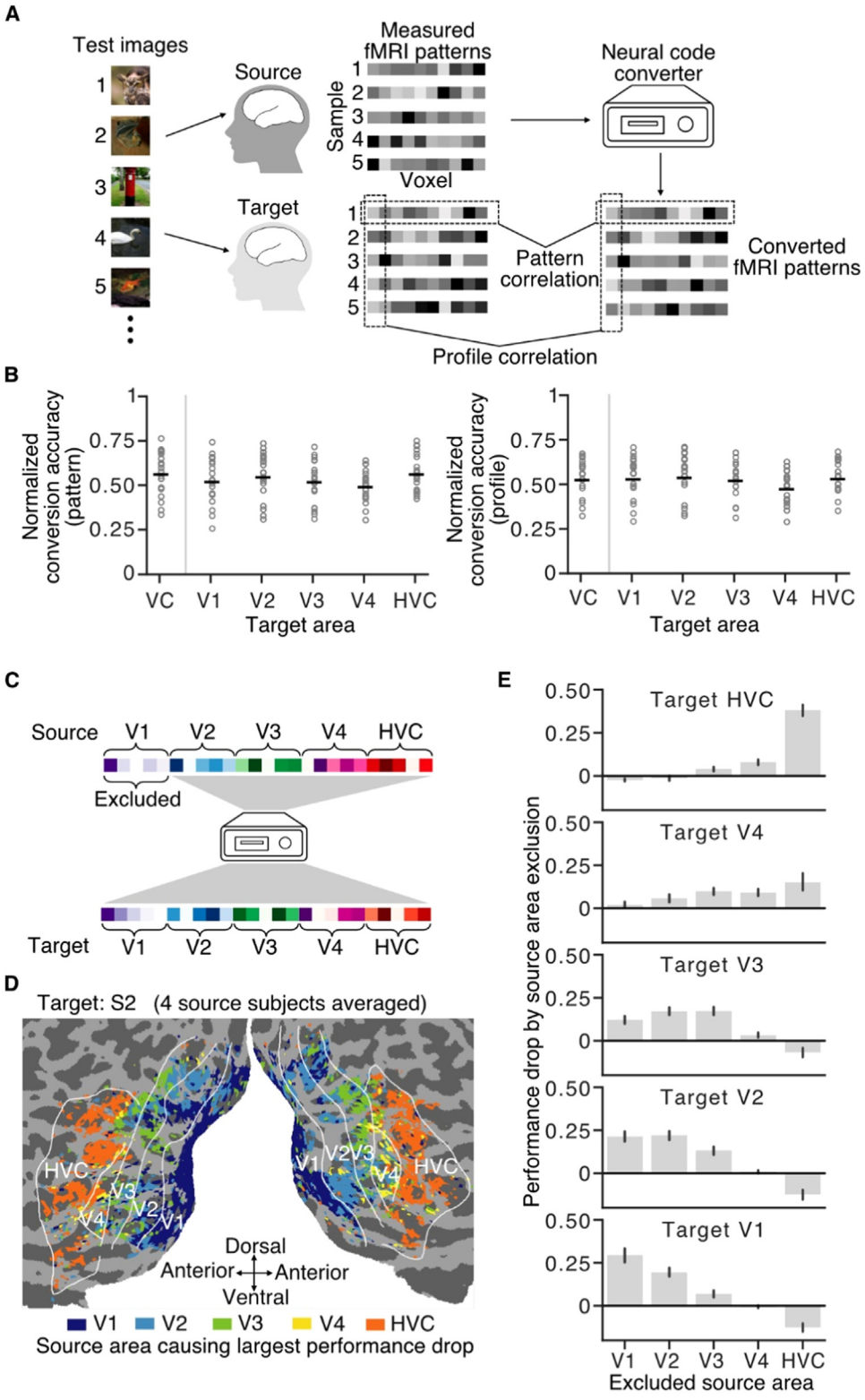


Fig. 2. Performance of neural code converters and cortical hierarchical correspondence.
(A) Evaluations of neural code converters. Two evaluations were performed by computing the Pearson correlation coefficients: pattern and profile correlations.
(B) Conversion accuracy. Distributions of the normalized pattern or profile correlation coefficients of 20 individual pairs are shown for the VC and visual subareas. Each horizontal black dash indicates the mean value; each circle represents the correlation coefficients of an individual pair.
(C) Ablation analysis on neural code converters. The analysis was performed by excluding one source visual area from the prediction of target voxel activities.
(D) Cortical map of the effects of source area exclusion. The cortical map is shown for one target subject (Subject 2). Each voxel on the target brain is colored by the index of the excluded visual area that caused the largest performance drop when testing with the natural image test dataset (performance drops were averaged across four source subjects for a single target subject; see Fig. S3 for other target subjects). Only voxels that generate reliable responses with noise ceilings above a threshold are shown (see Materials and Methods: "Noise ceiling estimation").
(E) Mean performance drop caused by source area exclusion. Each bar represents the mean performance drop averaged across voxels in a target area when a source area was excluded during prediction (averaged over 20 individual pairs; error bars, 95% confidence interval [C.I.] from 20 individual pairs).

the 50 natural test images (Fig. 2A). The pattern correlation for an image was defined as the mean of 24 samples (converted) × 24 samples (measured) = 576 correlation coefficients. The profile correlation for each voxel was defined as the mean of 24 repetitions (converted) × 24 repetitions (measured) = 576 correlation coefficients. The obtained correlation coefficients were normalized by their noise ceilings to account for the noise in fMRI brain responses over repeated measurements with the same stimulus (Hsu et al., 2004; Lescroart and Gallant, 2019; see Materials and Methods: "Noise ceiling estimation"). To summarize the results, we further averaged the correlation coefficients across images and voxels for the pattern and the profile correlations, respectively, in each individual pair and each ROI.

Although our primary analyses focused on the samples within each conversion pair (Smith et al. 2018), group results, where each data point

represents an individual pair, are shown in main figures for illustrative summary purposes. The normalized pattern correlation coefficients in individual pairs are shown for different ROIs of the target subject in Fig. S2A (left), and their distributions across all conversion pairs are shown in Fig. 2B (left). The mean normalized pattern correlation for the whole VC was 0.56 ± 0.06 (mean with 95% C.I.) over 20 individual pairs, with the visual subareas showing comparable distributions. Examples of converted brain activity patterns are shown together with the targets brain activity patterns in Fig. S2B. The mean normalized profile correlation for VC was 0.53 ± 0.05 over 20 individual pairs (Fig. S2A right for individual pairs; Fig. 2B right for group results). The subareas also yielded distributions similar to those of the VC. The conversion accuracy was modest in both pattern and profile correlations across all visual subareas but comparable to the findings in the previous study (Yamada et al., 2015). Other methods of functional alignment showed similar conversion accuracies, with optimal transport showing higher accuracies (Fig. S3).

To see how the source visual areas influenced the conversion accuracy for each voxel in each target visual area, we excluded one of the source visual subareas (V1, V2, V3, V4, or HVC) from the input to the trained converter model (Fig. 2C). We evaluated the drop in performance (normalized profile correlation difference) relative to the performance when all source visual subareas were included (i.e., the whole VC). This ablation analysis revealed that the effects of the source area exclusions varied with the area in the target brain. The largest drop in performance of a target voxel was often caused by the exclusion of the corresponding source area (Fig. 2D, target S2; see Fig. S4A for the other subjects). On average, the peak of performance drop shifted from lower to higher excluded source areas along the hierarchy of the target areas (Fig. S4B for results of some individual pairs; Fig. 2E for group results). The results indicate that the machine learning-based neural code converter models automatically detect a “low-to-high” hierarchical correspondence between source and target visual areas even without explicit anatomical information.

2.3. DNN feature decoding

We next used DNN feature decoding analysis (Horikawa and Kamitani, 2017) to examine whether fine-grained representations of visual features were preserved in the converted fMRI activity patterns. Feature decoders had been trained to predict the DNN feature values of the stimuli using 6,000 training samples of a target subject’s fMRI activity patterns in both the whole VC and individual visual subareas. The feature decoders were applied to the converted brain activities to predict the DNN features of the test images (“Across-functional” condition; see Materials and Methods: “DNN feature decoding analysis”). Following the original paper (Shen et al., 2019b), we used the average fMRI data over the repetitions for each test image as the input to feature decoders. The decoding accuracy of each DNN unit was calculated as the Pearson correlation coefficient between the sequences of the decoded and true feature values for the test images. We further took the mean decoding accuracy over all DNN units in each layer.

To provide a comparison, we performed the same analysis with anatomically aligned brain activity. The source subject’s fMRI images were aligned to the target’s anatomical template and then used for DNN feature decoding (“Across-anatomical”; see Materials and Methods: “Anatomical alignment”). We also compared the results to those obtained from the standard within-individual decoding, where DNN features were predicted using the decoders trained on the same subject’s data (“Within”).

We first evaluated feature decoding performance obtained from the whole VC (in the target space) of the converted fMRI activity. The results of the neural code converter (Across-functional) showed lower but comparable performance with the within-individual results, with similar trends across layers, both in individual pairs and at the group level (Fig. S5A for results of individual pairs; Fig. 3A for group results). Anatomical alignment (Across-anatomical) had the poorest performance among the three conditions, with accuracies below 0.1 in most layers, both in individual pairs and at the group level. The results show that the neural code converters have an advantage over anatomical alignment in DNN feature decoding. Other methods of functional alignment showed similar DNN decoding accuracies, with optimal transport showing lower accuracies (Fig. S6).

We next performed decoding analyses on each DNN unit using voxels from individual visual areas (V1–V4 and HVC in the target space) and identified the visual area that gave the highest decoding accuracy for each unit (“top visual area”), following Nonaka et al. (2021). We then computed the distribution of the top visual area across DNN units in a given layer. We observed a shift of the peak area, from lower to higher areas, along the DNN hierarchy in all conditions (Fig. S5B for results of individual pairs; Fig. 3B for group results). To quantify the degree of hierarchical correspondence between brain areas and DNN layers, we used the decoding-based brain hierarchy (BH) score (Nonaka et al., 2021), which is based on the rank correlation between the hierarchical levels of the DNN layer and the top brain area across DNN units (Fig. 3C; see Materials and Methods: “Brain hierarchy (BH) score”). The results of the within-individual condition replicated the previous findings with a BH score of around 0.5 (Horikawa and Kamitani, 2017; Nonaka et al., 2021). Despite the low accuracies in feature decoding with anatomical alignment (Across-anatomical; Fig. 3A), the hierarchical correspondence was largely preserved when quantified by the BH score (Fig. 3B, C). This is presumably because anatomical alignment maps a macroscopic organization of hierarchical visual areas between subjects, and the relative amount of information about hierarchy is preserved. The inter-individual conversion (Across-functional) showed a lower but substantial degree of hierarchical correspondence even though the converter was blind to cortical hierarchy information during training.

2.4. Visual image reconstruction

After confirming that multiple levels of DNN feature representations can be decoded from converted brain activity, we next sought to determine if we could reconstruct visual images via DNN features decoded from converted brain activities (deep image reconstruction, Shen et al., 2019b; see Materials and Methods: “Visual image reconstruction”). Along with the natural images, we also performed the reconstruction analysis on the artificial images of simple geometric shapes (see Materials and Methods: “fMRI datasets”).

We first show examples of the reconstructions from VC for the Within, Across-anatomical, and Across-functional conditions (Fig. 4A). The reconstructed images obtained in the Within and Across-functional conditions captured the main characteristics of the presented images, including the shapes and colors of the objects, while reconstructions with anatomical alignment (Across-anatomical) showed neither a recognizable shape nor color of the objects in the presented images (see Figs. S7 and S8 for other examples of natural images and artificial images). Other methods of functional alignment also produced similar reconstructions, but optimal transport slightly underperformed compared with others (Fig. S9). Here, we only present reconstructions from the average fMRI data over all the repetitions (24 and 20 repetitions for natural and artificial images, respectively). The results with the average of different numbers of repetitions are available in the supplemental information (Fig. S10). Notably, even fMRI data of a single repetition could produce discernible reconstructions, with the visual quality increasing with more repetitions.

To quantitatively evaluate our reconstruction results, we performed a pairwise identification analysis in which the pixel or DNN feature pattern of a reconstruction was used to identify the true stimulus between two alternatives by choosing the one with a more correlated pattern (see Materials and Methods: “Identification analysis”). DNN feature patterns were extracted using the AlexNet model (Krizhevsky et al., 2012), which is different from the DNN used in our reconstruction method (VGG19

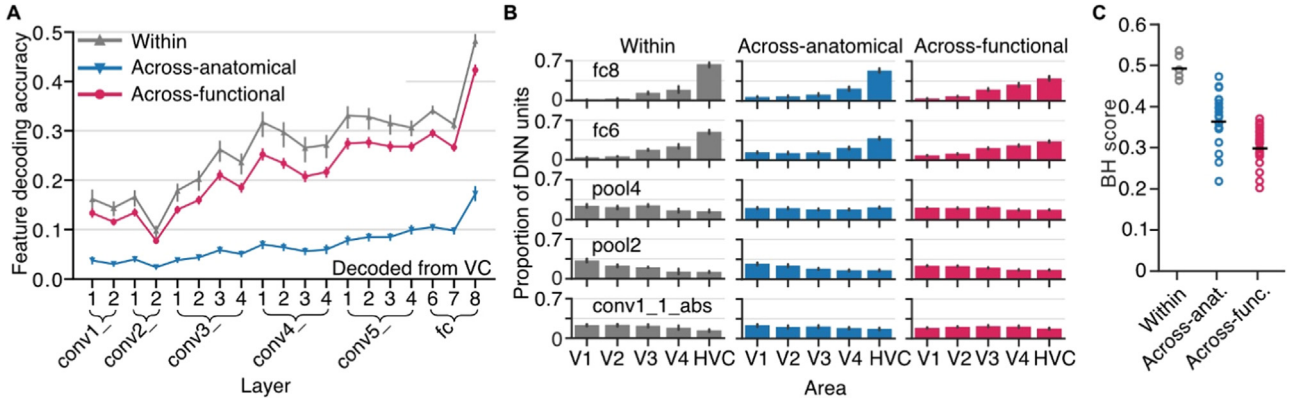


Fig. 3. DNN feature decoding and hierarchical representation.
(A) DNN feature decoding accuracy from the whole visual cortex (VC). Decoding accuracies for each layer of the VGG19 model are shown for the Within, Across-anatomical, and Across-functional conditions (error bars, 95% C.I. from five subjects for the Within condition, and from 20 individual pairs for the Across-anatomical and Across-functional conditions).
(B) Proportion the "top visual area" (best decodable area for each DNN unit) across DNN units in each layer. Only five representative layers are shown. Each bar indicates the mean proportion of DNN units over five subjects for the Within condition or over 20 individual pairs for the Across-anatomical and Across-functional conditions (error bars, 95% C.I. from five subjects or 20 pairs.).
(C) Brain hierarchy (BH) score. The horizontal black dashes indicate the mean BH score over subjects or pairs; each circle represents the BH score for a subject or a pair.

model). The identification was repeated for multiple false alternatives to obtain the accuracy for each reconstruction. For group analysis, the mean identification accuracy was calculated over all reconstructions in each pair. While the within-individual condition (Within) showed overall superior performance both for natural and artificial images, neural code conversion (Across-functional) greatly outperformed anatomical alignment (Across-anatomical) both in individual pairs and at the group level (Fig. S11 for individual pairs; Fig. 4B for group results).

2.5. Visual subarea-wise conversion

To examine whether constraining neural code conversion to respect cortical hierarchy could improve visual image reconstruction, we performed subarea-wise conversion that predicted the activity values of a voxel in a target area only from the source subject's corresponding source area (Fig. 5A). All individual pairs showed comparable conversion accuracies to the whole VC conversion, with the mean pattern correlation being 0.58 ± 0.07 and the mean profile correlation being 0.55 ± 0.06 for VC (Fig. S12A for individual pairs; Fig. S12B for group results). We then performed DNN feature decoding and visual image reconstruction using whole VC, and compared the results with the whole VC conversions (*c.f.*, Figs. 3 and 4).

In the DNN feature decoding of the natural images, the subarea-wise conversion showed similar but slightly lower decoding accuracy than the whole VC conversion across layers in all individual pairs (Fig. S12C) and at the group level (Fig. 5B; ANOVA on the means of individual pairs, effect of conversion type with the DNN layer as a between-subject factor, $F(1, 361) = 1959$, $p < .001$, $\eta_p^2 = 0.84$; see Materials and Methods: "Statistics"). Similar results were obtained for the artificial images in some individual pairs and at the group level (Fig. S12C for individual pairs; Fig. S12D for group results; ANOVA on the means of individual pairs, $F(1, 361) = 260.6$, $p < .001$, $\eta_p^2 = 0.42$).

Reconstructed images obtained from subarea-wise conversions exhibited a visual quality similar to those of the whole VC conversions (Fig. 5C). In the identification analysis of the natural images (Fig. 5D), only 2/20 pairs showed significantly higher accuracies for the subarea-wise conversion; 6/20 pairs showed higher significant accuracies for the whole conversion (Fig. S12E; ANOVA in individual pairs; effect of conversion type with the DNN layer feature as a between-subject factor). At the group level, the subarea-wise conversion showed lower accuracies (Fig. 5D; ANOVA on the means of individual pairs, $F(1, 171) = 11.2$, $p < .001$, $\eta_p^2 = 0.062$). In the identification analysis of the artificial images, 3/20 pairs showed higher significant accuracies for the subarea-wise conversion; 2/20 pairs showed higher significant accuracies for the whole VC conversion (Fig. S12E; ANOVA in individual pairs), while no statistical difference was found at the group level (Fig. 5D; ANOVA on the means of individual pairs, $F(1, 171) = 3.87$, $p = 0.051$, $\eta_p^2 = 0.022$). These results indicate that constraining neural code conversion to respect cortical hierarchy does not seem to contribute to the improvement of visual image reconstruction. Rather, the flexibility of the mapping with the whole VC conversion could be beneficial as indicated by the slightly superior performance with the natural images.

2.6. Varying the number of training data

One potential benefit of the inter-individual analysis is the reduction of the number of data required for model training from novel test (source) subjects by using training data from other individuals. The results of the inter-individual analysis so far were obtained using 2,400 samples for converter training, we here investigated how the number of training samples affects image reconstruction quality by varying the number of data used for converter training (300, 600, 900, 1,200, 2,400, 3,600, 4,800, and 6,000 training samples) while using all data of the target subject for decoder training (6,000 samples). We also compare the results between the whole VC and the subarea-wise conversions.

The reconstructed images retained a discernible quality even with a reduction in the number of training samples. Specifically, using converters trained on 300 samples still produced recognizable images in both the whole VC and subarea-wise conversions (Fig. 6A; similar results were obtained for the artificial images, see Fig. S13A). This result indicates that image reconstruction using converters with a small number of training data is feasible, without the need to collect a full set of fMRI data for each subject.

The identification accuracies increased with the number of training samples, approaching the accuracy of the within-individual (Within) condition (see Fig. S14A for individual pairs and Fig. 6B for group results). The subarea-wise and whole VC conversions showed similar accuracies with more than 1,200 training samples, but the subarea-wise conversion outperformed the whole VC conversion with 1,200 or fewer training samples (ANOVA within individual pairs at each training sam-

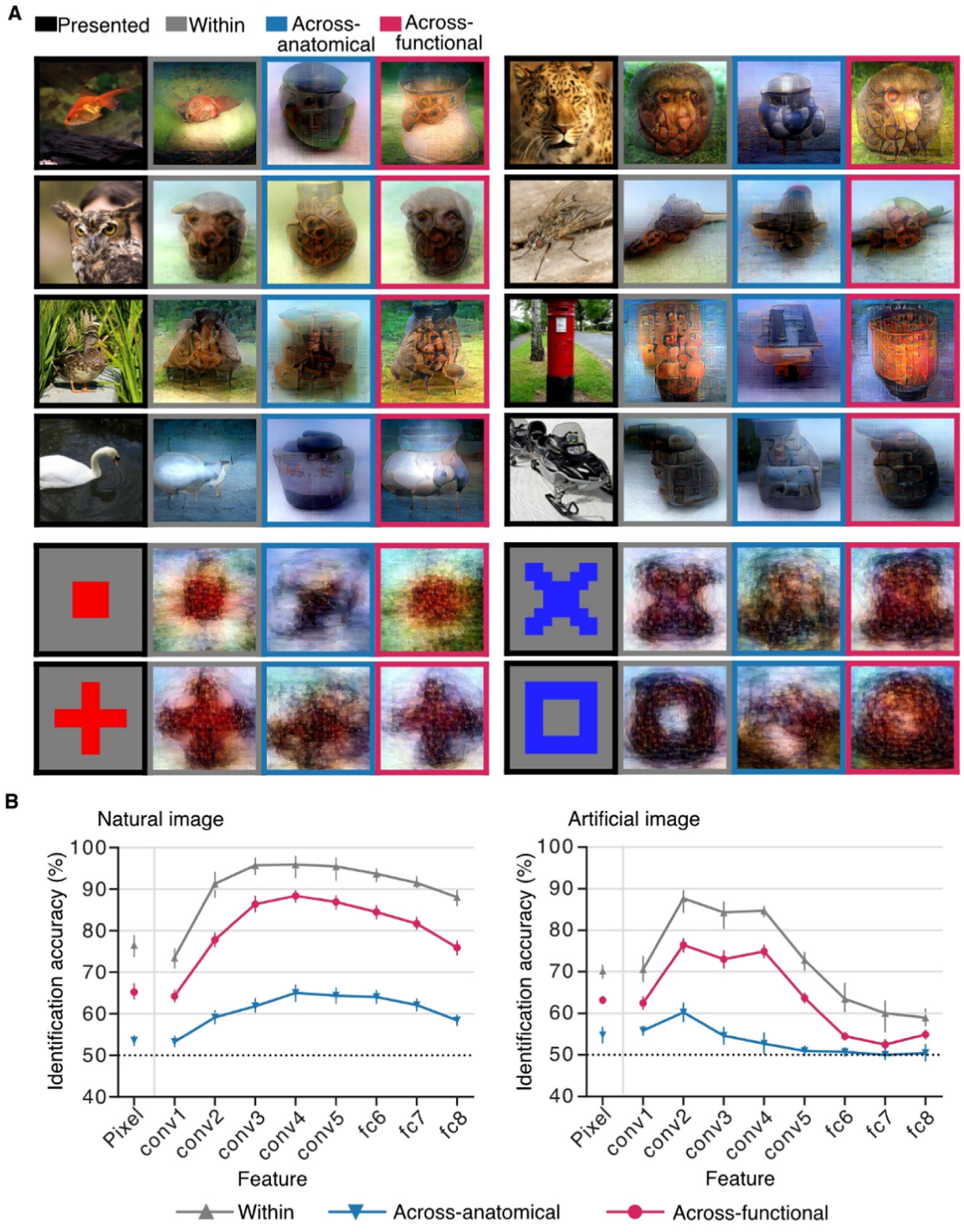


Fig. 4. Reconstructed images and evaluations.
(A) Within and across-individual reconstructions from the whole visual cortex (VC). The reconstructions shown under the three analytical conditions for each stimulus image were all from the same source subject. The results for different stimulus images are from different source subjects.
(B) Identification accuracy based on pixel values and extracted DNN feature values. A mean identification accuracy was calculated over all reconstructed images for each subject or individual pair. DNN features of images were extracted from the eight layers of the AlexNet model (left, natural images; right, artificial images; error bars, 95% C.I. from five subjects or 20 pairs; dotted lines, chance level = 50%).

ple number, effect of conversion type, $p < .05$ in 18, 10, 4, 3, 2, 1, 4, and 1 out of 20 pairs for the eight training sample numbers, respectively; group analysis on the mean accuracies of individual pairs, $p < .05$ at 300, 600, and 900 samples; Bonferroni-corrected by eight). Similar results were obtained for the artificial images (Fig. S14B for individual pairs; ANOVA within individual pairs, effect of conversion type, $p < .05$ in 10, 8, 5, 2, 2, 1, 1, and 2 out of 20 pairs for the eight training sample numbers, respectively; Fig. S13B for group results; group analysis on the mean accuracies of individual pairs, $p < .05$ at 300, 600, and 900 samples; Bonferroni-corrected by eight). Overall, the neural code conversion

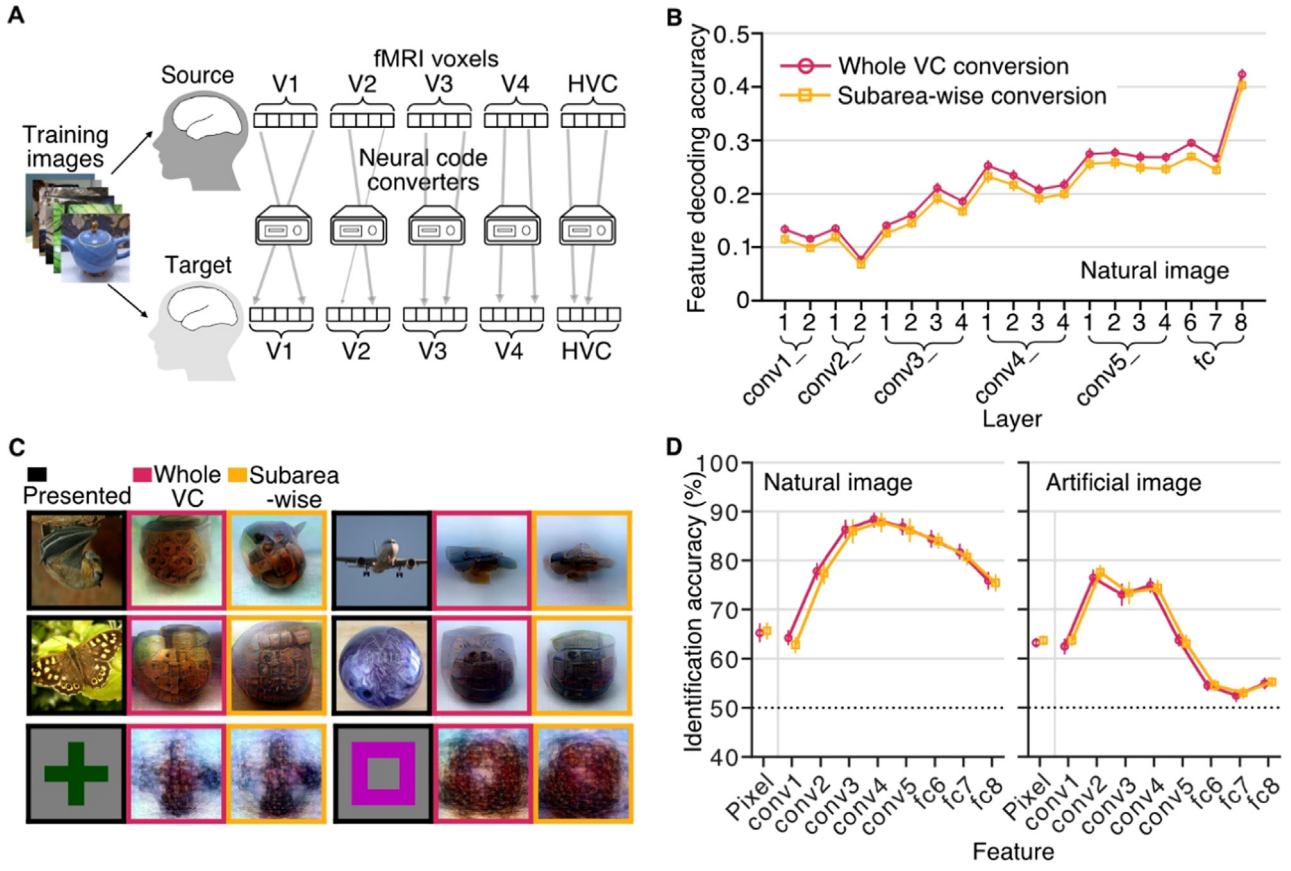


Fig. 5. Whole VC vs. subarea-wise conversion.
(A) Illustration of the subarea-wise conversion. A converter model was trained on a set of fMRI responses to an identical stimulus sequence. Activity values of a voxel in a target area were predicted only from source subject brain activity patterns in the voxel's corresponding source area.
(B) DNN feature decoding accuracy. The whole VC conversion and subarea-wise conversion were evaluated using DNN feature decoding of natural images (error bars, 95% C.I. from 20 individual pairs).
(C) Reconstructed natural and artificial images.
(D) Identification accuracies based on pixel values and extracted DNN feature values. DNN features of images were extracted from the eight layers of the AlexNet model (left, natural images; right, artificial images; error bars, 95% C.I. from 20 individual pairs; dotted lines, chance level = 50%).

with the cortical hierarchy constraint does not improve reconstruction, but it is beneficial when the number of training data is limited.

2.7. Pooling data from multiple subjects

Finally, because the neural code conversion allowed us to pool the data of multiple subjects into a single target subject brain space, we examined the pooling effect on the performance of the inter-individual visual image reconstruction. For a pair of a source and a target subject, we pooled all data from the other three subjects into the target brain space (4 subjects × 6,000 samples = 24,000 samples in total; whole VC conversion; Fig. 7A). We re-trained the decoders on this pooled data and called the decoders "multiple-subject feature decoders," in contrast to the "single-subject feature decoders," which were trained on the target subject in the native brain space. For the neural code converter training between the source subject's data and the pooled data, 2,400 samples of the source subject were paired with each set of 2,400 samples from the four pooled subjects. The converted brain activity from the source subject underwent DNN feature decoding with the multiple-subject feature decoders and then visual image reconstruction. The results were compared with those generated from the single-subject feature decoders.

DNN feature decoding analysis on the natural images showed a small improvement in accuracy across all layers in the multi-subject condition as compared with the single-subject condition. The multiple-subject condition yielded better performance than the single-subject condition both in individual pairs (Fig. S15A) and at the group level (Fig. 7B; ANOVA, effect of decoder type, $F(1, 361) = 1968$, $p < .001$, $\eta_p^2 = 0.85$). Similar results were obtained for artificial images, with the multiple-subject condition showing higher accuracies (see Fig. S16A for individual pair results and Fig. S16B for group results; ANOVA, effect of decoder type, $F(1, 361) = 172$, $p < .001$, $\eta_p^2 = 0.32$). Reconstructed images obtained using both the single- and multiple-subject decoders showed recognizable visual quality, but the visual qualities were not substantially different between the two conditions (Fig. 7C; see Fig. S16C for artificial images). In the identification analysis of the reconstructed natural images, the multiple-subject condition showed slightly higher accuracies than the single-subject condition (Fig. S15B for individual pairs; ANOVA, effect of decoder type, $p < .05$ in 10/20 individual pairs; Fig. 7D for group results; effect of decoder type, $F(1, 171) = 75.6$, $p < .001$, $\eta_p^2 = 0.30$). Similar results were obtained for artificial images, with the multiple-subject condition showing slightly higher identification accuracies than the single-subject condition (Fig. S16D for individual pairs; ANOVA, effect of decoder type, $p < .05$ in 6/20 individual pairs; Fig. S16E for group results; effect of decoder type, $F(1, 171) = 35.9$, $p < .001$, $\eta_p^2 = 0.17$).

To examine the impact of limited data availability on the benefits of pooling multiple-subject data, we conducted a similar analysis using only 300 training samples for the source subject, reflecting situations where data collection is restricted due to cost constraints (Fig. S17).

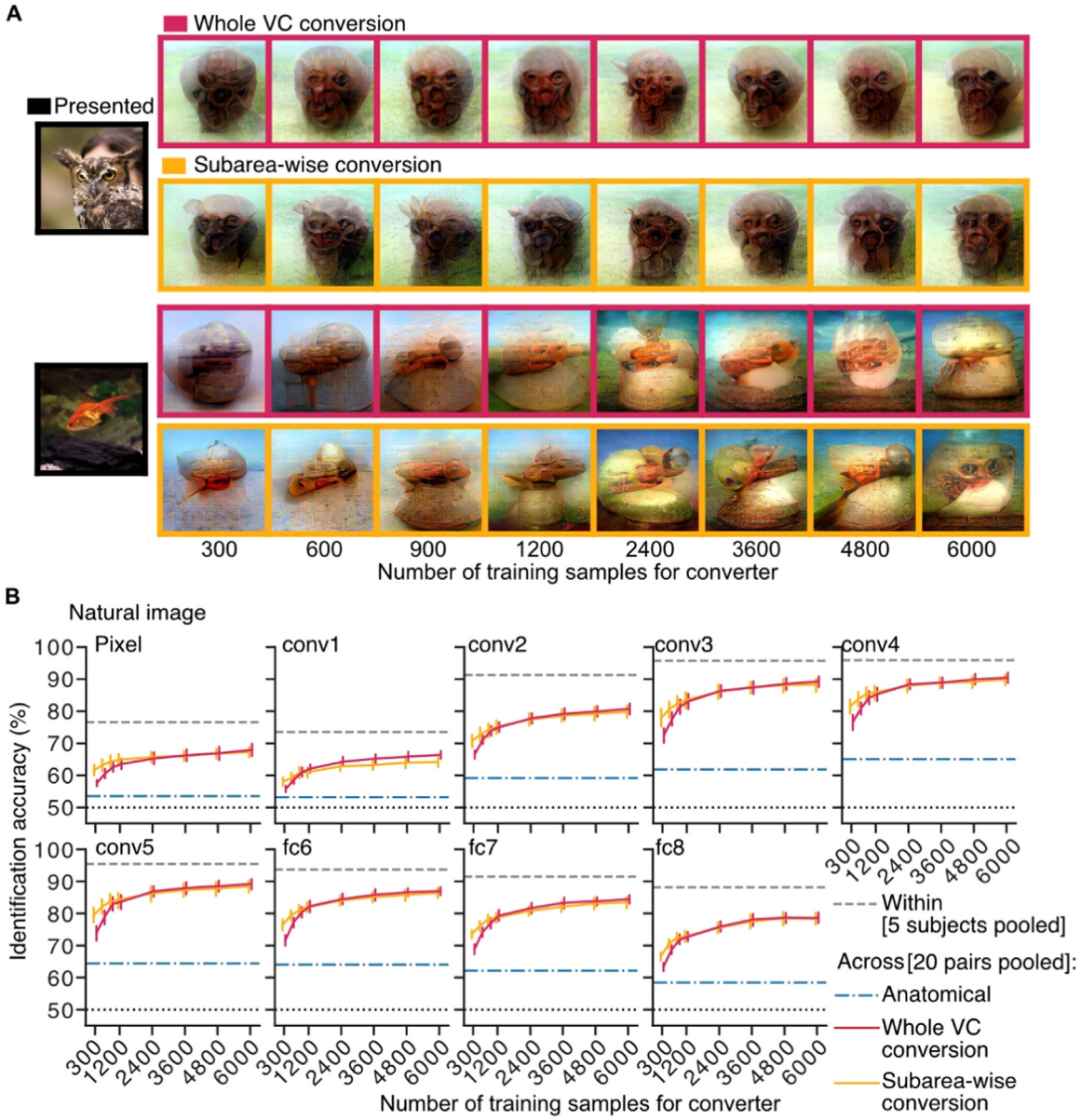


Fig. 6. Effect of the number of training data for the converter.
(A) Reconstructed images. All reconstructed images were produced from the same subject pair (source: Subject 2, target: Subject 3).
(B) Identification accuracy. Identification accuracies were calculated with the pixel values and the extracted DNN feature values (AlexNet) from the reconstructed natural images with varying numbers of training data for the whole VC and subarea-wise converters. The results are shown together with those from the within-individual condition (Within) and the anatomical alignment (Across-anatomical) (error bars, 95% C.I. from 20 individual pairs for whole VC and subarea-wise conversions; dotted lines, chance level = 50%).

There was also a slight improvement for the identification accuracies of the reconstructed natural images in some of the pairs and at the group level when using the multiple-subject feature decoders (Fig. S17D for individual pairs; ANOVA, effect of decoder type, $p < .05$ in 5/20 individual pairs; Fig. S17E for group results; effect of decoder type, $F(1, 171) = 28.3$, $p < .001$, $\eta_p^2 = 0.14$). These results indicate that pooling multiple-subject data is somewhat beneficial for improving the accuracy of inter-individual decoding and reconstruction, even when data availability is limited. However, the visual quality of the reconstructed images was not largely improved.

We additionally performed different pooling data paradigms by projecting directly other subjects' data to a subject's brain space and retraining decoders on the pooled data to examine if this pooling method could yield any improvement. DNN feature decoding analysis and visual image reconstruction were then performed using the subject's data. However, the results showed no improvement (Figs. S18, S19).

3. Discussion

This study aimed to investigate whether and how hierarchical and fine-grained visual information could be converted while preserving perceptual content across individuals using methods of pairwise functional alignment. The study started by showing that methods of pairwise functional alignment can accurately convert a source subject's brain activity into a target subject's brain space by evaluation using the pattern and profile correlations. The ablation analysis on the converters with the exclusion of voxels from various source visual subareas showed that the

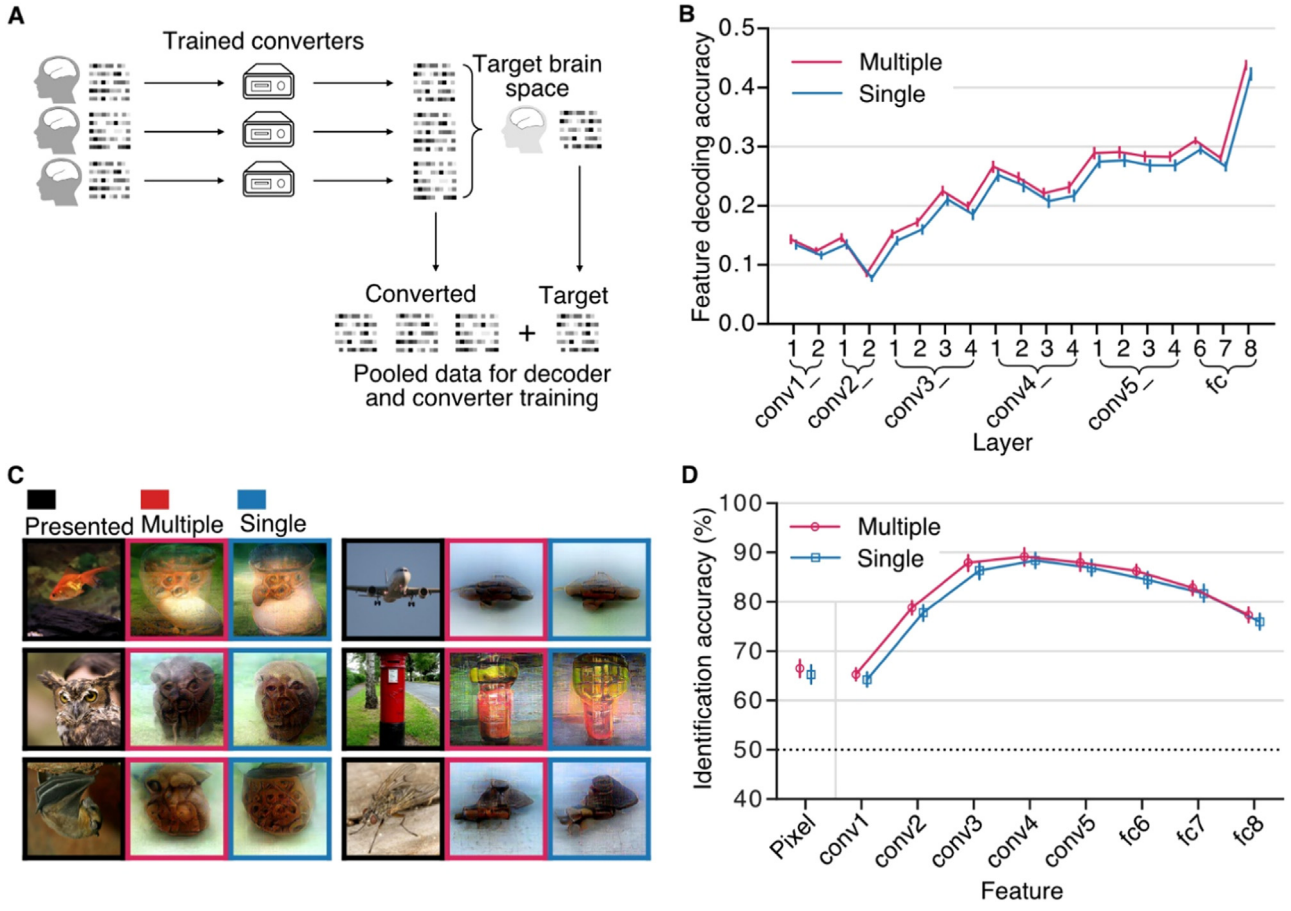


Fig. 7. Pooling data from multiple subjects.
(A) Illustration of the pooling procedure. For a pair of a source (not shown) and a target subject, the training data of the other three subjects were converted into the target subject's brain space and DNN feature decoders were re-trained on the converted data of the three subjects plus the target subject's data (24,000 samples).
(B) DNN feature decoding accuracy obtained via multiple- and single-subject feature decoders. The multiple-subject feature decoders were trained on pooled data, while the single-subject feature decoders were trained on a single subject's data. The accuracies were obtained from the source subjects test dataset of natural images (error bars, 95% C.I. from 20 individual pairs).
(C) Reconstructed natural images for the multiple- and single-subject conditions.
(D) Identification accuracies with the natural images. The identification analysis was performed using the pixel values and the extracted DNN feature values of the reconstructions obtained via multiple- and single-subject feature decoders (error bars, 95% C.I. from 20 individual pairs; dotted lines, chance level = 50%).

converters automatically detected the hierarchical correspondences of visual subareas between individuals. Decoding the converted brain activity into DNN features revealed the correspondence between visual subareas and DNN layers. Visual images were reconstructed from the converted brain activity with recognizable shapes and colors of the objects in the presented images. While the whole VC conversion slightly outperformed the subarea-wise conversion in the inter-individual visual image reconstruction with sufficient training data, the subarea-wise conversion performed better with minimal data. The results indicate that the whole VC conversion preserves the hierarchical structure that is explicitly assumed in the subarea-wise conversion. Even with a small number of training data, the converters preserved minimally sufficient information for visual image reconstruction. Pooling data from multiple subjects helped achieve slightly higher accuracy in the visual image reconstruction, even though the visual quality was not greatly improved. Our analyses demonstrate that hierarchically organized fine-grained visual features that enable visual image reconstruction are preserved in the converted brain activity, allowing efficient reconstruction of visual images without training subject-specific models.

We have shown that the neural code converters automatically detected the hierarchical correspondence of visual subareas between two individuals without explicitly labeling the visual areas (Fig. 2B, C). Previous studies of functional alignment mainly targeted a specific brain area, such as V1 or the inferior temporal cortex (Yamada et al., 2015; Haxby et al., 2011). Other studies functionally aligned a large region of the cortex (Bilenko and Gallant, 2016; Van Uden et al., 2018), but their subsequent analyses targeted different research questions such as the retinotopic organization and the semantic information, leaving the hierarchical correspondences of visual subareas remained undiscussed. Our results explicitly demonstrate that machine learning-based neural code converters can learn the hierarchical correspondence of visual subareas between two individuals. In addition, we observed some inter-regional predictions (e.g., source V1 could predict voxel values in target V2, Fig. 2C), but the prediction accuracy decreased as the cortical distance between two areas increased, presumably indicating that two close areas share some common information.

By decoding the converted fMRI activity patterns into DNN features and reconstructing them as visual images via the decoded DNN features (Figs. 3 and 4), we showed that hierarchically organized fine-grained visual features that enable visual image reconstruction are preserved in the neural code conversion. Previous studies have mainly focused on some specific features, such as object categories, image contrast,

retinotopy, and semantics (Haxby et al., 2011; Yamada et al., 2015; Bilenko and Gallant, 2016; Van Uden et al., 2018), but whether a set of hierarchical fine-grained features is preserved after functional alignment remained unknown. The results of DNN feature decoding on multiple levels of DNN layers showed that the converted fMRI activity patterns held multiple levels of fine-grained visual features (Fig. 3). Moreover, successful visual image reconstruction further confirmed that the converted fMRI activity patterns preserved sufficient perceptual content for reconstructing visual images (Fig. 4).

The subarea-wise conversion marginally underperformed the whole VC conversion with sufficient training data in the inter-individual visual image reconstruction (Fig. 5), with the whole VC conversions achieving slightly higher DNN feature decoding accuracy and marginally higher identification accuracy in the reconstruction. However, the subarea-wise conversion did outperform the whole VC conversion when there was little data (Fig. 6). This result shows that when sufficient training data are available, the whole VC conversion can implicitly learn the information about explicit labels of visual subareas, without the need to explicitly impose the hierarchy constraint.

Training a full visual image reconstruction model requires an fMRI dataset that is costly and takes a long time to collect. In this study, the DNN feature decoders were trained on 6,000 data samples, which took approximately 800 mins of data collection time. In fMRI studies, this long data collection time is impractical for most people. However, by trading off some visual quality of the reconstructed images, it is possible to collect fewer data samples, for instance, 300 samples, to train a neural code converter and perform inter-individual visual image reconstruction. In particular, the neural code converter is designed to capture the relationships between individuals' voxels across a range of visual scenes, and could potentially be used in combination with other decoding models. The inter-individual decoding method with the neural code converter has the potential to reduce the time and costs of fMRI data collection.

Our results show that pooling data from multiple subjects did not greatly enhance visual image reconstruction performance (Fig. 7). A possible reason is the variability of data quality, with some subjects' data leading to relatively poor visual quality in the reconstructed images. Poor quality data limited the capability of the decoders to leverage the pooled data and resulted in a limited improvement in visual image reconstruction performance. Furthermore, linear regression models may not be able to fully address the feature mismatch of brain activity patterns between individuals, and more advanced methods may be necessary to solve this problem (Li et al., 2021). Despite our findings indicating that pooling data did not lead to great improvements in visual image reconstruction quality, it is still a promising direction for future fMRI research.

We observed that the inter-individual image reconstruction did not outperform the within individual image reconstruction (Figs. 4 and 5). As mentioned above, this could be due to the linear constraint used for conversion, which may be too restrictive to capture more complex statistical relationships, such as nonlinearity between brain activity patterns. Additionally, a brain's response to a stimulus comprises a consistent stimulus-evoked response across individuals, an idiosyncratic stimulus-evoked response and a noise component (Nastase et al., 2019). The brain decoders might leverage the idiosyncratic responses that could not be converted across subjects, as well as noise components. As a result, the inter-individual visual image reconstruction thus underperformed the within-individual visual image reconstruction.

Our results showed that it is possible to translate brain activity patterns across individuals while retaining sufficient information to visualize the perceived stimulus. This presents an efficient approach for reconstructing visual images without the need to train subject-specific models, especially for complex and data-intensive reconstruction models. By reducing the amount of data required for model training, our method could help to promote the use of brain-machine/computer interfaces that communicate with our internal world.

4. Materials and methods

4.1. fMRI datasets

4.1.1. Subjects

In this study, Subject 1–3 correspond to the three subjects in Shen et al. (2019b) and the dataset was reused. Subject 4 (male, age 22) and Subject 5 (male, age 27) participated in our additional experiments for the test natural-image and artificial-image sessions. The dataset of the training natural-image session of Subject 4 and 5 was reused from Horikawa and Kamitani (2022). All subjects provided written informed consent for participation in the experiments, in accordance with the Declaration of Helsinki, and the study protocol was approved by the Ethics Committee of Advanced Telecommunications Research Institute International (ATR).

4.1.2. Stimuli

The natural image stimuli in Horikawa and Kamitani (2017) were selected from 200 representative categories in the ImageNet dataset (2011, fall release; Deng et al., 2009). The natural training images were 1,200 images taken from 150 object categories, and the natural test images were 50 images taken from the remaining 50 object categories. The artificial image stimuli used in Shen et al. (2019b) consisted of 40 combinations of five shapes (square, small frame, large frame, plus sign, and cross sign) and eight colors (red, green, blue, cyan, magenta, yellow, white, and black).

4.1.3. Experimental design

In both Horikawa and Kamitani (2017), and Shen et al. (2019b), fMRI signals were measured while subjects viewed a sequence of visual images. The visual images had a central fixation spot and were flashed at a frequency of 1 Hz. Each presentation of an image lasted for 8 s in a stimulus block with four volume scans (Repetition time [TR] = 2 s). The subjects were instructed to maintain fixation on the central fixation spot and click a button when two sequential blocks presented the same image.

The test natural-image session and test artificial-shape session consisted of 24 and 20 runs, respectively. Each run consisted of 55 and 44 stimulus blocks comprising 50 and 40 blocks of different images, and 5 and 4 randomly interspersed repetition blocks, along with additional 32-s and 6-s rest periods at the beginning and the end. The 50 natural images and 40 artificial images were presented in random order in each run.

4.1.4. fMRI data preprocessing

The following description is provided by fMRIPrep (https://fmriprep.org/en/1.2.1/citing.html). The results included in this manuscript are based on the data preprocessed using fMRIPrep version 1.2.1 (Esteban et al., 2019) and a Nipype-based tool (Gorgolewski et al., 2011, 2017). Each T1w (T1-weighted) volume was corrected for INU (intensity non-uniformity) using N4BiasFieldCorrection v2.2.0 (Tustison et al., 2010) and skull-stripped using antsBrainExtraction.sh v2.2.0 (using the OASIS template). Brain surfaces were reconstructed using recon-all from FreeSurfer v6.0.1 (Dale et al., 1999), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (Klein et al., 2017). Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov et al., 2009) was performed through nonlinear registration with the antsRegistration tool of ANTs v2.2.0 (Avants et al., 2008), using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter

(WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (Zhang et al., 2001; FSL v5.0.9).

Functional data were slice time corrected using 3dTshift from AFNI v16.2.07 (Cox et al., 1996) and motion corrected using mcflirt (FSL v5.0.9; Jenkinson et al., 2002). This was followed by co-registration to the corresponding T1w using boundary-based registration (Greve et al., 2009) with 9 degrees of freedom, using bbregister (FreeSurfer v6.0.1). Motion correcting transformations, BOLD-to-T1w transformation, and T1w-to-template (MNI) warp were concatenated and applied in a single step using antsApplyTransforms (ANTs v2.2.0) using Lanczos interpolation.

Physiological noise regressors were extracted by applying CompCor (Behzadi et al., 2007). Principal components were estimated for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). A mask to exclude signals with cortical origin was obtained by eroding the brain mask, ensuring that it only contained subcortical structures. Six tCompCor components were then calculated including only the top 5% variable voxels within that subcortical mask. For aCompCor, six components were calculated within the intersection of the subcortical mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run. Frame-wise displacement (Power et al., 2013) was calculated for each functional run using the implementation of Nipype.

Many internal operations of fMRIPrep use Nilearn (Abraham et al., 2014), principally within the BOLD-processing workflow. For more details of the pipeline see http://fmriprep.readthedocs.io/en/latest/workflows.html.

The coregistered data to the T1w space were then re-interpolated to $2 \times 2 \times 2$ mm voxels. The data samples were first shifted by 4-s (two volumes) to compensate for the hemodynamic delay, followed by regression to remove nuisance parameters such as a constant baseline, linear trend, and six head motion parameters from each voxel amplitude for each run. The data samples were then despiked to reduce extreme values (beyond ±3 standard deviations for each run) and were averaged within each 8-s trial (four volumes).

4.2. Regions of interest (ROIs)

Regions V1, V2, V3, and V4 were delineated using the standard retinotopy experiment (Engel et al., 1994; Sereno et al., 1995) in each subject's naive brain space. The higher visual cortex (HVC) was defined as a contiguous region covering the LOC, FFA, and PPA, which were identified using conventional functional localizers (Kourtzi and Kanwisher, 2000; Kanwisher et al., 1997; Epstein and Kanwisher, 1998). The whole visual cortex (VC) was defined as the combined regions of V1, V2, V3, V4, and HVC.

4.3. Anatomical alignment

For the analyses with anatomical alignment, the subjects' structural and functional images were nonlinearly normalized to a standard space: the ICBM 152 Nonlinear Asymmetrical template version 2009c (MNI152NLin2009cAsym [MNI space]; see Materials and Methods: "fMRI data preprocessing"). The T1w reference image was spatially normalized to MNI space by the ANTs (Avants et al., 2008) and the functional data were coregistered to this normalized T1w reference image. The coregistered data were then re-interpolated to $2 \times 2 \times 2$ mm voxels. Furthermore, ANTs were used to normalize the ROI masks of V1, V2, V3, V4, and HVC in their native space to the brain in MNI space. In the inter-individual analysis, if a voxel of a source subject and a voxel of a target subject shared the same coordinates, the fMRI activity of the source voxel was considered to be that of the corresponding target voxel. Thus, the voxels of a source subject covered by a ROI mask of a target subject were selected as the input to the model.

4.4. Methods of functional alignment

4.4.1. Neural code converter

The neural code converter model for each pair of subjects comprised a set of regularized linear regression models (ridge regression), each trained to predict the activities of an individual voxel of one subject (target) from the brain activity patterns of another subject (source) given the same stimuli. A converter takes a source subject's brain activity pattern $\mathbf{x}_i \in (\mathbb{R})^m$ consisting of m voxels' values, and predicts the target brain activity pattern $\mathbf{y}_i = \mathbf{W}\mathbf{x}_i + \mathbf{b}$, where $\mathbf{y}_i \in (\mathbb{R})^n$ is the converted brain activity pattern consisting of n voxels' values; $\mathbf{W} \in (\mathbb{R})^{n \times m}$ is the conversion matrix and $\mathbf{b} \in (\mathbb{R})^n$ is the bias vector. The converter is trained to minimize the objective function

$$\sum_i^N \left|\left|\mathbf{y}_i - \left(\mathbf{W}\mathbf{x}_i + \mathbf{b}\right)\right|\right|^2 + \lambda||\mathbf{W}||^2,$$

where $\mathbf{y}_i$ is the measured target subject's brain activity pattern for the i-th sample, N is the number of training samples, λ is the regularization parameter, and $||\cdot||$ represents the Frobenius norm.

To optimize the performance of visual image reconstruction, we fine-tuned the regularization parameter through 5-fold cross-validation on the training data. At each fold, the brain activity in the validation set was converted to the target's brain space and then decoded into DNN features. The decoded DNN features were used to calculate an identification accuracy that measured how well a decoded DNN feature pattern can identify the true stimulus between two alternatives (see Materials and Methods: "Identification analysis"). The regularization parameter was optimized in a grid-search manner to maximize the identification accuracy, which is linked to the performance of visual image reconstruction. 500 units instead of all from each layer of the VGG19 model were chosen to save the computational time, and were randomly chosen because there was no a priori knowledge about which DNN units lead to a better visual image reconstruction.

4.4.2. Procrustes transformation

Procrustes transformation is a transformation that includes translation, rotation and uniform scaling, and preserves the shape of a geometric object. It was first introduced by Haxby et al. (2011) in the hyperalignment analysis. Considering the source and target subjects' brain activity patterns $\mathbf{x}_i \in (\mathbb{R})^m$ and $\mathbf{y}_i \in (\mathbb{R})^n$, Procrustes transformation estimates an orthogonal transformation matrix $\mathbf{W} \in (\mathbb{R})^{n \times m}$ to minimize

$$\sum_i^N ||\mathbf{y}_i - \mathbf{W}\mathbf{x}_i||^2,$$

with the constraint $\mathbf{W}^T\mathbf{W} = \mathbf{I}$ where N is the number of training samples. Please refer to Bazeille et al. (2021) for more discussions.

4.4.3. Optimal transport

Optimal transport is related to the question of how one could transform a probability distribution into another probability distribution with the least cost. It was first applied to the functional alignment in Bazeille et al. (2019). Defining $\mathbf{X} = (\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_m)$ and $\mathbf{Y} = (\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_n)$ with $\mathbf{a}_i,\ \mathbf{b}_j \in (\mathbb{R})^N$ representing a sequence of a voxel response to N stimuli, optimal transport tries to find a transformation matrix $\mathbf{W}^*$ such that

$$\mathbf{W}^* = \underset{\mathbf{W}}{\operatorname{argmin}}\left(\sum_{ij} \mathbf{W}_{ij}\left|\left|\mathbf{b}_i - \mathbf{a}_j\right|\right|^2 - \epsilon h(\mathbf{W})\right)$$

with the constraints $\sum_j \mathbf{W}_{ij} = 1/m$ and $\sum_i \mathbf{W}_{ij} = 1/n$.

The entropic term

$$h(\mathbf{W}) = -\sum_{ij} \mathbf{W}_{ij}\left(\log\left(\mathbf{W}_{ij}\right) - 1\right)$$

regularizes the optimal transport problem and ϵ controls the strength of regularization. The regularization parameter was optimized as in the

neural code converter. We used *fmrialign* (https://parietal-inria.github.io/fmralign-docs/index.html) package in the analysis. Please refer to Bazeille et al. (2019) for more mathematical details.

4.4.4. Template-based pairwise alignment via hyperalignment

Template-based pairwise alignment via hyperalignment first estimates a common template among subjects using hyperalignment (Haxby et al., 2011), and then construct a pairwise transformation by first mapping a source subject's brain activity onto the template, followed by an inverse mapping from the template to a target subject's brain space. In the first iteration, the hyperalignment algorithm first selects an initial target subject as a template, then aligns the second subject's fMRI responses to the template using Procrustes transformation. The template is then updated as the mean of the current template and the newly aligned fMRI responses. The same procedure is repeated for additional subjects. In the second iteration, each subject's original response is aligned to the mean aligned responses of other subjects. The mean aligned response is recalculated and treated as a template. In the last step, each subject's response is aligned to the template, and an orthogonal transformation matrix is obtained for each subject.

Although the hyperalignment algorithm can estimate the shared space of more than two subjects, we used it only between two subjects, as in the neural code converter analysis.

4.5. Noise ceiling estimation

Repeated measures of the brain responses to an identical stimulus are subject to measurement noise in fMRI data, inevitably lowering the prediction accuracy. To account for the noise, we adopted the noise ceiling estimation used by Lescroart and Gallant (2019; see also Hsu et al., 2004). The noise ceiling was obtained by averaging the profile or pattern correlation coefficients between repetitions of the same stimuli within a subject. This noise ceiling estimation is based on the rationale that no model can predict better than the subject's own responses. Thus, the noise ceilings reflect the maximum performances of the converter models and were used to normalize the raw prediction accuracies of the converter models by dividing the raw accuracies by the noise ceilings.

Samples/voxels with corresponding noise ceilings below a threshold (99th percentile point in the distribution from random pairs) were excluded from the performance evaluation of the conversion because they could not be reliably measured. But all voxels were included in the downstream DNN feature decoding analysis to prevent information leakage.

4.6. DNN model

We used the VGG19 DNN model (Simonyan and Kisserman, 2014) implemented using the Caffe library (Jia et al., 2014). This model is pre-trained for the 1,000-class object recognition task using the images from ImageNet (Deng et al., 2009; the pre-trained model is available from https://github.com/BVLC/caffe/wiki/Model-Zoo). The model consists of 16 convolutional layers and three fully connected layers. All the input images to the model were rescaled to 224 × 224 pixels. Following Shen et al. (2019b), outputs from individual units before rectification were used as target variables in the DNN feature decoding analysis. The number of units in each layer is as follows: conv1_1 and conv1_2, 3,211,264; conv2_1 and conv2_2, 1,605,632; conv3_1, conv3_2, conv3_3, and conv3_4, 802,816; conv4_1, conv4_2, conv4_3, and conv4_4, 401,408; conv5_1, conv5_2, conv5_3, and conv5_4, 100,352; fc6 and fc7, 4,096; and fc8, 1,000.

We used the AlexNet DNN model (Krizhevsky et al., 2012) implemented using the Caffe library to extract DNN features from the reconstructed images and the presented image. This model is also pre-trained similarly (available from https://github.com/BVLC/caffe/tree/master/models/bvlc_alexnet). The model consists of five convolutional layers and three fully connected layers. The number of units in each layer is as follows: conv1, 290,400; conv2, 186,624; conv3 and conv4, 64,896; conv5, 43,264; fc6 and fc7, 4,096; and fc8, 1,000.

4.7. DNN feature decoding analysis

For each DNN unit, we trained a ridge linear regression model (DNN feature decoder) that takes an fMRI activity pattern induced by a stimulus as input and predicts a feature value of the stimulus. The ridge regularization parameter was set to 100, and both the feature values and voxel values were normalized before model training. We then performed a voxel selection procedure to select the top 500 voxels with the highest Pearson correlation coefficients between the sequences of feature values and voxel responses of all voxels for training. The trained decoders were tested on the average fMRI pattern over repetitions to increase the signal-to-noise ratio of the fMRI signal. For details of the feature decoding, please refer to the works of Horikawa and Kamitani (2017, 2022) and Shen et al. (2019b).

4.8. Brain hierarchy (BH) score

The BH score was originally designed to measure the degree to which an artificial neural network is hierarchically similar to the human brain (Nonaka et al., 2021). In this study, we adopted the decoding-based BH score to see whether the hierarchical similarity is preserved after inter-individual conversion. The DNN features of randomly selected 1,000 units of each layer are decoded from the fMRI pattern of one of the five visual areas: V1–V4 and the HVC. For each unit, the visual area showing the best decoding accuracy was identified and was called the "top visual area." The first layer, the last layer, and three randomly sampled intermediate layers were used to calculate a Spearman rank correlation coefficient between the hierarchical levels of the five DNN layers (coded as 0 through 4) and the top visual area (coded as V1: 0, V2: 1, V3: 2, V4: 3, and HVC: 4) across DNN units. This sampling procedure was repeated 10,000 times, and the mean Spearman rank correlation coefficient was taken as the BH score. See Nonaka et al. (2021) for more details.

4.9. Visual image reconstruction

We used an image reconstruction method (deep image reconstruction) proposed by Shen et al. (2019b). The method optimizes pixel values of an input image based on a set of DNN features given as a target. Given the decoded DNN features from multiple layers, an image was reconstructed by solving the following optimization problem (Mahendran et al., 2015):

$$\mathbf{v}^* = \underset{\mathbf{v}}{\operatorname{argmin}} \left(\frac{1}{2} \sum_{l}^{L} \beta_l \left|\left| \boldsymbol{\varphi}_l(\mathbf{v}) - \mathbf{u}_{il} \right|\right|^2 \right),$$

where $\mathbf{v} \in (\mathbb{R})^{224 \times 224 \times 3}$ is a vector whose elements are the pixel values of an image (width × height × RGB channels); L is the total number of layers; $\boldsymbol{\varphi}_l$ is the function that maps the image to the DNN feature vector of the l-th layer; $\mathbf{u}_{il}$ is the decoded DNN feature vector of the l-th layer for the i-th sample; and β_l is the parameter that weights the contribution of the l-th layer, which was set to be $1/||\mathbf{u}_{il}||^2$.

A natural image prior is applied by introducing a generative adversarial network called the deep generator network (DGN) to enhance the naturalness of the image (Nguyen et al., 2016). The optimization problem becomes

$$\mathbf{z}^* = \underset{\mathbf{z}}{\operatorname{argmin}} \left(\frac{1}{2} \sum_{l}^{L} \beta_l \left|\left| \boldsymbol{\varphi}_l(G(\mathbf{z})) - \mathbf{u}_{il} \right|\right|^2 \right),$$

where G is the DGN and $\mathbf{z}$ is a latent vector. The reconstructed image is obtained by $\mathbf{v}^* = G(\mathbf{z}^*)$. The DGN is a pre-trained generator provided by Dosovitskiy and Brox (2016; available from https://github.com/dosovits/caffe-fr-chairs).

The solution to the above optimization problem is considered to be the reconstructed image from the brain activity pattern. Following Shen et al. (2019b), natural images were reconstructed using the DGN, and the objective function was optimized by stochastic gradient descent with momentum with 200 iterations, whereas the artificial images were reconstructed without the DGN and the objective function was optimized by a limited-memory BFGS algorithm with 200 iterations (Le et al., 2011; Liu and Nocedal, 1989; Gatys et al., 2016).

4.10. Identification analysis

Identification analysis was used to evaluate image reconstruction quality. Presented images were identified using the similarity in either image pixels or DNN features, which were reshaped into a one-dimensional feature vector. The feature vector of a reconstructed image was used to compare the true feature vector of the presented image with the false alternative of another image. The comparison was counted as correctly identified if the feature vector of the reconstruction has a higher Pearson correlation coefficient with the true feature vector than with the false alternative. The identification was repeated for multiple false alternatives for each reconstruction. For natural images, the identification was repeated with 49 alternatives for each reconstruction, resulting in 50 images × 49 comparisons = 2450 comparisons in total. The identification accuracy for a reconstructed image was defined as the proportion of correct identification.

During cross-validation to optimize the regularization parameters for the neural code converters, we used a set of decoded DNN features concatenated from multiple layers to calculate the identification accuracies and evaluate the performance (see Materials and Methods: “Methods of functional alignment”). The candidate images for comparisons were a subset of the 1200 images presented in the training image session.

4.11. Statistics

We performed statistical analyses primarily on the data samples in each pair of subjects to examine the effect of each conversion and its prevalence across pairs (Ince et al., 2022). We additionally performed group level analyses using mean values from 20 individual pairs for summary purposes or when within-pair analysis was not applicable. In some analyses, the results with converted brain activity were compared with those without conversion (within-individual), in which the data from five subjects were similarly processed.

In the evaluation of conversion for each pair of subjects, the pattern correlation coefficients for 50 visual stimuli were used to calculate the mean conversion accuracy (pattern) and its 95% confidence interval, while the profile correlation coefficients for all voxels were used to calculate the mean conversion accuracy (profile) and its 95% confidence interval. At the group level, the mean conversion accuracies (pattern/profile) from 20 individual pairs were used to calculate the group mean and its 95% confidence interval.

In the DNN feature decoding analysis for each conversion or each subject (within), the decoding accuracies (profile correlations in individual units) for all DNN units were used to calculate the mean decoding accuracy and its 95% confidence interval. At the group level, the mean decoding accuracies from 20 individual pairs or five subjects (within) were used to calculate the group mean and its 95% confidence interval.

In the evaluation of the brain hierarchy, a BH score was computed for each conversion or each subject (within). At the group level, the BH scores of 20 individual pairs or five subjects (within) were averaged to obtain the mean BH score.

In the identification analysis of reconstructed images for each conversion or each subject (within), the identification accuracies for individual reconstructed images were used to calculate the mean identification accuracy and its 95% confidence interval. At the group level, the mean identification accuracies from 20 individual pairs or five subjects (within) were used to calculate the group mean and its 95% confidence interval.

In the comparison of the subarea-wise conversion with the whole VC conversion, we performed ANOVA on DNN feature decoding accuracies and identification accuracies with the conversion type as a repeated measure factor and the DNN layer as a between-subject factor. Since millions of DNN units and their decoding accuracies (profile correlations) always lead to statistical significance, we only performed group level ANOVA on DNN feature decoding accuracies to compute the F scores, p values, and the effect size, in which the data points were the mean DNN feature decoding accuracies of 20 individual pairs. For identification, the accuracies with individual reconstructed images were used as the data points in the ANOVA analysis to compute the F scores, p values, and the effect size in each individual pair. At the group level, mean identification accuracies from 20 individual pairs were used as data points to compute the F scores, p values, and the effect size.

Similar ANOVA tests were applied to the pooling data analysis to examine the effect of decoder type (multiple- and single-subject).

Data and code availability

The experimental code and data that support the findings of this study are respectively available from our repository (code for inter-individual deep image reconstruction including neural code converter, Procrustes transformation, optimal transport and hyperalignment: https://github.com/KamitaniLab/InterIndividualDeepImageReconstruction, code for feature decoding: https://github.com/KamitaniLab/dnn-feature-decoding, code for image reconstruction: https://github.com/KamitaniLab/DeepImageReconstruction, code for BH score calculation: https://github.com/KamitaniLab/BHscore) and open data repository (OpenNeuro: https://openneuro.org/datasets/ds003993/versions/1.0.0 for Subject 1–3, https://openneuro.org/datasets/ds003430/versions/1.2.0 for the dataset of training natural-image session for Subject 4 and 5, and https://openneuro.org/datasets/ds001506/versions/1.3.1 for the dataset of test natural-image and artificial-image sessions for Subject 4 and 5).

Ethics statement

All subjects provided written informed consent for participation in the experiments, in accordance with the Declaration of Helsinki, and the study protocol was approved by the Ethics Committee of Advanced Telecommunications Research Institute International (ATR).

Funding

This research was supported by grants from Japan Society for the Promotion of Science (JSPS) KAKENHI (https://www.jsps.go.jp) Grant Numbers JP20H05705/20H05954 to YK, the New Energy and Industrial Technology Development Organization (NEDO; https://www.nedo.go.jp) Grant Number JPNP20006 to YK, and JST CREST (https://www.jst.go.jp/kisoken/crest/) Grant Number JPMJCR18A5/JPMJCR22P3 to YK. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration of Competing Interest

Advanced Telecommunications Research Institute International (ATR) and Honda Motor Co., Ltd. hold a patents (US9020586B2) on the methods of neural code conversion. YK is one of the inventors of the patent.

Credit authorship contribution statement

Jun Kai Ho: Methodology, Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization.

Tomoyasu Horikawa: Methodology, Validation, Investigation, Writing – original draft, Writing – review & editing. **Kei Majima:** Methodology, Writing – review & editing. **Fan Cheng:** Methodology. **Yukiyasu Kamitani:** Conceptualization, Validation, Investigation, Resources, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgments

The authors would like to thank Jong-Yun Park, Shangfeng Jin, Ken Shirakawa, Shuntaro C. Aoki, Mitsuaki Tsukamoto, and Misato Tanaka for helpful comments on the manuscript. The data used in the study were collected using the fMRI scanner and related facilities of Kokoro Research Center, Kyoto University. We thank Kimberly Moravec, PhD, from Edanz (https://jp.edanz.com/ac) for editing a draft of this manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2023.120007.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., et al., 2014. Machine learning for neuroimaging with scikit-learn. Front. Neuroinform. 8, 14. doi:10.3389/fninf.2014.00014.

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. 12, 26–41. doi:10.1016/j.media.2007.06.004.

Bazeille, T., DuPre, E., Richard, H., Poline, J.B., 2021. Thirion B. An empirical evaluation of functional alignment using inter-subject decoding. Neuroimage 245, 118683. doi:10.1016/j.neuroimage.2021.118683.

Bazeille, T., Richard, H., Janati, H., Thirion, B., 2019. Local Optimal Transport For Functional Brain Template Estimation. Information Processing in Medical Imaging. Springer, Hong Kong, p. 11492. doi:10.1007/978-3-030-20351-1_18 June.

Behzadi, Y., Restom, K., Liau, J., Liu, T.T., 2007. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. Neuroimage 37, 90–101. doi:10.1016/j.neuroimage.2007.04.042.

Bilenko, N.Y., Gallant, J.L., 2016. Pyrcca: regularized kernel canonical correlation analysis in Python and its applications to neuroimaging. Front. Neuroinform. 10, 49. doi:10.3389/fninf.2016.00049.

Blumensath, T., Jbabdi, S., Glasser, M.F., Van Essen, D.C., Ugurbil, K., Behrens, T.E.J., et al., 2013. Spatially constrained hierarchical parcellation of the brain with resting-state fMRI. Neuroimage 76, 313–324. doi:10.1016/j.neuroimage.2013.03.024.

Chen, P.-H., Chen, J., Yeshurun, Y., Hasson, U., Haxby, J., Ramadge, P.J., 2015. A reduced-dimension fMRI shared response model. Adv. Neural Inf. Process Syst. 28, 460–468. http://papers.nips.cc/paper/5855-a-reduced-dimension-fmri-shared-response-model.pdf.

Cox, R.W., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29, 162–173. doi:10.1006/cbmr.1996.0014.

Dale, A., Fischl, B., Sereno, M.I., 1999. Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. Neuroimage 9, 179–194. doi:10.1006/nimg.1998.0395.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L., 2009. ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; June 2009, Miami, FL, USA. IEEE, pp. 248–255. doi:10.1109/CVPR.2009.5206848.

Dosovitskiy, A., Brox, T., 2016. Generating images with perceptual similarity metrics based on deep networks. Adv. Neural Inf. Process Syst. 29, 658–666. https://arxiv.org/abs/1602.02644.

Engel, S.A., Rumelhart, D.E., Wandell, B.A., Lee, A.T., Glover, G.H., Chichilnisky, E.-.J., et al., 1994. fMRI of human visual cortex. Nature 369, 525. doi:10.1038/369525a0.

Epstein, R., Kanwisher, N., 1998. A cortical representation of the local visual environment. Nature 392, 598–601. doi:10.1038/33402.

Esteban, O., Markiewicz, C.J., Blair, R.W., Moodie, C.A., Isik, A.I., Erramuzpe, A., et al., 2019. fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat. Methods 16, 111–116. doi:10.1038/s41592-018-0235-4.

Fischl, B., Rajendran, N., Busa, E., Augustinack, J., Hinds, O., Yeo, B.T.T., et al., 2008. Cortical folding patterns and predicting cytoarchitecture. Cereb. Cortex 18, 1973–1980. doi:10.1093/cercor/bhm225.

Fonov, V.S., Evans, A.C., McKinstry, R.C., Almli, C.R., Collins, D.L., 2009. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. Neuroimage 47, S102. doi:10.1016/S1053-8119(09)70884-5.

Gatys, L.A., Ecker, A.S., Bethge, M., 2016. Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. December 2016, Las Vegas, NV, USA. IEEE, pp. 2414–2423. doi:10.1109/CVPR.2016.265.

Gorgolewski, K., Burns, C.D., Madison, C., Clark, D., Halchenko, Y.O., Waskom, M.L., et al., 2011. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. Front. Neuroinform. 5, 13. doi:10.3389/fninf.2011.00013.

Gorgolewski, K.J., Esteban, O., Ellis, D.G., Notter, M.P., Ziegler, E., Johnson, H., et al., 2017. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python doi:10.5281/zenodo.581704.

Greve, D.N., Fischl, B., 2009. Accurate and robust brain image alignment using boundary-based registration. Neuroimage 48, 63–72. doi:10.1016/j.neuroimage.2009.06.060.

Güçlü, U., van Gerven, M.A.J., 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J. Neurosci. 35, 10005–10014. doi:10.1523/JNEUROSCI.5023-14.2015.

Güçlü, U., van Gerven, M.A.J., 2017. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. Neuroimage 145, 329–336. doi:10.1016/j.neuroimage.2015.12.036.

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. Cereb. Cortex 26, 2919–2934. doi:10.1093/cercor/bhw068.

Haxby, J.V., Guntupalli, J.S., Connolly, A.C., Halchenko, Y.O., Conroy, B.R., Gobbini, M.I., et al., 2011. A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron 72, 404–416. doi:10.1016/j.neuron.2011.08.026.

Horikawa, T., Kamitani, Y., 2022. Attention modulates neural representation to render reconstructions according to subjective appearance. Commun. Biol. 5, 1–12. doi:10.1038/s42003-021-02975-5.

Horikawa, T., Kamitani, Y., 2017. Generic decoding of seen and imagined objects using hierarchical visual features. Nat. Commun. 8, 15037. doi:10.1038/ncomms15037.

Hsu, A., Borst, A., Theunissen, F.E., 2004. Quantifying variability in neural responses and its application for the validation of model predictions. Netw.: Comput. Neural Syst. 15, 91–109. doi:10.1088/0954-898X_15_2_002.

Hubel, D.H., Wiesel, T.N., 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol. 160, 106–154. doi:10.1113/jphysiol.1962.sp006837.

Ince, R.A.A., Kay, J.W., Schyns, P.G., 2022. Within-participant statistics for cognitive science. Trends. Cogn. Sci. 26, 626–630. doi:10.1016/j.tics.2022.05.008.

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17, 825–841. doi:10.1006/nimg.2002.1132.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., et al., 2014. Caffe: Convolutional architecture For Fast Feature Embedding. arXiv:1408.5093 [Preprint] [cited 2021 Nov 8].

Kanwisher, N., McDermott, J., Chun, M.M., 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. 17, 4302–4311. doi:10.1523/JNEUROSCI.17-11-04302.1997.

Klein, A., Ghosh, S.S., Bao, F.S., Giard, J., Häme, Y., Stavsky, E., et al., 2017. Mindboggling morphometry of human brains. PLoS Comput. Biol. 13, e1005350. doi:10.1371/journal.pcbi.1005350.

Kourtzi, Z., Kanwisher, N., 2000. Cortical regions involved in perceiving object shape. J. Neurosci. 20, 3310–3318. doi:10.1523/JNEUROSCI.20-09-03310.2000.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process Syst. 25, 1106–1114. https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

Laumann, T.O., Gordon, E.M., Adeyemo, B., Snyder, A.Z., Joo, S.J., Chen, M.-Y., et al., 2015. Functional system and areal organization of a highly sampled individual human brain. Neuron 87, 657–670. doi:10.1016/j.neuron.2015.06.037.

Le, Q.V., Ngiam, J., Coates, A., Lahiri, A., Prochnow, B., Ng, A.Y., 2011. On optimization methods for deep learning. In: Proceedings of the 28th International Conference on International Conference on Machine Learning; June 2011, Bellevue, Washington, USA. Omnipress, pp. 265–272.

Lescroart, M.D., Gallant, J.L., 2019. Human scene-selective areas represent 3D configurations of surfaces. Neuron 101, 178–192. doi:10.1016/j.neuron.2018.11.004.

Li, D., Du, C., Wang, S., Wang, H., He, H, 2021. Multi-subject data augmentation for target subject semantic decoding with deep multi-view adversarial learning. Inf. Sci. (Ny) 547, 1025–1044. doi:10.1016/j.ins.2020.09.012.

Liu, D.C., Nocedal, J., 1989. On the limited memory BFGS method for large scale optimization. Math. Program 45, 503–528. doi:10.1007/BF01589116.

Mahendran, A., Vedaldi, A., 2015. Understanding deep image representations by inverting them. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; June 2015, Boston, MA, USA. IEEE, pp. 5188–5196. doi:10.1109/CVPR.2015.7299155.

Mishkin, M., Ungerleider, L.G., 1982. Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. Behav. Brain Res. 6, 57–77. doi:10.1016/0166-4328(82)90081-X.

Nastase, S.A., Gazzola, V., Hasson, U., Keysers, C., 2019. Measuring shared responses across subjects using intersubject correlation. Soc. Cogn. Affect. Neurosci. 14, 667–685. doi:10.1093/scan/nsz037.

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., Clune, J., 2016. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. Adv. Neural Inf. Process Syst. 29, 3387–3395. Available from https://arxiv.org/abs/1605.09304 .

Nonaka, S., Majima, K., Aoki, S.C., Kamitani, Y., 2021. Brain hierarchy score: which deep neural networks are hierarchically brain-like? iScience 24. doi:10.1016/j.isci.2021.103013.

Power, J.D., Mitra, A., Laumann, T.O., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., 2013. Methods to detect, characterize, and remove motion artifact in resting state fMRI. Neuroimage 84, 320–341. doi:10.1016/j.neuroimage.2013.08.048.

Schönemann, P.H., 1966. A generalized solution of the orthogonal procrustes problem. Psychometrika 31, 1–10. doi:10.1007/BF02289451.

Sereno, M.I., Dale, A.M., Reppas, J.B., Kwong, K.K., Belliveau, J.W., Brady, T.J., et al., 1995. Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. Science 268, 889–893. doi:10.1126/science.7754376.

Shen, G., Dwivedi, K., Majima, K., Horikawa, T., Kamitani, Y., 2019a. End-to-end deep image reconstruction from human brain activity. Front. Comput. Neurosci 13, 21. doi:10.3389/fncom.2019.00021.

Shen, G., Horikawa, T., Majima, K., Kamitani, Y., 2019b. Deep image reconstruction from human brain activity. PLoS Comput. Biol 15, 1006633. doi:10.1371/journal.pcbi.1006633.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks For Large-Scale Image Recognition. arXiv:1409.1556v1 [Preprint][cited 2021 Nov 8].

Smith, P.L., Little, D.R., 2018. Small is beautiful: in defense of the small-N design. Psychon. Bull. Rev. 25, 2083–2101. doi:10.3758/s13423-018-1451-8.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., et al., 2010. N4ITK: improved N3 bias correction. IEEE. Trans. Med. Imaging 29, 1310–1320. doi:10.1109/TMI.2010.2046908.

Van Essen, D.C., 2005. A population-average, landmark- and surface-based (PALS) atlas of human cerebral cortex. Neuroimage 28, 635–662. doi:10.1016/j.neuroimage.2005.06.058.

Van Essen, D.C., 2004. Surface-based approaches to spatial localization and registration in primate cerebral cortex. Neuroimage 23, S97–107. doi:10.1016/j.neuroimage.2004.07.024.

Van Uden, C.E., Nastase, S.A., Connolly, A.C., Ma, F.L., Hansen, I., Gobbini, M.I., et al., 2018. Modeling semantic encoding in a common neural representational space. Front. Neurosci. 12, 437. doi:10.3389/fnins.2018.00437.

Watson, J.D.G., Myers, R., Frackowiak, R.S.J., Hajnal, J.V., Woods, R.P., Mazziotta, J.C., et al., 1993. Area V5 of the human brain: evidence from a combined study using positron emission tomography and magnetic resonance imaging. Cereb. Cortex 3, 79–94. doi:10.1093/cercor/3.2.79.

Yamada, K., Miyawaki, Y., Kamitani, Y., 2015. Inter-subject neural code converter for visual image representation. Neuroimage 113, 289–297. doi:10.1016/j.neuroimage.2015.03.059.

Yamada, K., Miyawaki, Y., Kamitani, Y., 2011. Neural Code Converter for Visual Image Representation. In: International Workshop on Pattern Recognition in NeuroImaging, Seoul, South Korea. IEEE, pp. 37–40. doi:10.1109/PRNI.2011.13 May.

Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., DiCarlo, J.J., 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. USA. 111, 8619–8624. doi:10.1073/pnas.1403112111.

Zhang, Y., Brady, M., Smith, S, 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE. Trans. Med. Imaging 20, 45–57. doi:10.1109/42.906424.