Inter-individual deep image reconstruction

Jun Kai Ho[a¶], Tomoyasu Horikawa[b¶], Kei Majima[a], 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

*Corresponding author: Yukiyasu Kamitani

E-mail: kamitani@i.kyoto-u.ac.jp

¶These authors contributed equally to this work.

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. Though 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 evaluated machine learning models called neural code converters that predict a target subject's brain activity pattern from a source subject given the same stimulus by decoding hierarchical visual features and reconstructing perceived images. The training data consisted of fMRI data obtained with identical sets of natural images presented to pairs of individuals. The converters were trained 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 brain 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 when relatively small numbers of data were used for converter training. The conversion also allowed pooling data across multiple individuals for decoder training, leading to a slight improvement over the decoders trained on a single individual. These results demonstrate that the conversion learns hierarchical correspondence and preserves the fine-grained representations of visual features, enabling visual image reconstruction using decoders trained on other individuals.

Highlights

- The neural code converter converts the brain activity patterns across individuals with moderate conversion accuracy and learns reasonable visual area correspondences at the same level between individuals.
- The converted brain activity patterns can be decoded into hierarchical DNN features to reconstruct visual images, even though the converter is trained with a small number of data.
- The information of hierarchical and fine-scale visual features that enable visual image reconstruction are preserved after conversion.

Keywords

Visual image reconstruction; decoding; visual hierarchy; functional alignment; fMRI

1. Introduction

Sensory information is generally thought to be processed through a hierarchical pathway that detects topographically organized local simple features in the early stages and then progressively global complex features in the later stages, leading to holistic perception. In the ventral visual pathway, a stimulus is first 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.

The recent advances of deep neural networks (DNNs) have enabled detailed analyses of hierarchical feature representations across 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 visual cortices process increasingly complex visual features along the ventral neural pathway, similar to how DNNs process image features. Furthermore, perceptual content encoded in brain responses have been successfully reconstructed as images via DNN-based reconstruction algorithms (Shen, Horikawa et al., 2019; Shen, Dwivedi et al., 2019). The deep image reconstruction (Shen, Horikawa et al., 2019) consists of two steps: 1) DNN feature decoders are trained to predict the DNN features of an image from the brain activity pattern of a subject given that image as a stimulus; 2) an initial image is iteratively optimized such that its DNN features resemble the predicted DNN features. However, these predictive models require optimization by machine learning algorithms using training data derived from hours of experiments that measure the brain responses to hundreds or thousands of images. Furthermore, in most cases, a model trained on one subject does not generalize to test data from 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 compensate for individual differences. Human brain anatomy differs across individuals in terms of shape, size, and local anatomical landmarks. Furthermore, functional brain area parcellation that clusters voxels/vertices with similar properties produces similar brain areas on the individual level but also exhibits distinct topological features (Blumensath et al., 2013; Laumann et al., 2015). The retinotopic organization that shows a correspondence between the visual field and VC was used to delineate functional visual areas (Engel et al., 1994; Sereno et al., 1995). The delineated visual areas are often similar but not the same across individuals. Anatomical alignment could mitigate the anatomical difference by matching the 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., 2015; Chen et al., 2015; Bilenko and Gallant, 2016; Guntupalli et al., 2016). Methodologies of functional alignment can be categorized into two types: (a) direct linear mapping between the voxels of two subjects, for example, a neural code converter (Yamada et al., 2015); (b) a shared latent space among subjects is constructed, and mapping between subjects is mediated by the shared space, for example, hyperalignment (Haxby et al., 2011). Those functional alignment methods have uncovered common neural representations across individuals concealed under substantial individual variations in brain responses. However, those 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 a model trained on other individuals. Furthermore, the 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 whether and how fine-grained neural representations of hierarchical visual features can be converted between individuals while preserving the encoded perceptual content. For this purpose, we evaluate a functional alignment method (the neural code converter; Yamada et al., 2015) using the decoding of hierarchical DNN features (Horikawa and Kamitani, 2017) and reconstruction of perceived images (deep image reconstruction; Shen et al., 2019). We also evaluated the method of hyperalignment (Haxby et al., 2011), but we mainly discuss the results obtained using the neural code converters because the methods mediated by latent space (hyperalignment) make it difficult to interpret the correspondence of visual subareas between subjects. We first construct 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 train 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 pre-trained converter transforms the brain activity into the target subject's brain space (Fig. 1B). The converted brain activity is analyzed 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 accurately convert the fMRI responses of a source subject to those of a target subject. Notably, while the converter training for target voxels is performed with the voxels in the entire VC of the source subject without explicit information about cortical hierarchy, the converters can automatically learn the hierarchical correspondence of visual subareas at the same levels for the source and target subjects. Furthermore, 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 even after conversion. We show that image reconstruction using decoded features originating from the source subject's brain via the conversion produces faithful reconstructions of the viewed images even with small numbers of data for converter training. We also demonstrate that the converter trained with sufficient training data but without the information about cortical hierarchy shows no distinct difference from the converter trained with that information. Lastly, we pool data from

multiple subjects by neural code conversions and show that DNN feature decoders trained on the pooled data achieve a slight improvement in decoding accuracy. These results demonstrate that the neural code converter can automatically find hierarchical correspondence and preserve the fine-grained representations of visual features across individuals, providing an efficient way to create visual image reconstructions even for novel individuals

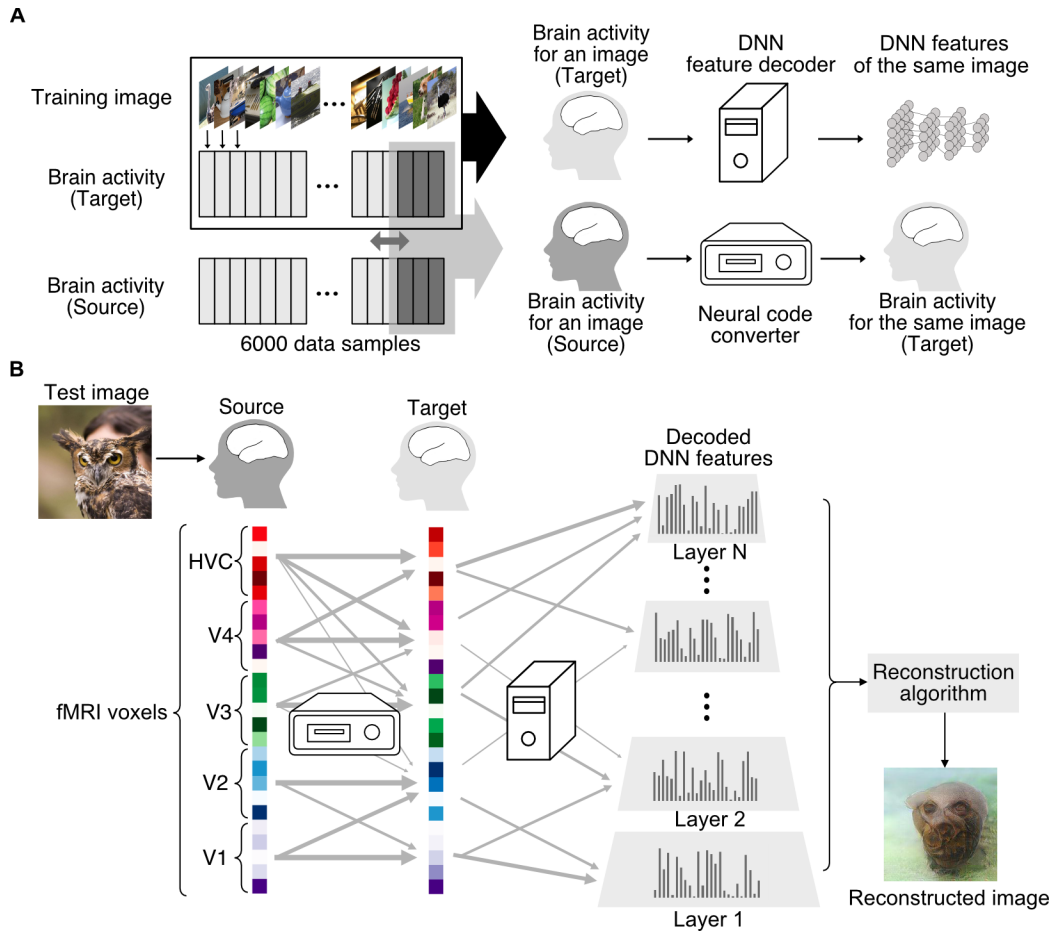

Fig. 1. Inter-individual deep image reconstruction.

(A) Training of the DNN feature decoder and neural code converter. DNN feature decoding models were trained on the 6,000 samples of measured fMRI activity of the target subject and the corresponding DNN features. A converter model is 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 image reconstruction. The converter model converts the source subject's stimulus-induced fMRI pattern to the target subject's brain space. The converted fMRI pattern is then decoded (or translated) into DNN feature patterns using the feature decoder. Finally, the decoded features are fed to a reconstruction algorithm to reconstruct the stimulus image perceived by the source subject.

2. Results

2.1 Neural code converter

We analyzed fMRI data of a total of five subjects, which consisted of five subjects of previously published data (Shen et al., 2019; Horikawa and Kamitani, 2022) and collected additional data for the two subjects in Horikawa and Kamitani (2022, see Materials and Methods: "fMRI datasets and stimuli"). The dataset consisted of fMRI data measured when subjects viewed 1,200 natural images in the training natural-image session, 50 natural images in the test natural-image session, and 40 artificial images in the test artificial-image session. Each image was presented during an 8-s stimulus block during which four volumes were scanned. We averaged the four volume scans in a stimulus block to create a single data sample. The presentation of each image was repeated 5, 24, and 20 times for 1,200 natural training images, 50 natural test images, and 40 artificial test images, respectively, resulting in 6,000 training samples, 1,200 test samples with natural images, and 800 test samples with artificial images. We followed Shen et al. (2019) and trained the DNN feature decoders using all 6,000 training samples in each of the five subjects. Note that the fMRI data collection for all training samples required approximately 800 minutes for each subject. In the current study, functional alignment models were trained using a varying subset (300, 600, 900, 1,200, 2,400, 3,600, 4,800, or 6,000 samples) of the same training samples. Unless otherwise noted, we show the results obtained using 2,400 training samples (two repetitions) as a representative case. The models were then tested using the data from the test natural-image session and the test artificial-image session.

Using these datasets, we first performed an analysis using the neural code converter to examine whether and how the visual information of one subject can be converted to that of another subject. In this analysis, we constructed a neural code converter model between each pair of subjects, using one as the target subject and the other as the source subject, which resulted in 20 subject pairs. A converter model comprises a set of regularized linear regression models (ridge regression), each trained to predict the activity of each voxel of a target subject's brain from a source subject's brain activity patterns in the VC (see Materials and Methods: "Neural code converter analysis"). Although we mainly present the neural code converter analysis in this study, we also used hyperalignment analysis for comparison. In contrast to neural code converter analysis, hyperalignment estimates a shared space among subjects by iteratively performing a

series of Procrustean transformations; the one-to-one mapping between subjects can then be constructed via the shared space (Fig. S1A; see Materials and Methods: “Hyperalignment analysis”). As for the converter analysis, we performed hyperalignment with only two subjects each time. The models were trained with 2,400 training samples and used to convert 1,200 test samples of a source subject’s brain activity patterns corresponding to the 50 natural test images into a target subject’s brain space or a shared space.

The VC consists of V1–V4, delineated by retinotopic analysis and ventral object-responsive areas from the left and right hemispheres. The ventral object-responsive areas include the lateral occipital complex (LOC; Kourtzi and Kanwisher, 2000), fusiform face area (FFA; Kanwisher et al., 1997), and parahippocampal place area (PPA; Epstein and Kanwisher, 1998). Each of them was obtained by localizer experiments (see Materials and Methods: “Regions of interest”). We defined the continuous region covering the LOC, FFA, and PPA as the HVC. In the converter and hyperalignment analysis, the left and right VCs were separately analyzed. All voxels were used as input without additional voxel selection (see Materials and Methods: “Neural code converter analysis”).

We evaluated the models using two methods: (a) pattern correlation, which is the spatial Pearson correlation coefficient between the converted and measured voxel patterns for each test image, and (b) profile correlation, which is the Pearson correlation coefficient between the sequences of converted and measured individual voxel responses to the 50 natural test images (Fig. 2A). Fifty different images were presented in each of the 24 runs of the test natural-image session, resulting in 24 samples for each of the 50 test images. Thus, the pattern correlation for each image was defined by the mean of 24 samples (converted) × 24 samples (measured) = 576 correlation coefficients. The profile correlation for each voxel was defined by the mean of 24 runs (converted) × 24 runs (measured) = 576 correlation coefficients. The obtained correlation coefficients were normalized by their noise ceilings (Hsu et al., 2004; Lescroart and Gallant, 2019). Each noise ceiling was estimated by the mean pattern or profile correlation coefficient between the measured voxel patterns or individual voxel responses of target subjects (see Materials and Methods: “Noise ceiling estimation”). To summarize the results, the correlation

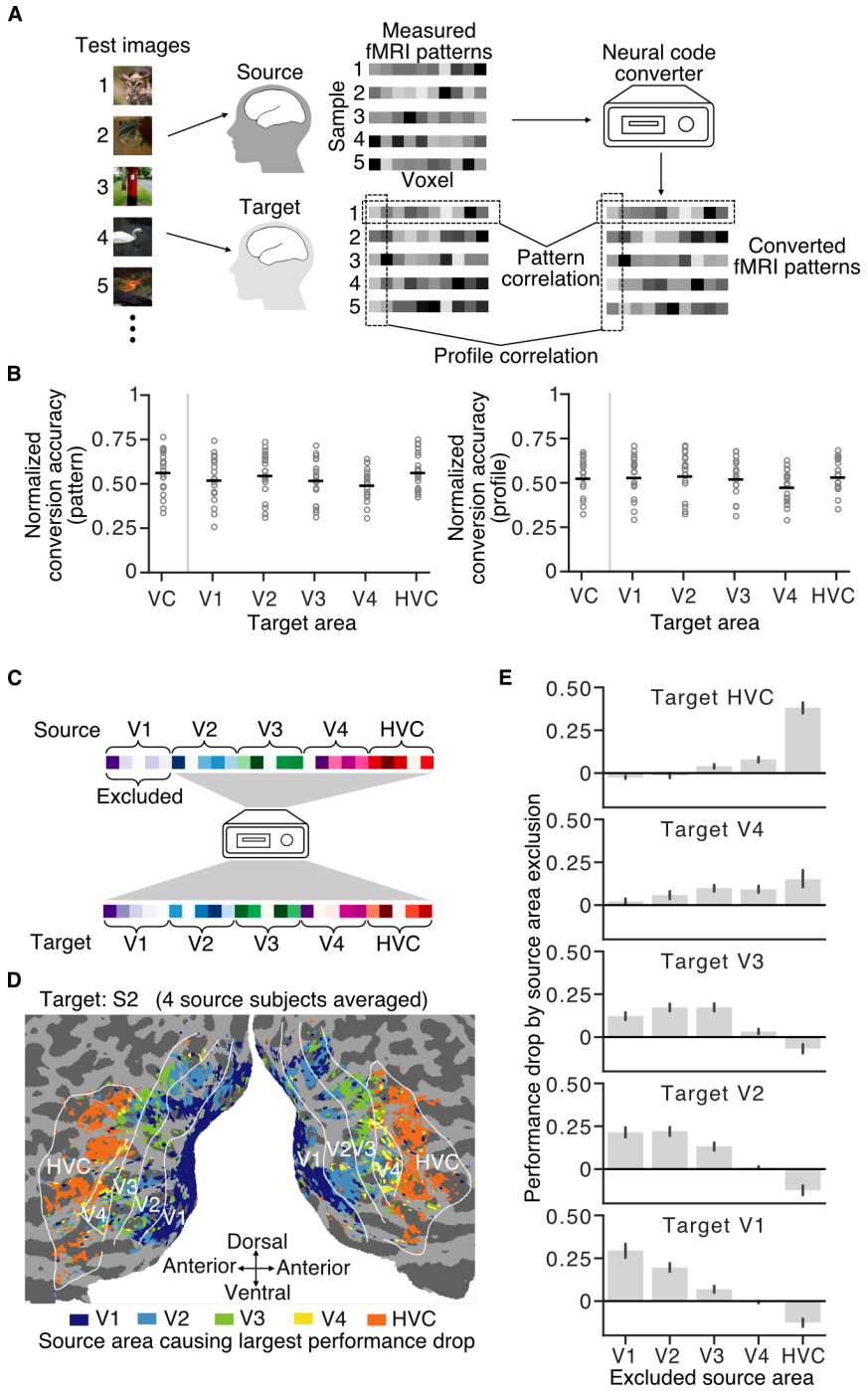

Fig. 2. Performance of neural code converters and cortical hierarchical correspondence. (A) Evaluations of neural code converters. Two evaluations were performed by computing the Pearson correlation coefficients: pattern and profile correlation.

(B) Conversion accuracy. Distributions of the normalized pattern or profile correlation coefficients are shown for the VC and visual subareas. Each horizontal black dash indicates the mean value; each circle represents the correlation coefficients of a subject 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 (S1). Each voxel on the target brain is colored by the index of the excluded visual area that caused the largest performance drop when the converter models were tested with the natural image test dataset (performance drops were averaged across four source subjects for a single target subject; see Fig. S2 for other target subjects). Only voxels that generate reliable responses are shown.
(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 subject pairs; error bars, 95% confidence interval [C.I.] across 20 subject pairs).

coefficients were averaged across images and voxels for the pattern and the profile correlations, respectively, in each subject pair for conversion.

In Fig. 2B, the distributions of the normalized pattern and profile correlation coefficients across all conversion pairs and images/voxels are shown for different regions of interest (ROIs). The mean normalized pattern correlation for the conversion with the whole VC was 0.56 ± 0.12 (mean ± standard deviation) over subject pairs, with the visual subareas showing comparable distributions (Fig. 2B left). The mean normalized profile correlation for VC was 0.53 ± 0.10 over 20 subject pairs (Fig. 2B right). The subareas also yielded distributions similar to those of the VC. The conversion accuracies were modest in both pattern and profile correlations across all visual subareas but comparable to those of the previous study (Yamada et al., 2015). The hyperalignment-based methods (conversion via shared space, VC, 0.55 ± 0.19 for pattern correlation and 0.52 ± 0.16 for profile correlation; see Fig. S1B and S1C) showed no clear advantage over the neural code converter.

To see how the visual areas of the source subjects contributed to the conversion accuracy for each voxel in each target visual area, we performed the same converter analysis while excluding one of the source visual subareas (V1, V2, V3, V4, or HVC) from the inputs of the converter

models (Fig. 2C). We evaluated the drop in performance (normalized profile correlation difference) with respect to the performance of the condition using all source visual subareas (*i.e.*, the whole VC). This ablation analysis revealed that the effects of the source area exclusions varied according to the location of the voxels in the target brain. The largest drop in performance of each target voxel was often caused by the exclusion of the corresponding source area (Fig. 2D, target S2; see Fig. S2 for other target subjects). On average, the maxima of the mean performance drop across voxels for target visual areas V1, V2, and HVC was observed when the corresponding source areas were excluded in the analysis, with the peaks shifted from lower to higher excluded source areas for lower to higher target areas (Fig. 2E). 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 information about the correspondences between them.

2.2 DNN feature decoding

Next, to examine whether fine-grained representations of visual features were preserved in the converted fMRI activity patterns, we performed DNN feature decoding analysis (Horikawa and Kamitani, 2017). We used the VGG19 DNN model, which consists of sixteen convolutional layers and three fully connected layers (Simonyan and Zisserman, 2014; see Materials and Methods: "DNN features"). Feature decoders were trained to predict the DNN features of the stimuli using 6,000 training samples of a target subject's fMRI activity patterns in either the whole VC or each visual subareas (V1–V4, HVC). The trained feature decoders were then tested using a source subject's fMRI activity patterns converted into the target subject brain space ("Across - functional [converter]" condition; see Materials and Methods: "DNN feature decoding analysis"). Following the original paper, we averaged the converted fMRI activity patterns over repetitions for each test image to enhance the signal-to-noise ratio. For each DNN unit, a feature decoder takes an averaged fMRI activity pattern as input and predicts a feature value. The decoding accuracy of each DNN unit was evaluated by a Pearson correlation coefficient between sequences of the decoded feature values and image feature values computed from the presented images (the natural test images). We further took the mean decoding accuracy over all DNN units in each layer.

For comparison, we performed DNN feature prediction with anatomically aligned brain activity. The source subject's fMRI images were aligned to the target's anatomical template, and DNN features were predicted using the decoders trained on the target subjects' data ("Across - anatomical"; see Materials and Methods: "Anatomical alignment" condition). We also compared the results with the within-individual decoding, in which DNN features were predicted using the decoders trained on the same subject's data ("Within" condition).

We first evaluated feature decoding performance obtained from the whole VC patterns of the converted fMRI activity. The prediction accuracy obtained from the neural code converter ("Across - functional") gradually increases from lower to higher layers, showing a similar trend to the within individual results ("Within"; Fig. 3A). The inter-individual decoding via the anatomical alignment ("Across - anatomical") performed the worst among the three conditions, with accuracies below 0.1 in most layers. The results showed that the neural code converters have an advantage over anatomical alignment in DNN feature decoding. Additional analyses showed that the neural code converter also outperformed the hyperalignment-based methods in DNN feature decoding (Fig. S1D).

We next performed feature decoding from individual visual areas (V1–V4 and HVC) and identified the visual area, called the "top" area, which gave the highest decoding accuracy, as in Nonaka et al. (2021). We then computed the distribution of the "top" area across DNN units in a given layer. We observed the shift of the peak area, from lower to higher areas, along the DNN hierarchy in all conditions (Fig. 3B). 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 correlations between the hierarchical levels across DNN units (Fig. 3C; see Materials and Methods: "BH score"). The results of the within individual condition ("Within") replicate the previous findings with a BH score around 0.5 (Horikawa and Kamitani, 2017; Nonaka et al., 2021). Although the anatomical alignment ("Across - anatomical") showed low overall accuracies in feature decoding (Fig. 3A), the hierarchical correspondence is largely preserved when quantified by the top decoding area for each DNN unit, presumably because anatomical alignment still roughly aligned the visual subareas therefore the cortical hierarchy information was inherited, even though the statistical

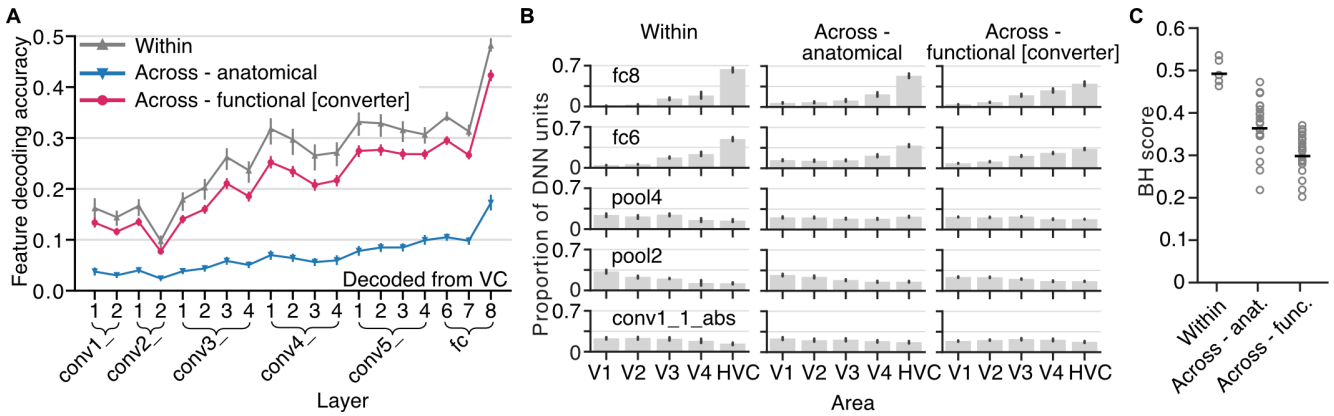

Fig. 3. DNN feature decoding and hierarchical representation by DNN.

(A) DNN feature decoding accuracy from the whole visual cortex (VC). Decoding accuracies for "Within," "Across - functional (converter)," and "Across - anatomical" decoding analyses from visual cortices are shown for each layer of the VGG19 model (error bars, 95% C.I. across five subjects for the "Within" condition, and 20 subject pairs for the "Across - functional [converter]" and "Across - anatomical" conditions).

(B) Proportion of DNN units for each "top" visual area, which is the best decodable visual area for a DNN unit. 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 subject pairs for the "Across - functional (converter)" and "Across - anatomical" conditions (error bars, 95% C.I. across five subjects for the "Within" condition, and 20 subject pairs for the "Across - functional [converter]" and "Across - anatomical" conditions).

(C) Brain hierarchy (BH) score. Each horizontal black dash indicates the mean BH score over 20 subject pairs; each circle represents the BH score for a subject or a subject pair.

relationships between voxels were not established. The inter-individual conversion ("Across – functional") showed a lower BH score than the other two conditions but still showed a substantial degree of hierarchical correspondence, even though the converter was blind to anatomical information during the training.

2.3 Visual image reconstruction

We have so far confirmed that multiple levels of DNN feature representations can be decoded from the converted brain activity patterns. We next tested to determine if we could reconstruct

visual images via DNN features decoded from the converted brain activity patterns using the image reconstruction method that renders hierarchical DNN features into visual images (deep image reconstruction, Shen et al., 2019). The visual image was reconstructed by iteratively optimizing an initial image such that the DNN features of the reconstructed image were as close as possible to the DNN features decoded from the whole visual cortical fMRI activity patterns (see Materials and Methods: "Visual image reconstruction analysis"). We performed the reconstruction analysis using two sets of visual images: natural images and artificial images. The natural image set consists of 50 images taken from the ImageNet database (Deng et al., 2009), and the artificial image set consists of 40 combinations of five shapes and eight colors (see Materials and Methods: "fMRI datasets and stimuli" for details of the images).

We compared the visual images reconstructed using the DNN features decoded from the VC in the "Within," "Across - functional (converter)," and "Across - anatomical" conditions. The reconstructed images obtained in the "Within" and "Across - functional (converter)" conditions captured the characteristics of the presented images, including the shapes and colors of the objects (Fig. 4A; see Fig. S3 and Fig. S4 for other examples of natural images and artificial images). Here, we only showed the images reconstructed from the average fMRI activity patterns across 24 repetitions; the images reconstructed from the converted fMRI activity patterns averaged across different numbers of repetitions are available in the supplemental information (Fig. S5). The reconstructed images obtained in the "Across - functional (converter)" condition showed comparatively good visual quality with respect to those obtained in the "Within" condition in terms of the shapes and colors of the objects. Reconstructed images obtained in the "Across - anatomical" condition showed neither a recognizable shape nor color of the objects in the presented images.

To quantitatively evaluate the quality of the reconstructed images, we performed a pairwise identification analysis based on image pixel values and DNN feature values of the presented and reconstructed images. In this analysis, we used the AlexNet model (Krizhevsky et al., 2012) to extract the DNN features of the presented and reconstructed images; correlation coefficients that were calculated from image pixel values or DNN feature values were used to identify a reconstructed image from a pair of images (a true [presented] image and a false image; see

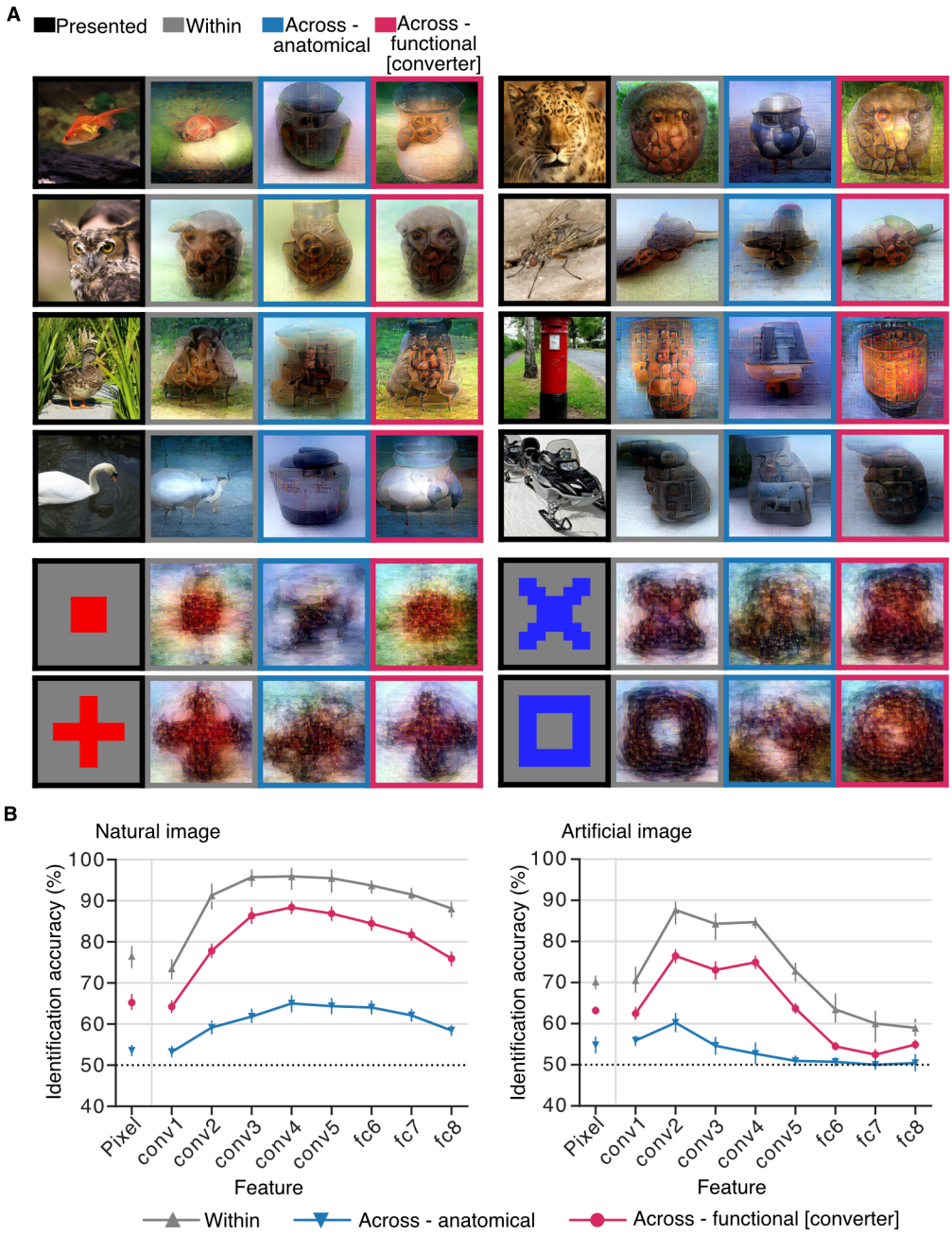

Fig. 4. Reconstructed images and evaluations.

(A) Within and across individual reconstructions from the whole visual cortex (VC). Images reconstructed under the three analytical conditions were all from the same source subject for each presented image. We show the reconstructed images produced from different source subjects for different presented images to avoid bias to any specific subject.

(B) Identification accuracy 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. across five subjects for the "Within" condition and across 20 pairs for the "Across - functional [converter]" and "Across - anatomical" conditions; dotted lines, chance level, 50%).

Materials and Methods: "Identification analysis"). The reconstructed image is correctly identified if its correlation with the true image is larger. The identification procedure was performed for each reconstructed image, and a mean identification accuracy was calculated over all reconstructed images for each subject. The term "identification accuracy" indicates the mean identification accuracy. The identification accuracies based on the pixel values and DNN feature values measure the similarity between the reconstructed and presented images as well as the discriminability between the reconstructed image and other (false) images at multiple levels of visual information. In the evaluations of the reconstructed natural or artificial images, the "Within" condition obtained the best results with the highest accuracies across all features. The "Across - functional (converter)" condition yielded lower accuracy than the "Within" condition but higher accuracy than the "Across - anatomical" condition (Fig. 4B). This result suggests that the reconstructed images contain multiple levels of visual information.

To examine whether explicit information about cortical hierarchy could improve visual image reconstruction, we performed subarea-wise conversion that predicts the activity values of a voxel in a target area only from the source subject's corresponding source area (Fig. 5A, see Fig. S6 for conversion accuracy). Using the converter trained in the subarea-wise manner, we performed the DNN feature decoding and the image reconstruction using the same methods used for the whole VC analysis (Fig. 3 and 4). We compared the results with those obtained from whole VC conversion, which predicts the activity values of a target voxel from the whole VC activity patterns of the source subjects.

In the DNN feature decoding of the natural test images, the whole VC condition showed similar but slightly higher decoding accuracy than the subarea-wise condition across layers (Fig. 5B; analysis of variance [ANOVA], effect of conversion type, $F(1, 361) = 1959$, $p < .001$, $\eta_p^2 = .84$). Reconstructed images obtained from subarea-wise conversions had a visual quality similar to those of the whole VC conversions (Fig. 5C). In the identification analysis of the reconstructed

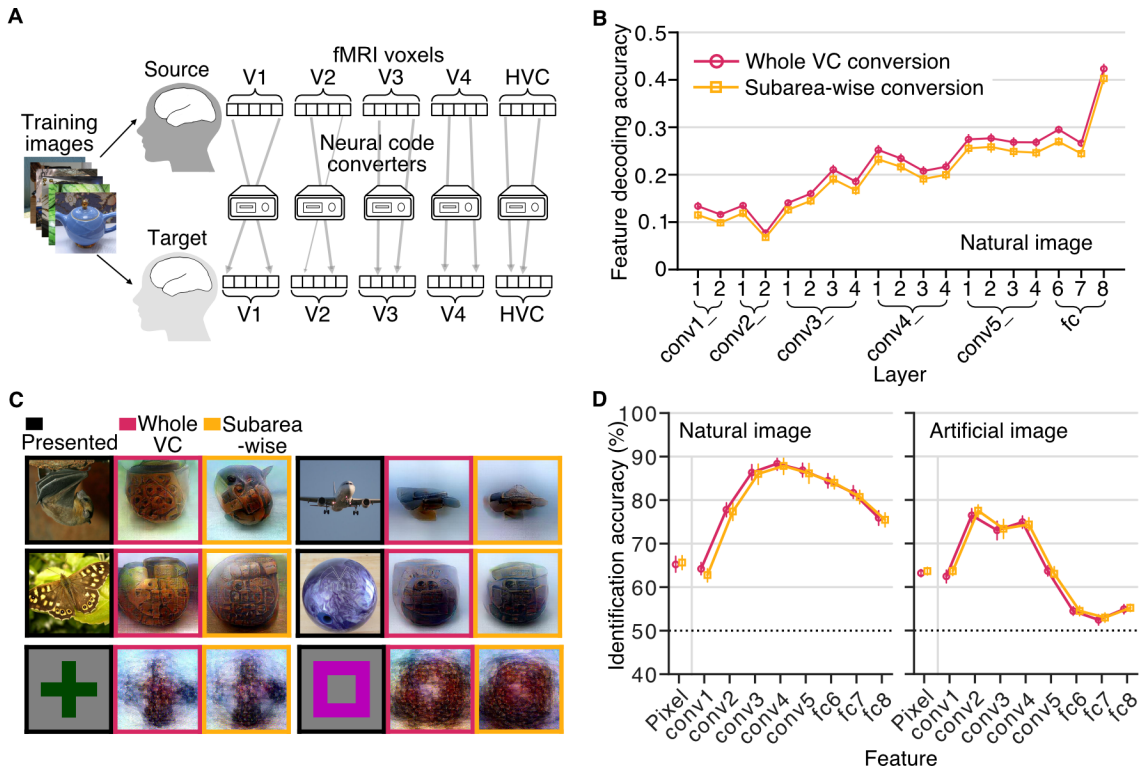

Fig. 5. Whole VC conversion and subarea-wise conversion.

(A) Illustration of the subarea-wise conversion. A converter model was trained with 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 the natural test image (error bars, 95% C.I. across 20 subject 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. across 20 subject pairs; dotted lines, chance level, 50%).

natural images and artificial images (*cf.*, Fig. 4), although the whole VC conversion slightly outperformed the subarea-wise condition (Fig. 5D), the effect size was very small (natural images, ANOVA, $F(1, 171) = 11.2$, $p < .001$, $\eta_p^2 = 0.062$; artificial images, ANOVA, $F(1, 171) =$

29.8, $p < .001$, $\eta_p^2 = 0.15$). Overall, the subarea-wise conversion did not show a consistent advantage over the whole VC conversion. Thus, the conversion with explicit information about the cortical hierarchy does not seem to significantly contribute to the improvement of visual image reconstruction.

One of the potential benefits of inter-individual analyses is to reduce the number of data required for model training from novel test (source) subjects by using training data from other individuals. Because 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 the 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 the available data of the target subjects for decoder training (6,000 samples). Note that it took 40 mins to collect 300 samples but 800 mins to collect all training samples. We then evaluated the reconstructed images by identification analysis.

While the visual quality of reconstructions proportionally decreased according to the decrement in the number of samples, the reconstructions obtained via converters trained with 300 samples still produced discernible natural images (Fig. 6A, see Fig. S7 for artificial images). The visual quality was somehow compromised with a small number of data but this result showed that performing image reconstruction using converters with a small number of training data is feasible, without the need for collecting a full set of fMRI data. The identification accuracies using pixel values and extracted DNN feature values increased with the number of training samples (Fig. 6B, natural images; see Fig. S7 for artificial images). The identification accuracy in the “Across - functional (converter)” condition was asymptotically closer to that in the “Within” condition but did not reach it even with the maximum number of samples tested here (6,000 samples). However, the “Across - functional (converter)” condition achieved higher accuracies than the “Across - anatomical” condition across all training sample conditions. Under the same number of training samples, the identification accuracies of the whole VC and subarea-wise conversions were similar when the number of training samples was more than 900. However, subarea-wise conversion outperformed whole VC conversion when the number of training samples was equal to or less than 900 (ANOVA with the type of conversion in each of

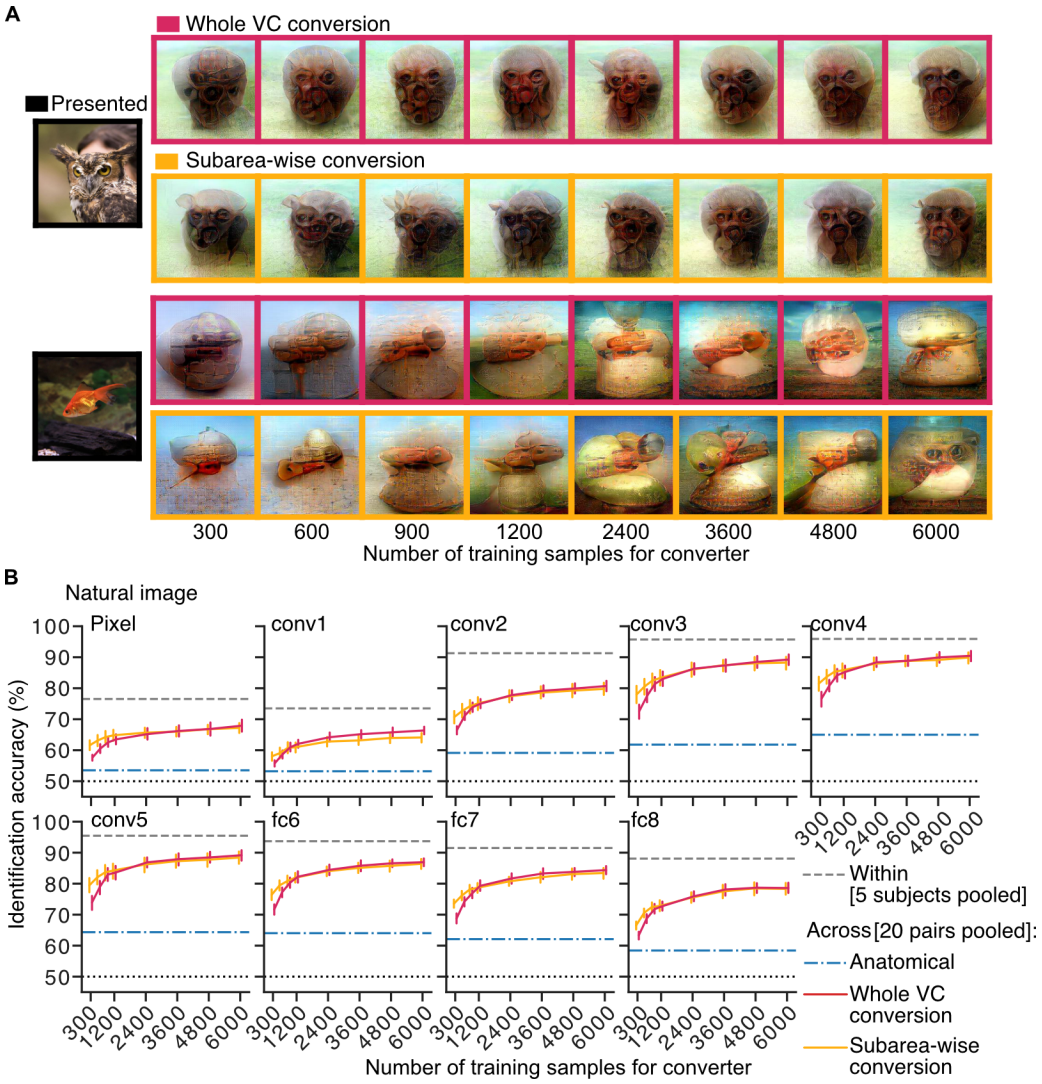

Fig. 6. Images reconstructed via converters trained on varying numbers of data.

(A) Images reconstructed via converters trained on varying numbers of data. All reconstructed images were produced from the same source subject (source: S2, target: S3)

(B) The effect of the number of training data on the identification accuracy. The converters were trained on a varying number of data. The identification analysis was performed using pixel values and extracted DNN feature values (AlexNet) of the reconstructed natural images (error bars, 95% C.I. across 20 subject pairs for whole VC and subarea-wise conversion conditions; dotted lines, chance level, 50%).

the eight training sample conditions [except for the 1,200-sample condition], $p < .001$, Bonferroni-corrected by eight). In the identification based on the conv1 feature values of the reconstructed natural images, the whole VC conversion with 4,800 training samples or more

showed higher accuracies than the subarea-wise conversion. Overall, the explicit information about cortical hierarchy does not generally improve reconstruction, but it may be beneficial when the number of training data is limited.

2.4 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 image reconstruction. This analysis left out one subject for testing and pooled the data from the other four subjects. As in our representative case, we took 2,400 training samples from the test subject. Within the four subjects, each subject was selected as the target subject once, and the other three subjects' training data were converted into the target subject's brain space (6,000 samples for each subject; whole VC conversion; Fig. 7A, see Materials and Methods: "Pooling data from multiple subjects" for details). Thus, there were 24,000 samples, including the original target subject data and the three subjects' converted data. We then trained DNN feature decoders on such pooled data and called the decoders "multiple-subject feature decoders," in contrast to the "single-subject feature decoders" trained with individual data in the native brain space. With 2,400 training samples of the test subject and pooled data from those four subjects, we trained a neural code converter to predict the fMRI activity patterns in the target brain space. We converted the test subject's data into the target brain space and then decoded the converted fMRI activity patterns into the DNN features and reconstructed visual images via the decoded DNN features. We compared the multiple-subject feature decoders with the single-subject feature decoders using inter-individual DNN feature decoding and visual image reconstruction.

The DNN feature decoding analysis on the natural images showed that the multiple-subject condition tended to achieve a small improvement in accuracy across all layers as compared with the single-subject condition (ANOVA, effect of decoder type, $F(1, 57) = 536$, $p < .001$, $\eta_p^2 = .90$; Fig. 7B for natural images; see Fig. S8 for artificial images). Reconstructed images obtained using both the single- and multiple-subject decoders showed recognizable visual quality, but the visual qualities were not substantially different (Fig. 7C). The pixel values and DNN features of the reconstructed natural images obtained in the multiple-subject condition showed slightly

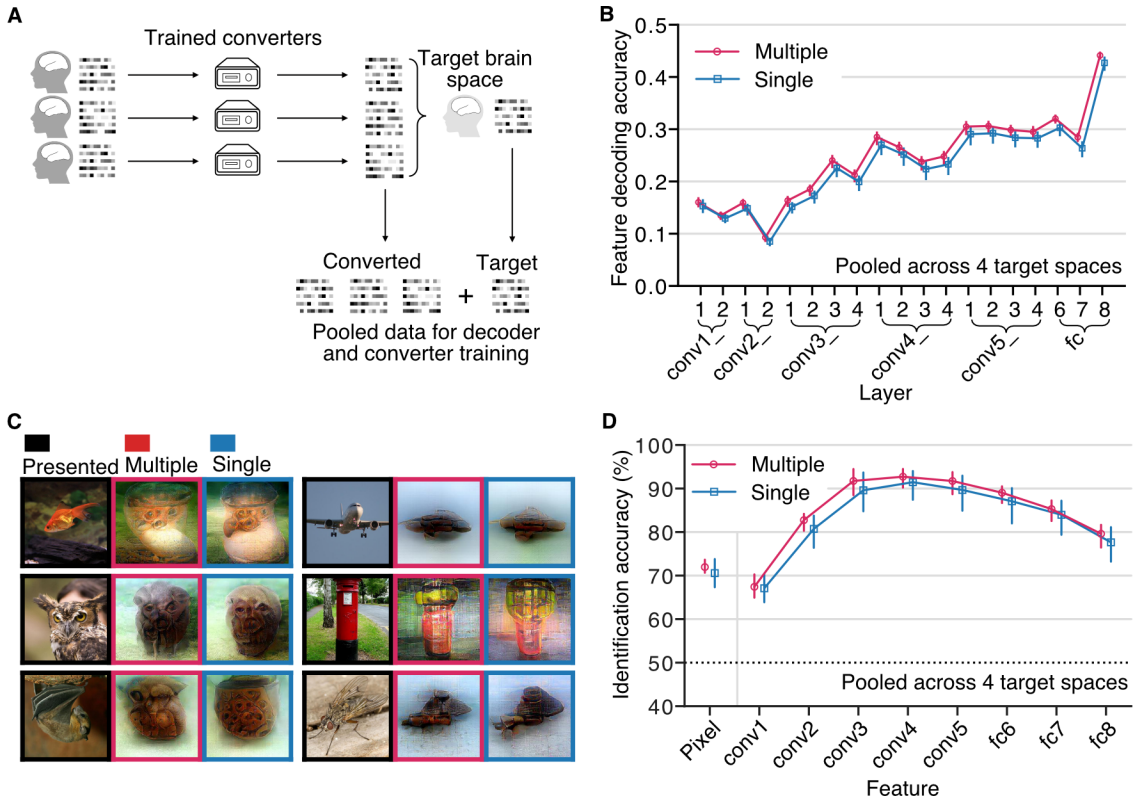

Fig. 7. Pooling data from multiple subjects.

(A) Illustration of the pooling procedure. Three subjects' training data were converted into one subject's brain space. The DNN feature decoders were trained with three subjects of converted data plus the target subject's data (24,000 samples). Among the four subjects (excluding the left-out test subject), each subject was selected as the target subject once, and hence there are four sets of pooled data in different brain spaces.

(B) DNN feature decoding accuracy obtained via multiple- and single-subject feature decoders. The multiple-subject feature decoders were trained on data pooled from four subjects; the single-subject feature decoders were trained on only one single subject's data. The accuracies were obtained from the test subject's test dataset of natural images (error bars, 95% C.I. across four target brain spaces).

(C) Reconstructed natural images obtained via multiple- and single-subject feature decoders.

(D) Identification accuracies obtained from the reconstructed natural images. The identification analysis was performed using pixel values and extracted DNN feature values of the reconstructed natural images obtained via multiple- and single-subject feature decoders (error bars, 95% C.I. across four target brain spaces; dotted lines, chance level, 50%).

higher identification accuracies than those obtained in the single-subject condition, as revealed by a similar ANOVA analysis (effect of decoder type, $F(1, 27) = 19.9$, $p < .001$, $\eta_p^2 = .42$; Fig. 7D for natural images; see Fig. S8 for artificial images). We also analyzed the case where there was only a limited number of data for the test subject (300 training samples) to reflect the circumstance that only a limited number of data can be collected for most people because of the high cost of collecting large numbers of fMRI data (see Fig. S9). These results indicate that pooling multiple-subject data is beneficial for slightly improving the accuracy of inter-individual image reconstruction, although the visual quality of the reconstructed images was not largely improved.

3. Discussion

This study investigated whether and how hierarchical and fine-grained visual information could be converted while preserving perceptual content across individuals using neural code converter analysis. We first showed that a neural code converter could predict a target subject's fMRI activity patterns from a source subject's fMRI activity patterns. The ablation analysis on the converters with the exclusion of voxels from various source visual subareas showed that the converters automatically detected the hierarchical correspondences of visual subareas between individuals. We decoded the converted brain activity patterns into DNN features and observed the correspondence between visual subareas and DNN layers even after the conversion. Visual images were reconstructed from the converted brain activity patterns with recognizable shapes and colors of the objects in the presented images. The conversions with and without explicit labels of visual areas (subarea-wise conversion and whole VC conversion) showed almost equivalent performances in the inter-individual visual image reconstruction, indicating that the whole VC conversion preserved 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 inter-individual 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 patterns, and neural code conversions provide an efficient way to reconstruct 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, like V1 or the inferior temporal cortex (Yamada et al., 2015; Haxby et al., 2011), whereas other studies targeted a large region including multiple levels of hierarchy (Bilenko and Gallant, 2016; Van Uden et al., 2018), but the hierarchical correspondences of visual subareas have not been discussed. Our results explicitly demonstrate that machine learning-based neural code converters can learn from fMRI activity patterns to detect the hierarchical correspondence of visual subareas between two individuals. However, some inter-regional predictions also existed (*e.g.*,

source V1 could predict voxel values in target V2, Fig. 2C). The strength of the inter-regional prediction decreases as the cortical distance between two areas increases (*e.g.*, there was no inter-regional prediction between V1 and HVC), presumably indicating that two areas close to each other 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 (Fig. 3 and 4), we showed that hierarchically organized fine-grained visual features that enable visual image reconstruction are preserved in neural code conversions. 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 that are sufficient to reconstruct visual images 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). Furthermore, visual image reconstruction via decoded DNN features successfully reconstructed the contours and colors of the objects in the presented images, confirming that the converted fMRI activity patterns preserved sufficient perceptual content to reconstruct visual images.

The subarea-wise conversion and whole VC conversion with sufficient training data showed almost equivalent performance in the inter-individual visual image reconstruction (Fig. 5 and 6). Moreover, the reconstructed images obtained via whole VC conversions tended to achieve slightly higher identification accuracy than those obtained via subarea-wise conversion (Fig. 6B, conv1). This result shows that subarea-wise conversion with explicit labels of visual subareas was not helpful in improving the reconstructed image quality. In other words, the whole VC conversion without explicit labels of the visual subareas enabled such information to be implicitly learned when sufficient training data are provided.

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 with 6,000 data samples, which took approximately 800 mins of data collection time. This is a long data

collection time in fMRI studies and is impractical for most people. The situation worsens when there are multiple decoding models trained on different datasets for other purposes. By trading off some visual quality of the reconstructed images, we could instead collect fewer data, for instance, 300 samples, to train a neural code converter and perform inter-individual visual image reconstruction. In particular, the neural code converter is thought to capture the relationships between the voxels of subjects under a variety of visual scenes, and it could potentially be used together with other decoding models. The inter-individual decoding method with the neural code converter could thus reduce the time and costs of fMRI data collection.

Our results show that pooling data from multiple subjects did not greatly improve the inter-individual image reconstruction (Fig. 7). One possible reason could be that the pooled fMRI data were the responses to the same set of the image dataset (1,200 natural images). It is well known in machine learning that training a model with more data can reduce the model's variance, but it might not reduce the model's bias if the training dataset is not diverse or truly representative. Another possible reason is that different subjects yield different data qualities, with some subjects' data leading to relatively poor visual quality in the reconstructed images. Data pooled with poorer quality data thus limited the decoder's ability to take advantage of the pooled data, resulting in a limited improvement in visual image reconstruction performance. Although pooling data did not lead to great improvements in visual image reconstruction quality, it is still a promising approach for future fMRI studies.

One observation was that the inter-individual image reconstruction did not outperform the within individual image reconstruction (Fig. 4 and 5). Because our converter models were based on linear regression models, the linear constraint used for conversion might have been too restrictive for learning more complex statistical relationships such as nonlinearity between brain activity patterns, thus limiting the performance of inter-individual image reconstructions. By contrast, the brain response to a stimulus can be treated as a sum of three mixtures: a consistent stimulus-evoked response across subjects, 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. Together with the noise components, the inter-individual image reconstruction thus underperformed on the within-individual images.

Our results showed that brain activity patterns can be translated across individuals while preserving sufficient information to visualize the perceived stimulus. This provides an efficient way to reconstruct visual images without training subject-specific models, especially when the reconstruction model is complex and data hungry. By reducing the need to collect a large number of data, our approach might prove useful in popularizing the use of brain-machine/computer interfaces that communicate with our internal world.

4. Materials and methods

4.1 Subjects

In addition to the three subjects in previous studies (Shen et al., 2019, see Materials and Methods: "fMRI datasets and stimuli"), two additional healthy subjects with normal or corrected-to-normal vision participated in our experiments: Subject 4 (male, age 22) and Subject 5 (male, age 27). For these two subjects, we reused a subset of previously published data for the training natural-image session (Horikawa and Kamitani, 2022) and additionally collected data for test natural-image and artificial-image sessions.

4.2 fMRI datasets and stimuli

We used the fMRI dataset originally collected by Shen et al. (2019) (https://openneuro.org/datasets/ds001506/versions/1.3.1) for Subjects 1–3 and the partial fMRI dataset collected by Horikawa and Kamitani (2022) (https://openneuro.org/datasets/ds003430/versions/1.2.0) for Subject 4 and 5. We collected additional data for test natural-image and artificial-image sessions for Subject 4–5 (https://openneuro.org/datasets/ds003993). These datasets collectively consist of fMRI responses to visual stimuli from a total of five subjects. The fMRI signals were measured while subjects viewed sequences of visual images (image presentation experiment). The image presentation experiment comprised a training natural-image session, test natural-image session, and test artificial-image session. The natural image stimuli 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 1,200 images were repeatedly presented five times in the training natural-image session, whereas the 50 images were repeatedly presented 24 times in the test natural-image session. The artificial image stimuli 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). The 40 artificial images were repeatedly presented 20 times in the test artificial-image session. Each presentation of an image lasted for 8 s in a stimulus block with four volume scans (Repetition time [TR] = 2 s). The Shen et al. (2019) dataset contains the data of the three image sessions for Subjects 1–3; the Horikawa and Kamitani (2022) dataset contains the data of the training natural-

image session for Subjects 4 and 5. We additionally collected the data of the test natural-image and artificial-image sessions for Subjects 4 and 5.

4.3 Experimental design

We conducted image presentation experiments consisting of test natural-image session and test artificial-image session for Subjects 4 and 5. The protocol is the same as the protocol used in Shen et al. (2019). Please refer to Shen et al. (2019) for details.

4.4 Image presentation experiment

The image presentation experiments consisted of two types of sessions: test natural-image session and test artificial-shape session. Each session consisted of 24 and 20 runs, respectively. Each run consisted of 55 and 44 stimulus blocks that last for 8 s, with these comprising 50 and 40 blocks of different images, and 5 and 4 randomly interspersed repetition blocks where the same image as in the previous block was presented. The 50 images were repeated 24 times in the test natural-image session and the 40 images were repeated 20 times in the test artificial-image session. The design is the same as the design used in Shen et al. (2019). Please refer to (2019) for details.

4.5 fMRI data preprocessing

The fMRI data were preprocessed through the fMRI data preprocessing pipeline provided by fMRIPrep (version 1.2.1; Esteban et al., 2019). A T1w reference image was generated through a custom pipeline that consisted of bias field correction, skull stripping, and brain tissue segmentation. The functional data of each run were motion corrected and slice-time corrected, followed by coregistration to T1w reference images in their native spaces. The coregistered data were then re-interpolated to 2 × 2 × 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 (training and test sessions, four volumes).

609

610 **4.6 Anatomical alignment**

611 We used fMRIPrep (Esteban et al., 2019) to nonlinearly normalize subjects' structural and
612 functional images to a standard space: MNI152NLin2009cAsym (MNI space). By default,
613 fMRIPrep uses Advanced Normalization Tools (ANTs, http://stnava.github.io/ANTs/), which
614 implements the symmetric image normalization method (Avants et al., 2008), to register images
615 to a template. Following standard fMRI data preprocessing (see Materials and Methods: "fMRI
616 data preprocessing"), the T1w reference image was further spatially normalized to MNI space by
617 the ANTs, and the functional data were coregistered to this normalized T1w reference image.
618 The coregistered data were then re-interpolated to 2 × 2 × 2 mm voxels. In the inter-individual
619 analysis, if a voxel of a source subject and a voxel of a target subject shared the same
620 coordinates, the fMRI activity of the source voxel was considered to be that of the corresponding
621 target voxel.

622

623 **4.7 ROIs**

624 For brains in native space, regions V1, V2, V3, and V4 were delineated using the standard
625 retinotopy experiment (Engel et al., 1994; Sereno et al., 1995). The HVC was defined as a
626 contiguous region covering the LOC, FFA, and PPA, which were identified using conventional
627 functional localizers (Kourtzi and Kanwisher, 2000; Kanwisher et al., 1997; Epstein and
628 Kanwisher, 1998). The VC was defined as the combined regions of V1, V2, V3, V4, and the
629 HVC.

630

631 For brains in MNI space, ANTs were used to normalize the ROI masks of V1, V2, V3, V4, and
632 HVC in their native space to the brain in MNI space. In the inter-individual analysis, a source
633 subject's voxels covered by an ROI mask of a target subject were selected as the input to the
634 model of the target subject.

635

636 **4.8 Neural code converter analysis**

637 The neural code converter model for each pair of subjects comprised a set of regularized linear
638 regression models (ridge regression), each trained to predict the activities of an individual voxel
639 of one subject (target) from the brain activity patterns of another subject (source) using fMRI

responses in the VC to the natural images. A converter takes the source subject's brain activity patterns $\mathbf{X}$ as input and predicts the target brain activity patterns $\hat{\mathbf{Y}} = \mathbf{W}\mathbf{X} + \mathbf{b}$, where $\mathbf{W}$ is the converter matrix and $\mathbf{b}$ is the bias vector. The converter minimizes the objective function $||\mathbf{Y} - \hat{\mathbf{Y}}||_2^2 + \lambda||\mathbf{W}||_2^2$, where $\mathbf{Y}$ is the true target subject's fMRI activity patterns, λ is the regularization parameter, and $||\cdot||$ represents the Frobenius norm.

Each hemisphere of the VC was used to perform the conversion across individuals. We performed a 5-fold cross-validation on the training data. The training dataset was split into five folds such that each fold of data contains fMRI responses to presented images different from others; four folds of data were used to train a converter, and the remaining fold of data was used as a validation dataset. We optimized the regularization parameter by first estimating the rough optimal value using six possible values (1, 10, 100, 1,000, 10,000, and 100,000); then further performed a grid search in a narrower range containing the roughly optimal values (e.g., 2, 4, 6, and 8 as well as 20, 40, 60, and 80 around the value of 10); the process was repeated a few times to refine the parameter down to unit digits. The regularization parameter was optimized for each source and target subject pair and was fixed across all voxels. Model performance was evaluated using the identification accuracy based on the decoded DNN feature values (see Materials and Methods: "Identification analysis"): we randomly selected 500 units (and the corresponding feature decoders of the target subjects) from each layer of the VGG19 model; we then decoded the converted fMRI activity patterns of the left-out validation dataset to the selected DNN features; the decoded DNN features of 19 layers were concatenated as a vector and the identification accuracy was calculated using the concatenated decoded DNN feature values; we chose the regularization parameter that maximized the averaged accuracy over five folds. This optimization procedure aims to convert brain activity patterns such that a better performance in the downstream task could be obtained. The main goal is to achieve a better performance in the visual image reconstruction (see Materials and Methods: "Visual image reconstruction"), but it is computationally expensive to reconstruct images for a grid search of regularization parameters, so we used a set of decoded DNN features to calculate the identification accuracy as an indirect performance measure of the visual image reconstruction, recalling that the DNN features of a reconstructed image approximate the decoded DNN features.

4.9 Hyperalignment analysis

Similar to neural code converter analysis, hyperalignment analysis requires the fMRI responses of subjects given an identical sequence of stimuli. We only briefly mention the hyperalignment algorithm; please refer to Haxby et al. (2011) for details. The hyperalignment iteratively estimates a shared space by first aligning two subjects' fMRI responses using a Procrustean transformation; if there is an additional subject, that subject's fMRI responses are aligned to the mean aligned responses of the subjects whose fMRI responses have been aligned; in the second iteration, each subject's response is aligned to the mean aligned responses obtained in the first iteration; the mean aligned response is recalculated and treated as a template; in the last step, each subject's response is aligned to the mean template and an orthogonal transformation matrix is obtained.

Although the hyperalignment algorithm can estimate the shared space of more than two subjects, we performed hyperalignment with only two subjects, as we did in the neural code converter analysis. Each hemisphere of the VC was analyzed separately. In addition to aligning subjects' fMRI responses to the shared space, we also performed another type of one-to-one conversion via the shared space: a subject's responses were first mapped into the shared space, followed by the inverse mapping into the other subject's brain space. We call this "conversion via shared space," in contrast to the "alignment into shared space," where all subjects' responses are brought into the shared space.

4.10 Noise ceiling estimation

Noise in the fMRI data inevitably lowers 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 correlation coefficients between measured brain responses as

$$\gamma = \frac{2}{N(N-1)} \sum_{i \neq j}^{N} \mathrm{corr}(\mathbf{y}_i, \mathbf{y}_j).$$

When the noise ceiling of brain activity patterns to an image is estimated, $\mathbf{y}$ is a brain activity pattern induced by the image and N is the total number of repeated presentations of the image. When the noise ceiling of an individual voxel's responses is estimated, $\mathbf{y}$ is the vector of the

voxel's response to a sequence of images, and *N* is the total number of runs. The noise ceilings were then used to normalize the raw prediction accuracies of the converter models.

Samples with corresponding noise ceilings below a threshold were excluded from the performance evaluation by pattern correlation because the conversion performance could not be reliably evaluated. Similarly, voxels with a profile noise ceiling below a threshold were excluded from the performance evaluation by profile correlation, but they were included in the downstream DNN feature decoding analysis to prevent information leakage. The thresholds were defined by the 99th percentile point in the distribution of the correlations from randomly paired samples or voxels.

4.11 DNN features

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. Outputs from individual units (before rectification) were used as target variables in the DNN feature decoding analysis. The outputs of all units in each layer were treated as a vector and used in the identification and reconstruction analyses. 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, 100352; 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. This model is pre-trained for the 1,000-class object recognition task with the images from ImageNet (Deng et al., 2009; the pre-trained model is 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.

732

733 **4.12 DNN feature decoding analysis**

734 For each unit of the DNN, we trained a ridge linear regression model (DNN feature decoder) that
735 takes an fMRI activity pattern induced by a presented image as input and outputs a feature value
736 computed from the presented image. The ridge regularization parameter was set to 100. The
737 feature value and the voxel values were normalized before training the model. For the decoder of
738 each DNN unit, a voxel selection procedure was performed before training, in which the Pearson
739 correlation coefficients between the sequences of feature values and voxel responses of all
740 voxels were computed and the top 500 voxels having the highest correlations were selected. The
741 training dataset with the selected voxels was then used to train a decoder to decode the values of
742 an individual unit in the feature vectors of all DNN layers. When testing the trained decoders on
743 the test datasets, the fMRI samples corresponding to the same stimulus were averaged across
744 trials to increase the signal-to-noise ratio of the fMRI signal. For details of the feature decoding,
745 see Horikawa and Kamitani (2017) or Shen et al. (2019).

746

747 **4.13 BH score**

748 The BH score was originally designed to measure the degree to which an artificial neural
749 network is hierarchically similar to the human brain. Here, we adopted the decoding-based BH
750 score; the DNN features are decoded from the fMRI signals in one of the five visual areas: V1–
751 V4 and the HVC. For each unit, the visual area showing the best decoding accuracy is identified
752 and is called the "top" area. Here, the decoding accuracy refers to the mean Pearson correlation
753 coefficient between the predicted and true DNN unit activations sequences over stimuli. The first
754 layer and the last layer are fixed, and three intermediate layers are randomly sampled from the
755 DNN. For each sampling, a Spearman rank correlation coefficient between the layer number and
756 the top area number (V1: 0, V2: 1, V3: 2, V4: 3, and HVC: 4) across units in the given DNN is
757 calculated. This sampling procedure is repeated 10,000 times, and the mean Spearman rank
758 correlation coefficient is taken as the BH score. In this study, the conv1_1_abs (the absolute
759 value was taken for the conv1_1), pool1–pool5, and fc6–fc8 layers of the VGG19 model were
760 used and 1,000 units were randomly selected from each layer to compute the BH score. For
761 details of the BH score, refer to Nonaka et al. (2021).

762

4.14 Visual image reconstruction analysis

We used an image reconstruction method (deep image reconstruction) proposed by Shen et al. (2019). 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 [37]:

$$\mathbf{v}^* = \underset{\mathbf{v}}{\mathbf{argmin}} \frac{1}{2} \sum_{l}^{L} \beta_l \|\mathbf{\Phi}_l(\mathbf{v}) - \mathbf{y}_l\|_2^2 ,$$

where $\mathbf{v} \in \mathbb{R}^{224 \times 224 \times 3}$ is a vector whose elements are the pixel values of an image (width $\times$ height $\times$ RGB channels); L is the total number of layers; $\mathbf{\Phi}_l$ is the function that maps the image to the DNN feature vector of the l-th layer; $\mathbf{y}_l$ is the decoded DNN feature vector of the l-th layer; and β_l is the parameter that weights the contribution of the l-th layer, which was set to be $1/\|\mathbf{y}_l\|_2^2$.

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 when reconstructing natural images (Nguyen et al., 2016). The optimization problem becomes

$$\mathbf{z}^* = \underset{\mathbf{z}}{\mathbf{argmin}} \frac{1}{2} \sum_{l}^{L} \beta_l \|\mathbf{\Phi}_l(G(\mathbf{z})) - \mathbf{y}_l\|_2^2 ,$$

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; trained model for fc7). Given decoded features from the 19 layers of the VGG19 model, the latent vector is initialized by a zero-value vector, and gradient descent with momentum algorithm optimizes the objective function (over 200 iterations).

The solution to the above optimization problem is considered to be the reconstructed image from the brain activity pattern. Following Shen et al. (2019), natural images were reconstructed using the DGN, and the objective function was optimized by stochastic gradient descent with momentum 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).

4.15 Identification analysis

Identification analysis was used to evaluate image reconstruction quality. Presented images were identified using either reconstructed image pixels or the DNN features extracted from reconstructed images. Before the identification analysis, the image pixels or DNN features were reshaped into a one-dimensional vector, which is called a feature vector for simplicity. The feature vector obtained from a reconstructed image was used to compare the true feature vector of the presented image with the false feature vector of another image. The comparison was counted as correctly identified if the Pearson correlation coefficient between the reconstructed and true feature vectors was higher than that between the reconstructed and false feature vectors. The same procedure was repeated until all other images were compared with the true presented image. Identification accuracy for a reconstructed image was defined as the proportion of comparison that were correctly identified. The same identification procedure was performed for each reconstructed image, and the mean identification accuracy was calculated over all reconstructed images. We use the term "identification accuracy" to indicate the mean identification accuracy for simplicity unless stated otherwise.

When optimizing the regularization parameters for the neural code converters, we used a set of decoded DNN features concatenated from multiple layers as a proxy of the DNN features extracted from the reconstructed images (see Materials and Methods: "Neural code converter analysis"). Following the same identification analysis, this set of decoded DNN features was used to calculate the identification accuracies to evaluate the performance of neural code converters. The candidate images were a subset of the 1,200 images presented in the training image session, depending on which fold of data was used as a validation dataset; for example, 2,400 training samples were split into five folds, with each fold containing 480 samples corresponding to 240 images.

We used the image pixel values of the reconstructed images or the DNN features extracted from the reconstructed images as the feature vector in the identification procedure to evaluate the

reconstruction quality. Each reconstructed image was tested with all pairs of images; for example, a reconstructed natural image was compared using 49 pairs, with each comparison consisting of one presented image and another (false) image from the remaining 49 natural test images, resulting in 50 × 49 = 2,450 comparisons. We used another neural network with a smaller network architecture (AlexNet, Krizhevsky et al., 2012; available from https://github.com/BVLC/caffe/tree/master/models/bvlc_alexnet; see also Materials and Methods: "DNN features"), to extract the visual features of both the reconstructed images and original images, and then performed the identification analysis on them layer by layer.

4.16 Pooling data from multiple subjects

First, we pooled data from multiple subjects to train the DNN feature decoders. In this analysis, we pooled the data of four subjects and left one subject as a test subject. Among the four subjects, one subject was selected as the target subject, and neural code converters for the other three subjects were trained using 6,000 samples from each subject. These three subjects' training data (6,000 samples for each subject) were then converted into the target brain space (Fig. 7A). Thus, there were 24,000 samples, including the original target subject data (6,000 samples) and the converted data of the three subjects (3 × 6,000 samples) in the target brain space. We then trained DNN feature decoders on the pooled data to obtain the multiple-subject feature decoders, in contrast to the single-subject feature decoders trained with individual data (6,000 samples) in a native brain space. The same analysis was repeated four times such that every subject of the four was selected as a target subject.

Second, we used a subset of training samples (2,400 or 300 samples) from the test subject to train a neural code converter. Training a neural code converter needs a sequence of source and target fMRI data responses to identical sequences of stimuli, but in this case, there were four sets of fMRI data (1 original + 3 converted) responding to the same set of stimuli in the target brain space. We thus paired these four sets of fMRI data to the 2,400 (or 300) samples from the test subject, resulting in 9,600 (or 1,200) pairs of samples. A neural code converter was then trained with this dataset to predict the fMRI activity patterns in the target brain space.

Finally, we performed inter-individual DNN feature decoding and visual image reconstruction. The test data of the test subject were converted into the target brain space using the trained neural code converter. We then decoded the converted fMRI activity patterns into the DNN features and reconstructed the visual images via the decoded DNN features.

4.17 Statistics

We used ANOVA to examine the effects of conversion types (whole VC and subarea-wise) on the decoding accuracies and identification accuracies with the conversion type as a repeated measure factor and DNN layer as a between-subject factor. We also used ANOVA to examine the effects of decoder types (multiple subject and single subject) on the decoding accuracies and identification accuracies with the decoder type as a repeated measure factor and DNN layer as a between-subject factor.

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: 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, https://openneuro.org/datasets/ds003430/versions/1.2.0, and https://openneuro.org/datasets/ds001506/versions/1.3.1).

Author contributions

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; **Yukiyasu Kamitani**: Conceptualization, Validation, Investigation, Resources, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition.

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 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

The authors have declared that no competing interests exist.

Acknowledgments

The authors would like to thank Jong-Yun Park, Fan Cheng, Shangfeng Jin, 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.

References

1. Avants BB, Epstein CL, Grossman M, Gee JC. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis. 2008;12: 26–41. doi: 10.1016/j.media.2007.06.004.
2. Bilenko NY, Gallant JL. Pyrcca: Regularized kernel canonical correlation analysis in Python and its applications to neuroimaging. Front Neuroinform. 2016;10: 49. doi: 10.3389/fninf.2016.00049.
3. Blumensath T, Jbabdi S, Glasser MF, Van Essen DC, Ugurbil K, Behrens TEJ, et al. Spatially constrained hierarchical parcellation of the brain with resting-state fMRI. NeuroImage. 2013;76: 313–324. doi: 10.1016/j.neuroimage.2013.03.024.
4. Chen P-H, Chen J, Yeshurun Y, Hasson U, Haxby J, Ramadge PJ. A reduced-dimension fMRI shared response model. Adv Neural Inf Process Syst. 2015;28: 460–468. Available from: http://papers.nips.cc/paper/5855-a-reduced-dimension-fmri-shared-response-model.pdf
5. Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; June 2009; Miami, FL, USA. IEEE; 2009. p. 248–255. doi: 10.1109/CVPR.2009.5206848.
6. Dosovitskiy A, Brox T. Generating images with perceptual similarity metrics based on deep networks. Adv Neural Inf Process Syst. 2016;29: 658–666. Available from: https://arxiv.org/abs/1602.02644
7. Engel SA, Rumelhart DE, Wandell BA, Lee AT, Glover GH, Chichilnisky E-J, et al. fMRI of human visual cortex. Nature. 1994;369: 525. doi: 10.1038/369525a0.
8. Epstein R, Kanwisher N. A cortical representation of the local visual environment. Nature. 1998;392: 598–601. doi: 10.1038/33402.
9. Esteban O, Markiewicz CJ, Blair RW, Moodie CA, Isik AI, Erramuzpe A, et al. fMRIPrep: A robust preprocessing pipeline for functional MRI. Nature Methods. 2019;16: 111–116. doi: 10.1038/s41592-018-0235-4.

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

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

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

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

14. Guntupalli JS, Hanke M, Halchenko YO, Connolly AC, Ramadge PJ, Haxby JV. A model of representational spaces in human cortex. Cerebral Cortex. 2016;26: 2919–2934. doi: 10.1093/cercor/bhw068.

15. Haxby JV, Guntupalli JS, Connolly AC, Halchenko YO, Conroy BR, Gobbini MI, et al. A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron. 2011;72: 404–416. doi: 10.1016/j.neuron.2011.08.026.

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

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

18. Hsu A, Borst A, Theunissen FE. Quantifying variability in neural responses and its application for the validation of model predictions. Network: Computation in Neural Systems. 2004;15: 91–109. doi: 10.1088/0954-898X_15_2_002.

19. Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology. 1962;160: 106–154. doi: 10.1113/jphysiol.1962.sp006837.

20. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, et al. Caffe: Convolutional architecture for fast feature embedding. arXiv:1408.5093 [Preprint]. 2014 [cited 2021 Nov 8]. Available from: https://arxiv.org/abs/1408.5093.

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

22. Kourtzi Z, Kanwisher N. Cortical regions involved in perceiving object shape. J Neurosci. 2000;20: 3310–8. doi: 10.1523/JNEUROSCI.20-09-03310.2000.

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

24. Laumann TO, Gordon EM, Adeyemo B, Snyder AZ, Joo SJ, Chen M-Y, et al. Functional system and areal organization of a highly sampled individual human brain. Neuron. 2015;87: 657–670. doi: 10.1016/j.neuron.2015.06.037.

25. Le QV, Ngiam J, Coates A, Lahiri A, Prochnow B, Ng AY. On optimization methods for deep learning. Proceedings of the 28th International Conference on International Conference on Machine Learning; June 2011; Bellevue, Washington, USA. Omnipress. 2011. p. 265–272.

26. Lescroart MD, Gallant JL. Human scene-selective areas represent 3D configurations of surfaces. Neuron. 2019;101: 178-192. doi: 10.1016/j.neuron.2018.11.004.

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

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

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

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

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

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

33. Sereno MI, Dale AM, Reppas JB, Kwong KK, Belliveau JW, Brady TJ, et al. Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. Science. 1995;268: 889–893. doi: 10.1126/science.7754376.

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

35. Shen G, Horikawa T, Majima K, Kamitani Y. Deep image reconstruction from human brain activity. PLoS Computational Biology. 2019;15: 1006633. doi: 10.1371/journal.pcbi.1006633.

36. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556v1 [Preprint]. 2014 [cited 2021 Nov 8] Available from: https://arxiv.org/abs/1409.1556.

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

38. Van Essen DC. Surface-based approaches to spatial localization and registration in primate cerebral cortex. NeuroImage. 2004;23: S97–107. doi: 10.1016/j.neuroimage.2004.07.024.

39. Van Uden CE, Nastase SA, Connolly AC, Ma FL., Hansen I, Gobbini MI et al. Modeling semantic encoding in a common neural representational space. Front Neurosci. 2018;12: 437. doi: 10.3389/fnins.2018.00437.

40. Watson JDG, Myers R, Frackowiak RSJ, Hajnal JV, Woods RP, Mazziotta JC, et al. Area V5 of the human brain: Evidence from a combined study using positron emission tomography and magnetic resonance imaging. Cerebral Cortex. 1993;3: 79–94. doi: 10.1093/cercor/3.2.79.

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

42. Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci USA. 2014;111: 8619–8624. doi: 10.1073/pnas.1403112111.