Inter-individual deep image reconstruction

Jun Kai Ho[1¶], Tomoyasu Horikawa[2¶], Kei Majima[1], Yukiyasu Kamitani[1,2*]

[1] Graduate School of Informatics, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan

[2] 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. While 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 contents. In this study, we evaluated machine learning models called neural code converters that predict one's brain activity pattern (target) from another's (source) given the same stimulus by the decoding of hierarchical visual features and the reconstruction of perceived images. The training data for converters consisted of fMRI data obtained with identical sets of natural images presented to pairs of individuals. Converters were trained using the whole visual cortical voxels from V1 through the ventral object areas, without explicit labels of visual areas. We decoded the converted brain activity patterns into hierarchical visual features of a deep neural network (DNN) using decoders pre-trained on the target brain and then reconstructed images via the decoded features. Without explicit information about visual cortical hierarchy, the converters automatically learned the correspondence between the visual areas of the same levels. DNN feature decoding at each layer showed higher decoding accuracies from corresponding levels of visual areas, indicating that hierarchical representations were preserved after conversion. The viewed images were faithfully reconstructed with recognizable silhouettes of objects even with relatively small amounts of data for converter training. The conversion also allows pooling data across multiple individuals, leading to stably high reconstruction accuracy compared to those converted between individuals. 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.

Author summary (150-200 words)

While the brains show individual differences, information of perceptual contents is thought to be processed and represented through hierarchical processing stages. The stimulus information encoded in brain activity patterns could be aligned/converted across individuals using functional alignment. However, it is unclear whether hierarchical and fine-grained features encoded in brain responses can be functionally converted across individuals while preserving encoded perceptual contents. Here, we performed image reconstruction from visual stimulus-induced brain activity converted from different individuals using a machine learning-based converter model. The converter model showed moderate conversion accuracy for voxels distributed across visual areas and reasonably learns visual areas' correspondence at the same level between individuals. We then decoded the converted activity patterns into hierarchical DNN features to reconstruct visual images and showed that the hierarchical and fine-scale visual features are preserved after conversion. Using converter models, pooling data from multiple individuals has demonstrated its effectiveness in achieving stably high reconstruction accuracy than those converted between individuals. Visual image reconstruction using decoders trained on other individuals can potentially reduce the burden of collecting data to train individual-specific decoders and might prove helpful in promoting brain-machine interfaces that externalize perceptual contents.

Introduction

Sensory information is generally thought to be processed through a hierarchical pathway that detects topographically organized local simple features at early stages and then progressively global complex features at 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 & Wiesel, 1962) [1], and is then further processed in 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 & Ungerleider, 1982) [2]. While the general principles such as topography and hierarchy appear to govern the organization of the visual cortex, 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 contents by simple mapping rules common across individuals.

Recent advances of deep neural networks (DNNs) have enabled detailed analyses of hierarchical feature representations across visual cortical areas (Yamins et al., 2014; Güçlü & van Gerven, 2015; Güçlü & van Gerven, 2017; Horikawa & Kamitani, 2017) [3–6]. 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 contents encoded in brain responses have been successfully reconstructed as images via DNN-based reconstruction algorithms (Shen, Horikawa, Majima, & Kamitani, 2019; Shen, Dwivedi, Majima, Horikawa, & Kamitani, 2019) [7, 8]. In

particular, deep image reconstruction (Shen, Horikawa et al., 2019) [7] that was used in this study consists of two steps: 1) DNN feature decoders are trained to predict the DNN features of an image from a brain activity pattern, given that image as the stimulus perceived by a subject; 2) a reconstructed image is iteratively optimized such that its DNN features resemble the predicted DNN features. However, these predictive models required optimization by machine learning algorithms using training data derived from hours of experiments to measure brain responses to hundreds or thousands of images. Further, 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 anatomical and functional alignment of different individuals' brains have been developed in the 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) [9, 10]. The retinotopic organization that shows a correspondence between the visual field and visual cortex could be used to delineate functional visual areas by using the traveling wave paradigm (Engel et al., 1994; Sereno et al., 1995) [11, 12]. The delineated visual areas are often similar but still different across individuals. Anatomical alignment could mitigate the anatomical difference by matching the anatomical features between brains (Fischl et al., 2008; van Essen, 2004 & 2005) [13–15], but it still cannot perfectly align the functional topography across individuals (Watson et al., 1993) [16]. Functional alignment adopts an anatomy-free approach by

learning statistical relationships between subjects' brain activity patterns (Haxby et al., 2011; Yamada, Miyawaki, & Kamitani, 2015; Chen et al., 2015; Bilenko & Gallant, 2016; Guntupalli et al., 2016) [17–21]. Methodologies of functional alignment could be categorized into two types: (a) direct linear mapping between voxels of two subjects, for example, neural code converter; (b) a shared latent space among subjects is constructed, and mapping between subjects could be mediated by the shared space, for example, hyperalignment. In particular, functional alignment methods have uncovered common neural representations across individuals concealed under substantial individual variations of brain responses. However, those investigations have been often focused on a few specific features, for example, object categories, image contrast, retinotopy, and semantics (Haxby et al., 2011; Yamada et al., 2015; Bilenko & Gallant, 2016; Van Uden et al., 2018) [17, 18, 20, 22], leaving it unclear whether distinct levels of fine-grained neural representations of hierarchical visual features can be converted across individuals so that one's perceptual experience can be reconstructed using the model trained on other individuals. Furthermore, the previous studies have separately performed alignments on different brain areas using rough anatomical correspondences across individuals. 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 contents. For this purpose, we evaluate the functional alignment methods (neural code converter; Yamada et al., 2015; hyperalignment, Haxby et al., 2011]) [18, 17] using the decoding of hierarchical DNN features (Horikawa & Kamitani, 2017) [6] and reconstruction of perceived

images (deep image reconstruction; Shen et al., 2019) [7]. We mainly discuss the results with 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 visual cortex of one subject (source) to individual voxel responses of another subject (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 through the DNN feature decoders pre-trained on the target subject, and then the decoded features are fed to the image generator to create images (Fig 1B).

In this study, we first show that machine learning-based converter models accurately convert fMRI responses of a source subject to those of a target subject. Notably, while the converter training for individual target voxels is performed from voxel patterns of the entire visual cortex without explicit information about cortical hierarchy, the converters can automatically learn the hierarchical correspondence of visual subareas at the same levels between the source and target subjects. Furthermore, feature decoding from the converted fMRI responses at each hierarchical DNN level shows greater accuracy in corresponding levels of visual subareas than other levels, indicating that fine-grained hierarchical feature representations are preserved even after conversion. We show that image reconstruction applied on decoded features originated from the source subject's brain via the conversion produces faithful reconstructions of viewed images even with small amounts of data for converter training. We also demonstrate that the converter

trained with explicit information about cortical hierarchy shows no distinct difference from the converter trained with sufficient training data and without that information. Lastly, we pool data from multiple subjects and show that DNN feature decoders trained on the pooled data could achieve stably high decoding accuracies. 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.

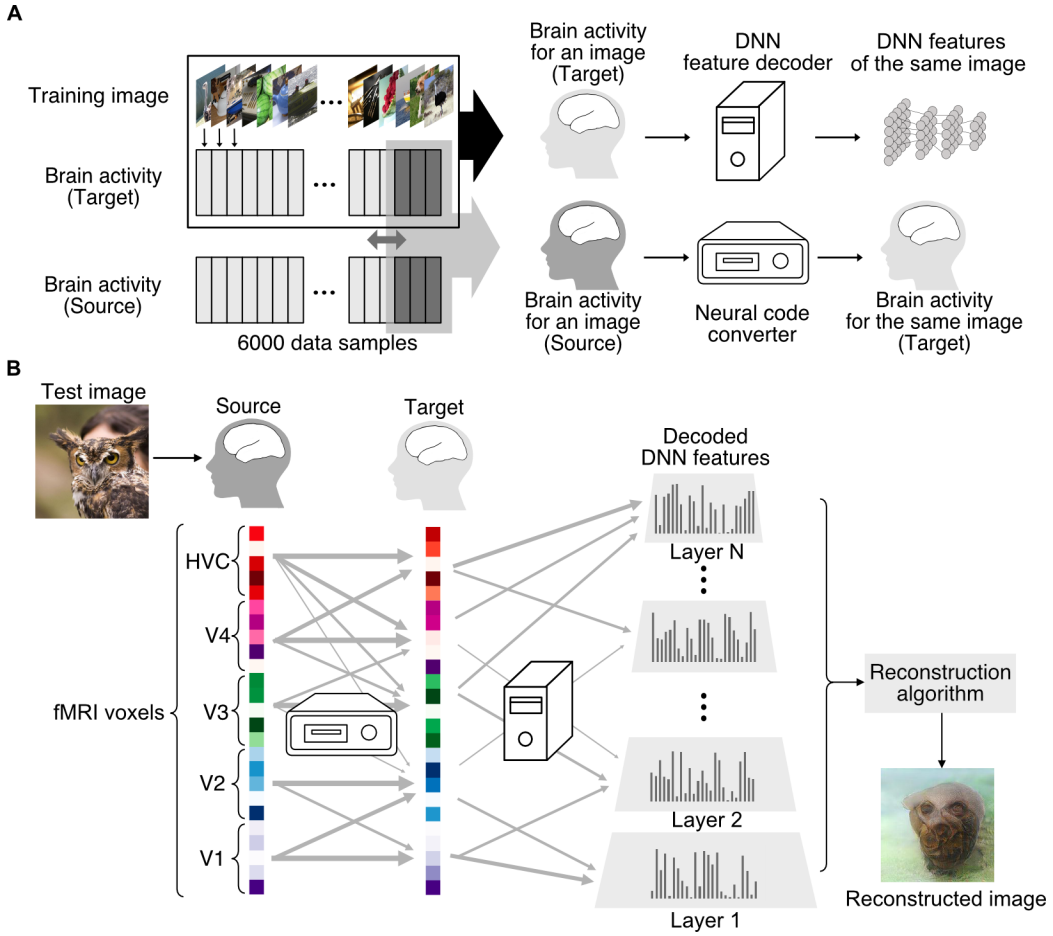

Fig 1. Inter-individual deep image reconstruction.
(A) Training of the DNN feature decoder and neural code converter. DNN feature decoding models are trained on the 6,000 samples of measured fMRI activity of the target subject and the corresponding DNN features. A converter model is trained on a subset 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 assumed in 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.

Results

We used fMRI datasets collected in previous studies (Shen et al., 2019; Horikawa, Aoki, Tsukamoto, & Kamitani, 2019) [7, 23], in which fMRI responses to visual stimuli were measured from five subjects (see Materials and Methods: "FMRI datasets and stimuli"). From the published datasets, we used fMRI data measured when subjects viewed 1,200 natural images in "training natural image session," 50 natural images in "test natural image session," and 40 artificial images in "test artificial image session." Each image was presented during an 8-s stimulus block in 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 training natural images, 50 test natural images, and 40 test artificial images, respectively, resulting in 6,000 training samples, 1,200 test samples with natural images, and 800 test samples with artificial images. We used the DNN feature decoders from [Shen et al., 2019], trained using the whole 6,000 training samples in each of the five subjects. Note that the fMRI data collection for the whole training samples required approximately 800 minutes in each subject. The current study trained functional alignment models 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 results obtained with 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 the neural code converter analysis to examine whether and how visual information of one subject can be converted to another subject. In this analysis, we constructed neural code converter models between each pair of subjects, one as the target

subject and the other as the source subject, resulting in 20 subject pairs. A converter model comprised a set of regularized linear regression models (ridge regression), each trained to predict each voxel's activity of a target subject from a source subject's brain activity patterns in the visual cortex (see Materials and Methods: "Neural code converter analysis"). Although we mainly showed the neural code converter analysis in this study, we also used hyperalignment analysis to compare. Hyperalignment instead estimated a shared space among subjects by iteratively performing a series of Procrustean transformations; the one-to-one mapping between subjects could be constructed via the shared space (S1A Fig; see Materials and Methods: "Hyperalignment analysis"). Similar to 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 test natural images into a target subject's brain space or a shared space.

The visual cortex (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), fusiform face area (FFA), and parahippocampal place area (PPA), each of them was obtained by localizer experiments. We defined a continuous region covering LOC, FFA, and PPA as the higher visual cortex (HVC). In the converter and hyperalignment analysis, the left and right VC 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 by two methods: (a) the pattern correlation, which is the spatial Pearson correlation coefficient between the converted and measured voxel patterns for each test image, and (b) the profile correlation, which is the Pearson correlation coefficient between the sequences of converted and measured individual voxel responses to the 50 test natural images (Fig 2A top). 50 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 label was defined by the mean of 24 samples × 24 samples = 576 correlation coefficients. The profile correlation for each voxel was defined by the mean of 24 runs × 24 runs = 576 correlation coefficients. The obtained correlation coefficients were normalized by their noise ceilings (Hsu, Borst, & Theunissen, 2004; Lescroart & Gallant, 2019) [24, 25]. The noise ceiling was estimated by the mean pattern or profile correlation coefficient between measured voxel patterns or individual voxel responses of target subjects (see Materials and Methods: "Noise ceiling estimation"). To summarize the results, the correlation coefficients were averaged across images and voxels for the pattern and the profile correlations, respectively, in each subject pair for conversion.

In Fig 2A (bottom), 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 visual cortex was 0.56 ± 0.12 (mean ± standard deviation) over subject pairs, with the visual subareas showing comparable distributions (Fig 2A bottom left). The mean normalized profile correlation for VC was 0.53 ± 0.10 over subject pairs. The subareas also showed similar distributions to VC. The conversion accuracies were modest in both pattern and profile

correlations across all visual subareas but comparable to the previous study (Yamada et al., 2015) [18]. The hyperalignment-based methods showed no clear advantage over the neural code converter (S1B & S1C Fig).

To see how visual areas of 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 to the converter models (Fig 2B top). We evaluated the degree of the performance drop (normalized profile correlation difference) from the condition using all source visual subareas (*i.e.*, whole VC). This sensitivity analysis showed that the effects of the source area exclusions varied according to the location of voxels in the target brain. The largest performance drop of each target voxel was often caused by the exclusion of the corresponding source area (Fig 2B bottom, target S1; see S2 Fig for other target subjects). On average, maxima of the mean performance drop across voxels for target visual areas V1, V2, 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 2C). The results indicate that the machine learning-based neural code converter models could automatically detect a "low-to-high" hierarchical correspondence between source and target visual areas even without explicit information of the correspondences between them.

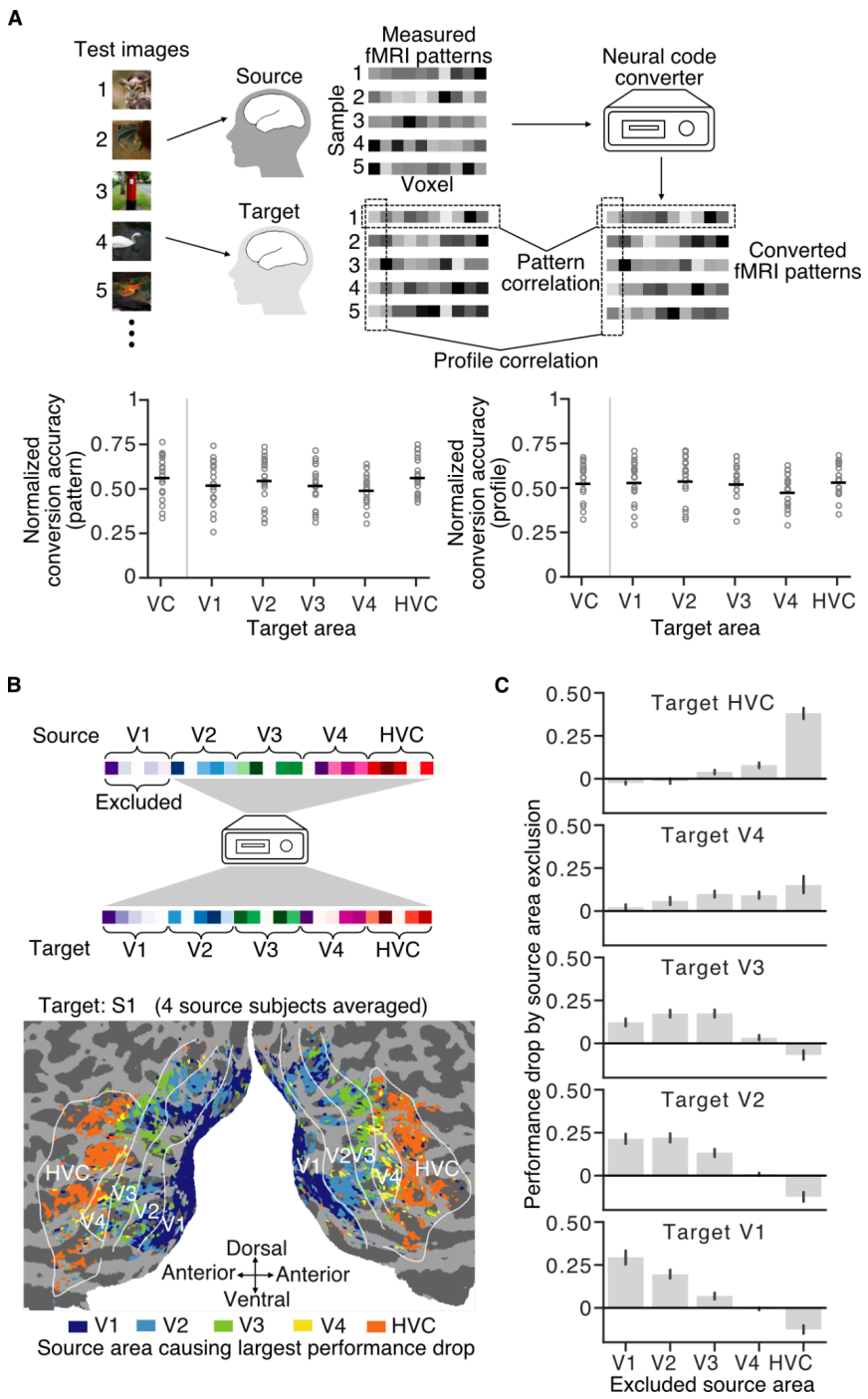

Fig 2. Performance of neural code converters and cortical hierarchical correspondence. (A) Evaluations of neural code converters. Two types of evaluations were performed by computing Pearson correlation coefficients: pattern and profile correlation (top). Distributions of normalized pattern or profile correlation coefficients are shown for VC and visual subareas (bottom). Each horizontal black dash indicates the mean value; each circle represents the correlation coefficients of a subject pair.

(B) Sensitivity analysis on neural code converters. The analysis is performed by excluding one source visual area on the prediction of target voxels' activities (top). The cortical map of the

effects of source area exclusion was shown for one target subject (S1, bottom). 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 test natural image dataset (performance drops were averaged across four source subjects for a single target subject; S2 Fig for other target subjects). Only voxels that generate reliable responses were shown.
(C) 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).

Next, to examine whether fine-grained representations of visual features were preserved in the converted fMRI activity patterns, we performed the DNN feature decoding analysis (Horikawa & Kamitani, 2017) [6]. We used the VGG19 DNN model that consists of sixteen convolutional layers and three fully-connected layers (Simonyan & Zisserman, 2014) [26] (see Materials and Methods: "Deep neural network features"). Feature decoders were trained using 6,000 samples of a target subject's fMRI activity patterns in either the whole VC or 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]"; see Materials and Methods: "DNN feature decoding analysis"). Following the original paper, we averaged the converted fMRI activity patterns across repetitions for each test image to enhance the signal-to-noise ratio. For each DNN unit, a feature decoder takes an 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 decoded feature values and image feature values computed from the presented images (test natural images).

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"). 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").

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 the anatomical alignment in DNN feature decoding. Additional analyses showed that the neural code converter outperformed the hyperalignment-based methods in DNN feature decoding, too (S1D Fig).

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) [27]. 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, which is based on the rank correlations between the hierarchical levels across DNN units (Fig 3C; Nonaka et al., 2021; see Materials and Methods: "Brain hierarchy score"). The results of the within individual condition ("Within") replicate the previous findings (Horikawa 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 of the cortical hierarchy information inherited through the process of anatomical alignment. The inter-individual conversion ("Across - functional") showed a lower but substantial degree of hierarchical correspondence, even though the converter was blind to anatomical information during the training.

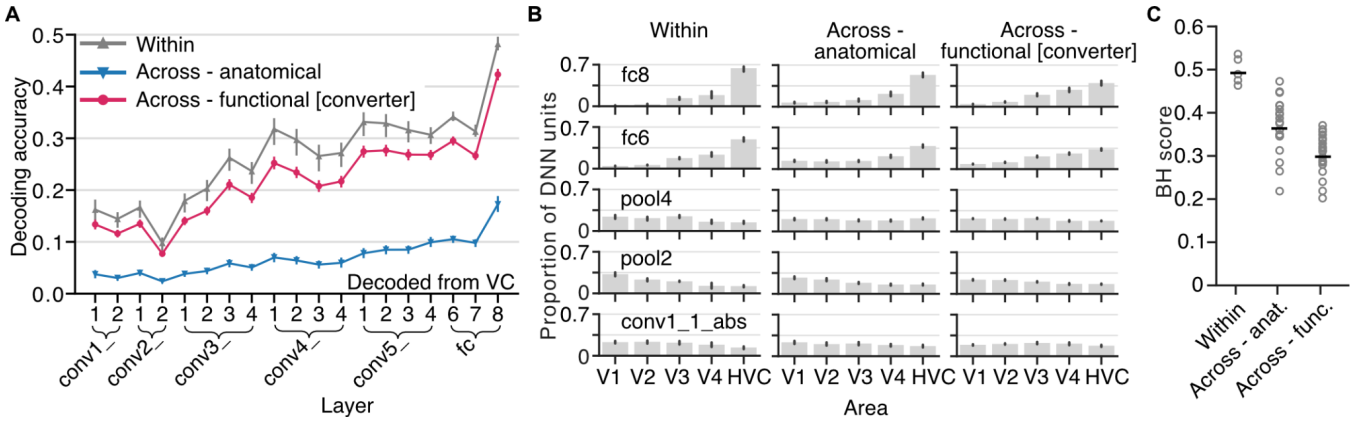

Fig 3. DNN feature decoding and hierarchical representation by DNN.
(A) DNN feature decoding accuracy from the whole visual cortex (VC). Decoding accuracies for "Within", "Across - functional [converter]" and "Across - anatomical" decoding analyses from visual cortices are shown for each layer of the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

Visual image reconstruction

We so far confirmed that multiple levels of DNN feature representations can be decoded from the converted brain activity patterns. Then, we went on to test if we can 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) [7]. The visual image was reconstructed 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) [28], 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 image reconstructions via the DNN features decoded from VC in the "Within," "Across - functional [converter]," and "Across - anatomical" conditions. Reconstructed images obtained in the "Within" and "Across - functional [converter]" captured the characteristics of the presented images, including shapes and colors of the objects (Fig 4A; see S3 Fig and S4 Fig 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 (S5 Fig). The reconstructed images obtained in the "Across - functional [converter]" condition showed comparative good

visual quality with those in the "Within" condition. Reconstructed images obtained in the "Across - anatomical" condition showed neither recognizable shape nor the color of the objects in the presented images.

To quantitatively evaluate the quality of reconstructed images, we performed a pairwise identification analysis based on image pixel values and DNN feature values. In this analysis, correlation coefficients that were calculated either from image pixel values or DNN feature values extracted from the reconstructed images by the AlexNet model (Krizhevsky, Sutskever, & Hinton, 2012) [29] were used to identify a reconstructed image from a pair of images (a true [presented] image and a false image; see Materials and Methods: "Identification analysis"). The identification accuracies based on pixel values and DNN feature values of all DNN layers measure the similarity between the presented and reconstructed images in multiple levels of visual information. In the evaluations of reconstructed natural or artificial images, the "Within" condition ranked the top with the highest accuracies across all features. The "Across - functional [converter]" condition showed 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.

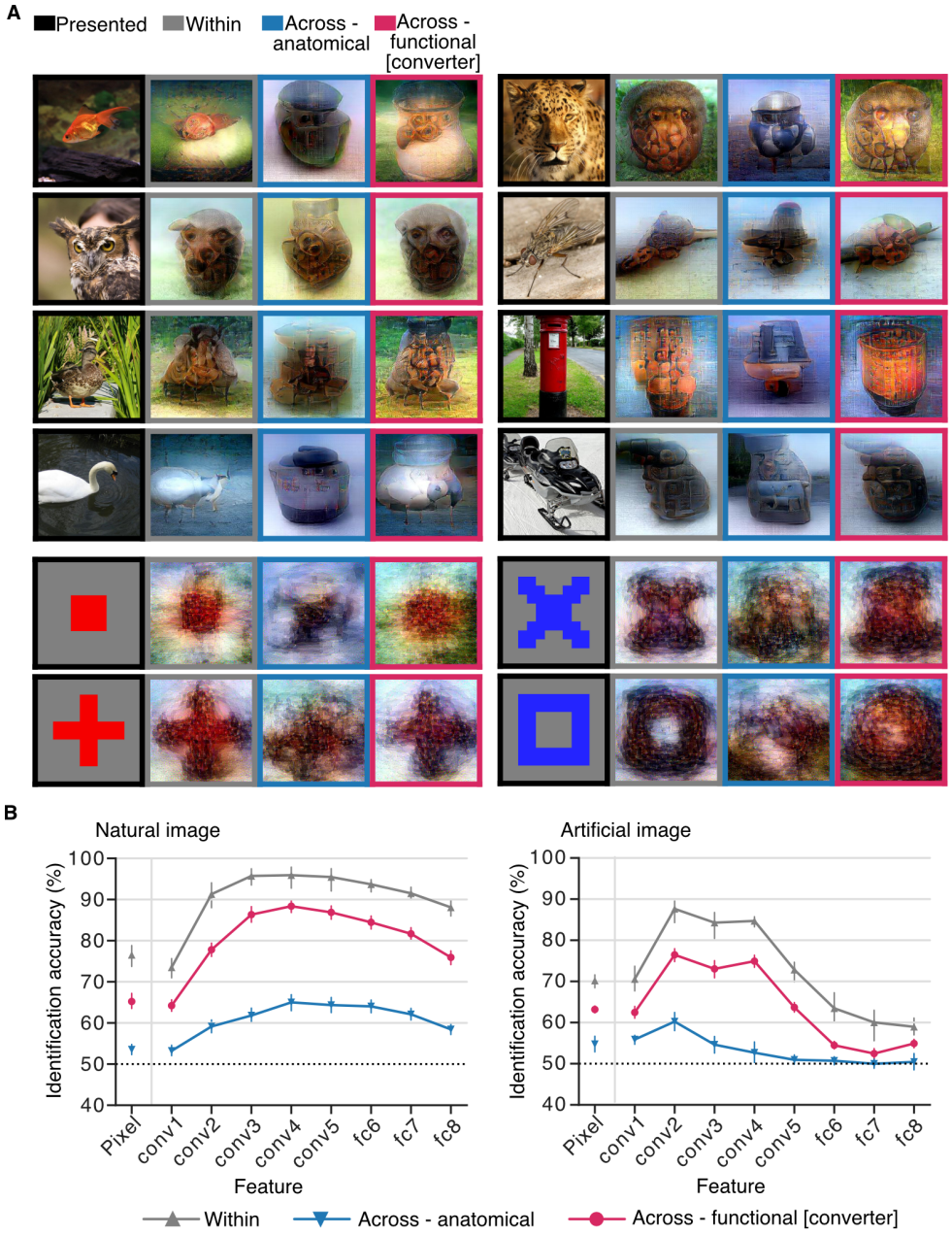

Fig 4. Reconstructed images and evaluations.
(A) Within and across individual reconstruction from the whole visual cortex. Reconstructed images of the three analytical conditions were all from the same source subject for each presented image. We show reconstructed images produced from different source 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 "Within" condition and across 20 pairs for "Across - functional [converter]" and "Across - anatomical" conditions; dotted lines, chance level, 50%).

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

In DNN feature decoding on the test natural images, the whole VC condition showed similar but slightly higher decoding accuracy than the subarea-wise condition across layers (Fig 5B; ANOVA, effect of conversion type, $F(1, 361) = 1959$, $p < .001$, $\eta_p^2 = .84$). Reconstructed images obtained from subarea-wise conversions showed a similar visual quality to whole VC conversions (Fig 5C). In the identification analysis of reconstructed natural images and artificial images (cf., Fig 4), although the whole VC conversion slightly outperformed subarea-wise condition (Fig 5D), the effect size was very small (ANOVA, $F(1, 171) = 11.2$, $p < .001$, $\eta_p^2 = 0.062$). Overall, the subarea-wise conversion did not show a significant advantage over the whole VC conversion. Thus, the conversion with explicit information about cortical hierarchy does not seem to significantly contribute to the improvement of visual image reconstruction.

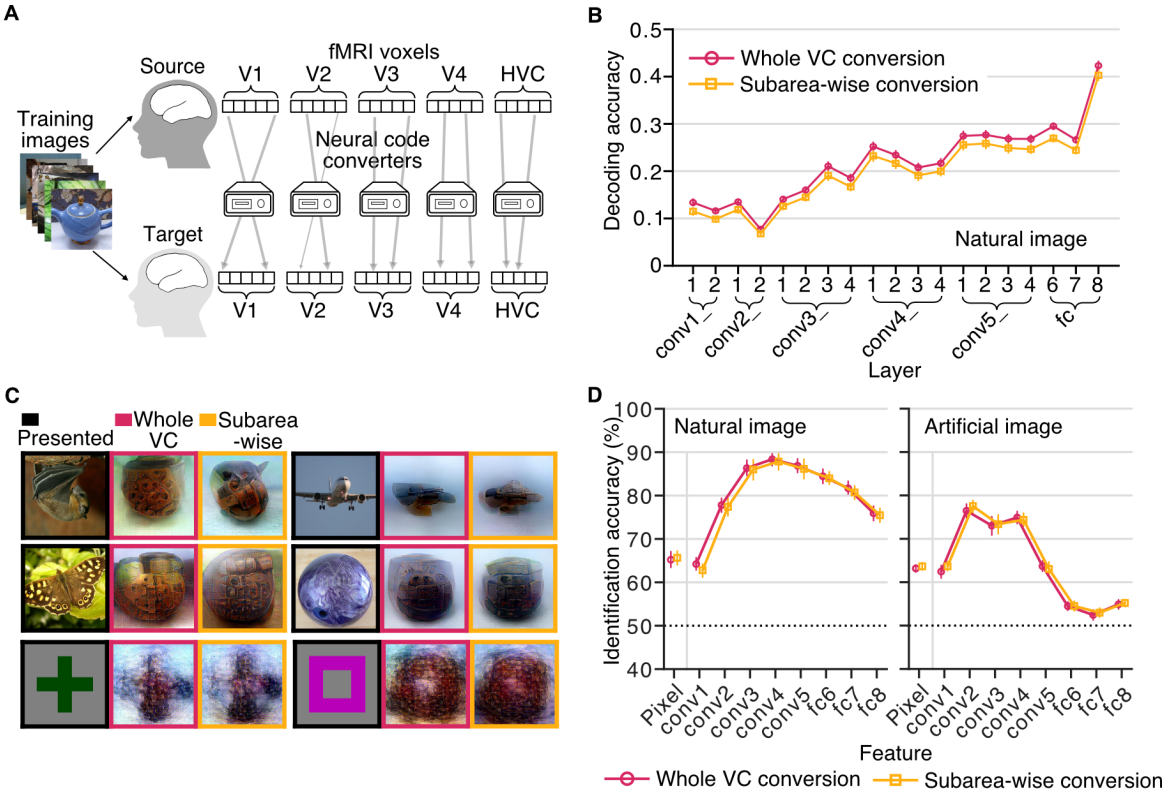

Fig 5. Whole VC conversion and subarea-wise conversion.
(A) Illustration of the subarea-wise conversion. A converter model is trained with a set of fMRI responses to an identical stimulus sequence. Activity values of a voxel in a target area are predicted only from source subjects' brain activity patterns in its corresponding source area.
(B) DNN feature decoding accuracy (error bars, 95% C.I. across 20 subject pairs).
(C) Reconstructed natural and artificial images. Conventions are the same as Fig 3A.
(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%).

One of the potential benefits of inter-individual analyses is to reduce the amount of data required for model training from novel test (source) subjects by substituting training data from other individuals. As 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 amount of training data (300, 600, 900, 1,200, 2,400, 3,600, 4,800, and 6,000 training samples) while using all the available data of 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.

While the visual quality of reconstructions proportionally decreased according to the decrement of the number of samples, the reconstructions obtained via converters trained with 300 samples still produced discernible natural images (Fig 6A, see S7 Fig for artificial images). This result suggests that we could compromise some visual quality and perform image reconstruction using converters instead of collecting fMRI data for a full reconstruction model. The identification accuracies using pixel values and extracted DNN feature values increased with the number of training samples (Fig 6B, natural image; see S7 Fig for artificial image). The identification accuracy in the "Across - functional [converter]" condition was asymptotically closer to that in the "Within" condition while not reaching 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 were similar between the whole VC and subarea-wise conversions 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. In the identification based on conv1 feature values of the reconstructed natural images, the whole VC conversion with equal to or more than 4,800 training samples showed higher accuracies than the subarea-wise conversion. Overall, the explicit information about cortical hierarchy does not improve reconstruction, but it may be beneficial when the number of training data is limited.

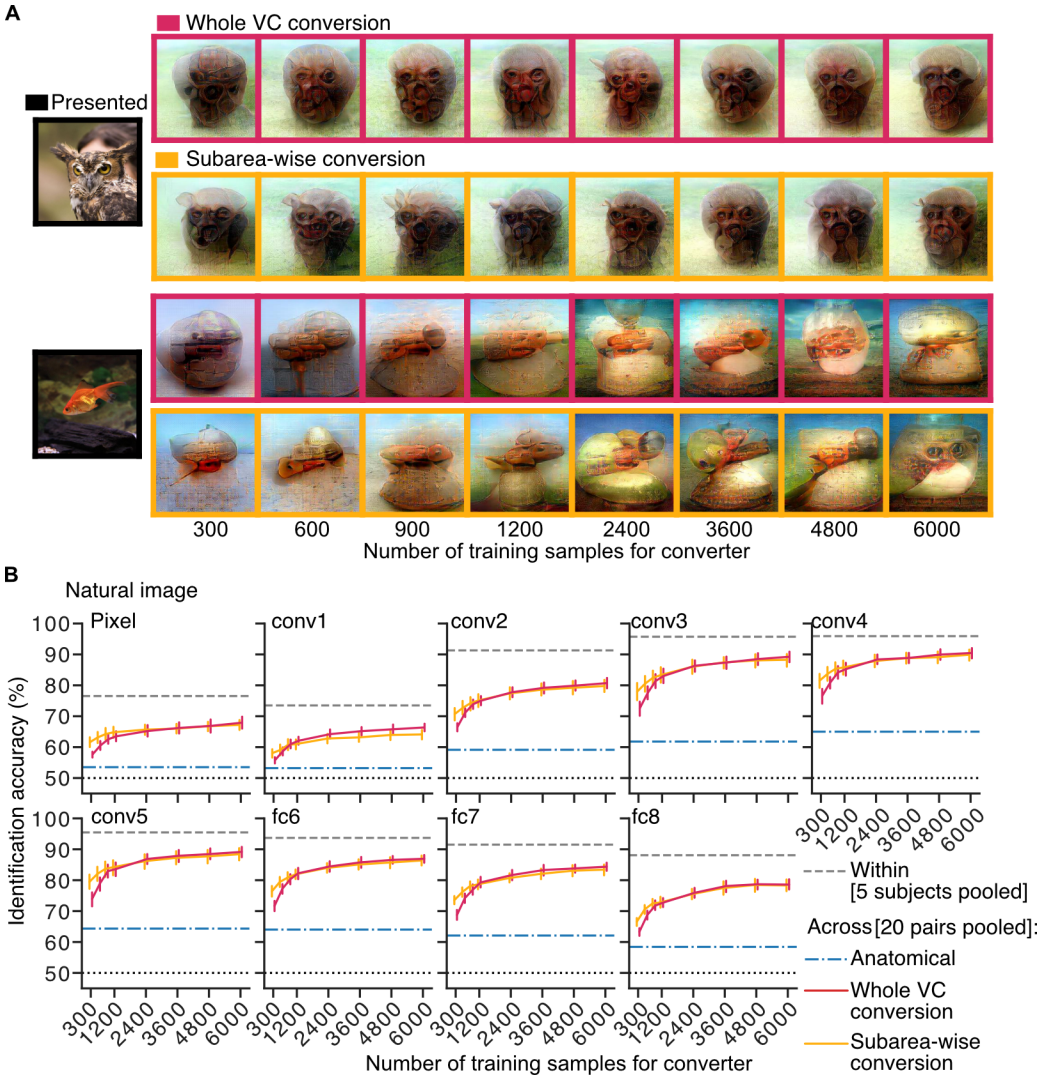

Fig 6. Reconstructed images via converters trained on varying amounts of data.
(A) Reconstructed images via converters trained on varying amounts of data. All reconstructed images were produced from the same source subject (source: S2, target: S5)
(B) The effect of the amount of training data on the identification accuracy. The converters were trained on a varying amount 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%).

As the neural code conversion allowed us to pool the data of multiple subjects into a single target subject brain space, we finally examined the pooling effect on the performance of the inter-individual image reconstruction. This analysis left out one subject for testing and pooled data

from the other four subjects. We assumed the test subject only had a limited amount of data (300 training samples and full test data) to reflect the circumstance that most people could only collect a limited amount of data because of the high cost of collecting large amounts of fMRI data. 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 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 native brain space. With 300 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 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 natural images showed that the multiple-subject condition tended to achieve small but stably high accuracy across all layers as compared with single-subject condition (ANOVA, effect of decoder type, $F(1, 57) = 182$, $p < .001$, $\eta_p^2 = .76$; Fig 7B for natural image; see S8 Fig for artificial image). Reconstructed images obtained in both single- and multiple-subject decoders showed recognizable visual quality (Fig 7C). Pixel values and DNN features of reconstructed natural images obtained in the multiple-subject condition did

not show higher identification accuracies than those obtained in the single-subject condition across all features as revealed by a similar ANOVA analysis ($F(1, 27) = 2.74$, $p = .11$, $\eta_p^2 = .09$; Fig 7D for natural images; see S8 Fig for artificial images). However, if single-subject decoders of a particular subject show particularly lower accuracies than others, pooling data into this subject's brain space did improve the image reconstruction (see S9 Fig). These results indicate that pooling multiple-subject data is beneficial to achieve stably high accuracy in the inter-individual image reconstruction, although the visual quality of reconstructed images was not largely improved.

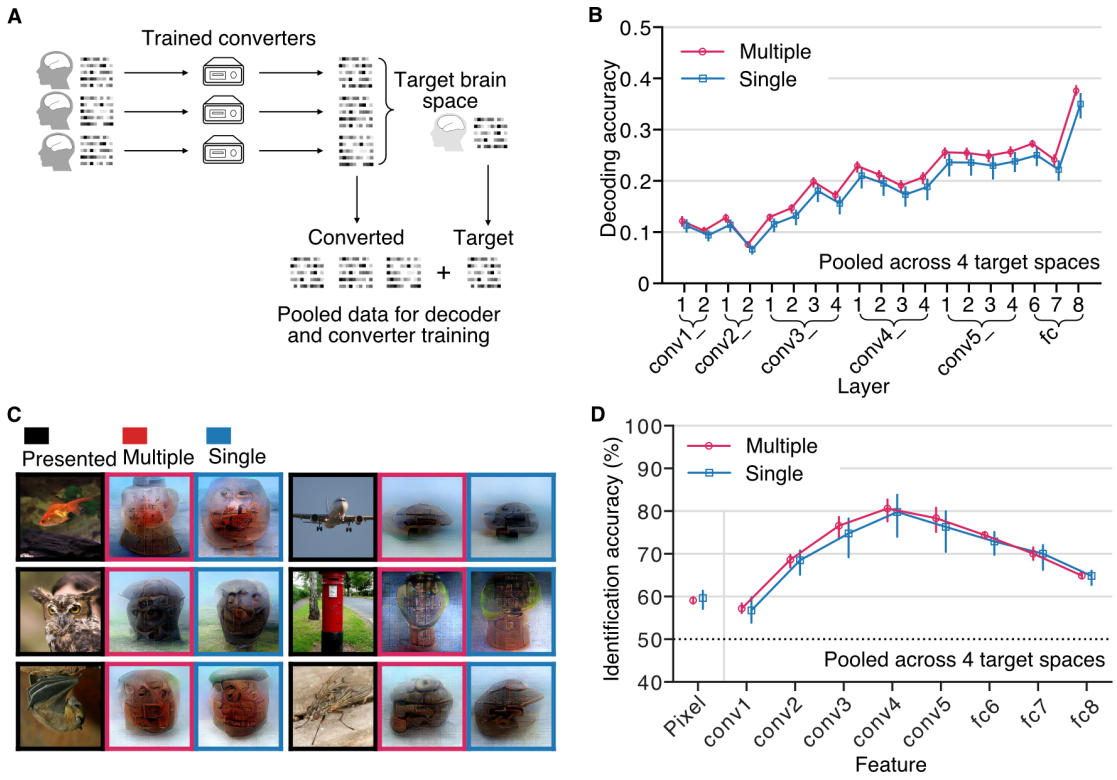

Fig 7. Pooling data across multiple subjects.
(A) Illustration of the pooling procedure. Three subjects' training data were converted into one subject's brain space. The DNN feature decoders were trained with three subjects of converted data plus the target subject's data (24,000 samples). Among the four subjects (excluding the left-out test subject), each subject was selected as the target subject once, so there are four sets of pooled data in different brain spaces.

(B) DNN feature decoding accuracy obtained via multiple- and single-subject feature decoders. Multiple-subject feature decoders trained on data pooled from four subjects; single-subject feature decoders trained on only one single subject's data. The accuracies were obtained from fMRI data corresponding to the test natural image session (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 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%).

Discussion

This study investigated whether and how the hierarchical and fine-grained visual information could be converted while preserving perceptual contents across individuals by 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 sensitivity analysis on the converters with the exclusion of voxels from a source visual subarea 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 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 amount of training data, the converters preserve minimally sufficient information for visual image reconstruction. Pooling data from multiple subjects helped achieve stably high 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 had mainly targeted a specific brain area, like V1 or inferior temporal (IT) cortex (Yamada et al., 2015; Haxby et al., 2011) [18, 17], while some other studies had targeted a large region including multiple levels of hierarchy (Bilenko & Gallant, 2016; Van Uden et al., 2018) [20, 22], but the hierarchical correspondence of visual subareas was not discussed. Our results explicitly demonstrate that machine learning-based neural code converters could 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., no inter-regional prediction between V1 and HVC), presumably indicating 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 (Figs 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 [17, 18, 20, 22], 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 contents 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 (Figs 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 showed 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 visual subareas could implicitly learn such information, provided sufficient training data.

Training a full visual image reconstruction model requires a lot of time and costs to collect an fMRI dataset. In this study, the DNN feature decoders were trained with 6,000 data samples, which took around 800 minutes of data collection time. This data collection time is long in fMRI studies and is inaccessible for most people. The situation even worsens when there are multiple decoding models trained on different datasets for other purposes. By trading off some visual quality of reconstructed images, we could instead collect less amount of data, say 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 of voxels between subjects under a variety of visual scenes, potentially being able to be used together with other

decoding models. The inter-individual decoding method with the neural code converter could thus reduce the time and costs for fMRI data collection.

Our results showed that pooling data from multiple subjects did not greatly improve the inter-individual image reconstruction (Fig 7). The possible reason could be that the pooled fMRI data are 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 could reduce the model's variance, but it might not reduce the model's bias if the training dataset is not diverse or not truly representative. Another possible reason is that different subjects show different data qualities, with some subjects' data showing relatively poor visual quality on reconstructed images. Pooling those data with poorer quality thus limits the decoder's ability to take advantage of pooled data, resulting in a limited improvement on visual image reconstruction. Although pooling data did not show great improvement in visual image reconstruction, it is still promising in other fMRI studies.

One observation was that the inter-individual image reconstruction did not outperform the within individual image reconstruction (Fig 4 & 5). As our converter models were based on linear regression models, such a linear constraint for conversion might be too restrictive to learn more complex statistical relationships, such as nonlinearity, between brain activity patterns, thus limiting the performances of inter-individual image reconstructions. On the other hand, 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, Gazzola, Hasson, & Keysers, 2019) [30]. 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 the within individual image reconstruction.

Our results showed that brain activity patterns could 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 burden of collecting a large amount of data, our approach might prove useful in popularizing the use of brain-machine/computer interfaces that communicate with our internal world.

Materials and methods

FMRI datasets and stimuli

We used fMRI datasets originally collected by Shen et al. (2019) [7] (https://openneuro.org/datasets/ds001246/versions/1.0.1) and Horikawa et al. (2019) [23] (https://openneuro.org/datasets/ds003430). These datasets collectively consisted of fMRI responses to visual stimuli (natural images and artificial images) measured from a total of five human subjects. The fMRI signals were measured while subjects viewed sequences of visual images (image presentation experiment). The image presentation experiment comprised the 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) [28]. The training natural images were 1,200 images taken from 150 object categories, and the test natural 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, while the 50 images were repeatedly presented 24 times in the test natural image session. The artificial image stimuli consisted of 40 combinations of 5 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 (TR = 2 s).

MRI data preprocessing

The fMRI data were preprocessed through the MRI data preprocessing pipeline provided by fMRIPrep (version 1.2.1; Esteban et al., 2019) [31]. A T1w reference image was generated

through a custom pipeline, including bias field correction, skull stripping, 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 regressing out nuisance parameters, including a constant baseline, a 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 SD for each run) and were averaged within each 8-s trial (training and test session, four volumes).

Anatomical alignment

We used fMRIPrep [31] to nonlinearly normalize subjects' structural and functional images to a standard space: MNI152NLin2009cAsym. By default, fMRIPrep uses Advanced Normalization Tools (ANTs, http://stnava.github.io/ANTs/), which implements the symmetric image normalization method (SyN; Avants, Epstein, Grossmann, & Gee, 2008) [32], to register images to a template. Following standard MRI data preprocessing (see Materials and Methods: "MRI data preprocessing"), the T1w reference image was further spatially normalized to the MNI-space by the ANTs, and the functional data were coregistered to this normalized T1w reference image. The coregistered data were then re-interpolated to 2 × 2 × 2 mm voxels. 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 as that of the corresponding target voxel.

Regions of interest (ROI)

For brains in native space, the VC was defined as the combined regions of V1, V2, V3, V4, and the higher visual cortex (HVC). V1, V2, V3, and V4 were delineated using the standard retinotopy experiment (Engel et al., 1994; Sereno et al., 1995) [11, 12]. The HVC was defined as a continuous region covering the lateral occipital complex (LOC), fusiform face area (FFA), and parahippocampal place area (PPA), which were identified using conventional functional localizers (Kourtzi & Kanwisher, 2000; Kanwisher, McDermott, & Chun, 1997; Epstein & Kanwisher, 1998)[33–35].

For brains in MNI space, 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, a source subject's voxels covered by an ROI mask of a target subject were selected as the input to the model of the target subject.

Neural code converter analysis

A neural code converter model for each pair of subjects comprised a set of regularized linear regression models (ridge regression), each trained to predict activities of an individual voxel of one subject (target) from brain activity patterns of another subject (source) using fMRI responses in the visual cortex 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 fMRI activity patterns; λ is the regularization parameter; $||\cdot||$ represents the Frobenius norm.

Each hemisphere of the visual cortex is 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, 1000, 10000, 100000); then further performing grid search in a narrower range containing the rough optimal value (e.g., 2, 4, 6, 8 and 20, 40, 60, 80 around 10); the process is repeated a few times to refine the parameter down to units digit. The regularization parameter was optimized for each source and target subjects pair and was fixed across all voxels. Model performance was evaluated by the identification accuracy based on decoded DNN feature values (see Materials and Methods: "Identification analysis"): we randomly selected 500 units (and also the corresponding feature decoders of 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 an identification accuracy was calculated using the concatenated decoded DNN feature values; we chose the regularization parameter which maximizes the averaged accuracy over five folds. This optimization procedure aims to convert brain activity patterns such that the downstream task is optimized. The main goal is to optimize 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 DNN features of a reconstructed image approximate decoded DNN features.

Hyperalignment analysis

Similar to neural code converter, hyperalignment requires fMRI responses of subjects given an identical sequence of stimuli. We only briefly mention the hyperalignment algorithm; please refer to Haxby et al. (2011) [17] for details. The hyperalignment iteratively estimates a shared space by first aligning two subjects' fMRI responses using Procrustean transformation; if there is any third subject, the subject's fMRI responses are aligned to the mean aligned responses of the first two subjects and so on for other subjects; in a second iteration, each subject's response is aligned to the mean aligned responses obtained in the first iteration; a mean aligned responses is recalculated and is 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 a shared space of more than two subjects, we performed hyperalignment with only two subjects, similar to the neural code converter analysis. Each hemisphere of the visual cortex 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 shared space: a subject's responses were first mapped into the shared space, followed by the inverse mapping into another subject's brain space. We called it "conversion via shared space," in contrast to the "alignment into shared space," where all subjects' responses are brought into the shared space.

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 in Lescroart & Gallant, 2019 [25] (see also, Hsu et al., 2004 [24]). The noise ceiling was obtained by averaging correlation coefficients between measured brain responses,

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

When the noise ceiling of brain activity patterns to an image is estimated, $\mathbf{y}$ is a brain activity pattern to the image, and *N* is the total number of repetitions presenting 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 since the conversion performance cannot be reliably evaluated. Similarly, voxels with 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 99 percentile point in the distribution of the correlations from randomly paired samples or voxels.

Deep neural network (DNN) features

We used the VGG19 deep neural network model (Simonyan & Kisserman, 2014) [26] implemented using the Caffe library (Jia et al., 2014) [36]. This model is pre-trained for the 1,000 classes object recognition task with the images from ImageNet (ILSVRC 2012) [28] (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. All units' outputs in each layer were treated as a vector and used in the identification and reconstruction analyses. The number of units in each layer is: conv1_1 and conv1_2, 3211264; conv2_1 and conv2_2, 1605632; conv3_1, conv3_2, conv3_3 and conv3_4, 802816; conv4_1, conv4_2, conv4_3 and conv4_4, 401408; conv5_1, conv5_2, conv5_3 and conv5_4, 100352; fc6 and fc7, 4096; and fc8, 1000.

We used the AlexNet deep neural network model (Krizhevsky et al. 2012) [29] implemented using the Caffe library to extract DNN features from reconstructed images. This model is pre-trained for the 1,000 class object recognition task with the images from ImageNet (ILSVRC 2012) [28] (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: conv1, 290400; conv2, 186624; conv3 and conv4, 64896; conv5, 43264; fc6 and fc7, 4096; and fc8, 1000.

DNN feature decoding analysis

For each unit of DNN, we trained a ridge linear regression model (DNN feature decoder) that takes an fMRI activity pattern induced by a presented image as input and outputs a feature value computed from the presented image. The ridge regularization parameter was set to 100. The feature value and the voxel values are normalized before training the model. For the decoder of

each DNN unit, a voxel selection procedure was performed before training, in which Pearson correlation coefficients between the sequences of feature values and voxel responses of all voxels were computed, and the top 500 voxels having the highest correlations were selected. The training dataset with the selected voxels was then used to train a decoder to decode the values of an individual unit in the feature vectors of all DNN layers. When testing the trained decoders on the test datasets, the fMRI samples corresponding to the same stimulus were averaged across trials to increase the signal-to-noise ratio of the fMRI signal. For details of the feature decoding, see Horikawa & Kamitani (2017) [6] or Shen et al. (2019) [7].

Brain hierarchy score

The brain hierarchy (BH) score is originally designed to measure the degree to which an artificial neural network is hierarchically similar to the human brain. Here we adopted the decoding based-BH score, the DNN features are decoded from the fMRI signals in one of the five visual areas: V1–V4 and higher visual cortex (HVC). For each unit, the visual area showing the best decoding accuracy is identified and is called the "top" area. Here, the decoding accuracy refers to the mean Pearson correlation coefficient between the predicted and true DNN unit activations sequences over stimuli. The first layer and the last layer are fixed, and three intermediate layers are randomly sampled from the DNN. For each sampling, a Spearman rank correlation coefficient between the layer number and the "top" area number (V1: 0, V2: 1, V3: 2, V4: 3, and HVC:4) across units in the given DNN is calculated. This sampling procedure is repeated 10000 times, and the mean Spearman rank correlation coefficient is taken as the BH score. In this study, conv1_1_abs (absolute value is taken), pool1–5, fc6–8 layers of VGG19 model are used and

random 1000 units are selected from each layer for the BH score computation. For details of the BH score, refer to Nonaka et al. (2021) [27].

Visual image reconstruction analysis

We used an image reconstruction method (deep image reconstruction) proposed in Shen et al. (2019) [7]. 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 optimization problem (Mahendran & Vedaldi, 2015) [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 × height × RGB channels); L is the total number of layers; $\mathbf{\Phi}_l$ is the function which 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 which 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, Dosovitskiy, Yosinski, Brox, & Clune, 2016) [38]. 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 & Brox (2016;

available from https://github.com/dosovits/caffe-fr-chairs; trained model for fc7) [39]. Given decoded features from 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 (200 iterations).

The solution to the above optimization problem is considered as the reconstructed image from the brain activity pattern. Following Shen et al. (2019) [7], natural images are reconstructed with DGN, and the objective function is optimized by stochastic gradient descent with momentum while artificial images are reconstructed without DGN and the objective function is optimized by limited-memory BFGS algorithm (L-BFGS) with 200 iterations (Le et al., 2011; Liu & Nocedal, 1989; Gatys, Ecker, & Bethge, 2016) [40–42].

Identification analysis

Identification analysis is used to evaluate image reconstruction quality. Presented images were identified using either reconstructed image pixels or DNN features extracted from reconstructed images. Before the identification analysis, image pixels or DNN features were reshaped into a one-dimensional vector, and it was termed a feature vector for simplicity. The feature vector obtained from a reconstructed presented image was used to compare the true feature vector of the presented image with a false feature vector of another image. The comparison was counted as correctly identified if the Pearson correlation coefficient between the reconstructed feature vector and true feature vectors was higher than that between the reconstructed and false feature vectors. The same procedure was repeated for each reconstructed presented image until all other images were compared with the true presented image. Identification accuracy for a reconstructed image

was defined as the proportion of correctly identified comparisons among all comparisons. The same identification procedure was repeated for each reconstructed image, and a mean identification accuracy was calculated across all reconstructed images. We used "identification accuracy" to indicate the mean identification accuracy for simplicity unless stated otherwise.

When optimizing the regularization parameters for neural code converters, we used a set of decoded DNN features concatenated from multiple layers as a proxy of the DNN features extracted from 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 identification accuracies that evaluate the performance of neural code converters. The candidate images were a subset of 1,200 images presented in the training image session, depending on which fold of data is 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 image pixel values of reconstructed images or DNN features extracted from reconstructed images as the feature vector in the identification procedure to evaluate the reconstruction quality. Each reconstructed image was tested with all pairs of the images; for example, a reconstructed natural image was compared with 49 pairs, with each comparison consisting of one presented image and another (false) image from the rest of 49 test natural images, resulting in 50 × 49 = 2,450 comparisons. For neat evaluation, we used another neural network of smaller network architecture (AlexNet [28]; available from https://github.com/BVLC/caffe/tree/master/models/bvlc_alexnet; see also Materials and

Methods: "Deep neural network features"), to extract visual features of both reconstructed images and original images, and then performed the identification analysis on them layer by layer.

Pooling data from multiple subjects

First, we pooled data from multiple subjects to train DNN feature decoders. In this analysis, we pooled data among 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 three subjects' converted data (3 × 6,000 samples) in the target brain space. We then trained DNN feature decoders on the pooled data and called the decoders multiple-subject feature decoders, in contrast to the single-subject feature decoders trained with individual data (6,000 samples) in native brain space. The same analysis was repeated four times such that every subject among the four was selected as a target subject.

Second, we used 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 now 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 300 samples from the test subject, resulting in 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 visual images via decoded DNN features.

Statistics

We used analysis of variance (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 used analysis of variance (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 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), and

open data repository (OpenNeuro: https://openneuro.org/datasets/ds001246, https://openneuro.org/datasets/ds003430, https://openneuro.org/datasets/ds001506/versions/1.3.1).

Acknowledgments

The authors would like to thank Jong-Yun Park, Fan Cheng, Shangfeng Jin, Shuntaro C. Aoki, and Mitsuaki Tsukamoto, and Misato Tanaka for helpful comments on the manuscript. The data used in the study were collected using the MRI scanner and related facilities of Kokoro Research Center, Kyoto University.

Author Contributions

Conceptualization: Yukiyasu Kamitani.

Formal analysis: Jun Kai Ho.

Investigation: Tomoyasu Horikawa, Yukiyasu Kamitani.

Methodology: Jun Kai Ho, Tomoyasu Horikawa, Kei Majima.

Project administration: Yukiyasu Kamitani.

Resources: Yukiyasu Kamitani.

Supervision: Yukiyasu Kamitani.

Validation: Jun Kai Ho, Tomoyasu Horikawa, Yukiyasu Kamitani.

Visualization: Jun Kai Ho

Writing - original draft: Jun Kai Ho, Tomoyasu Horikawa, Yukiyasu Kamitani.

Writing - review & editing: Jun Kai Ho, Tomoyasu Horikawa, Kei Majima, Yukiyasu Kamitani.

References

1. 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. https://doi.org/10.1113/jphysiol.1962.sp006837
2. 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. https://doi.org/10.1016/0166-4328(82)90081-X
3. 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 U S A*. 2014; 111: 8619–8624. https://doi.org/10.1073/pnas.1403112111
4. 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. https://doi.org/10.1523/JNEUROSCI.5023-14.2015
5. Güçlü U, van Gerven MAJ. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*. 2017; 145: 329–336. https://doi.org/10.1016/j.neuroimage.2015.12.036
6. Horikawa T, Kamitani Y. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*. 2017; 8: 15037. https://doi.org/10.1038/ncomms15037
7. Shen G, Horikawa T, Majima K, Kamitani Y. Deep image reconstruction from human brain activity. *PLoS Computational Biology*. 2019; 15: 1006633. https://doi.org/10.1371/journal.pcbi.1006633

8. 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. https://doi.org/10.3389/fncom.2019.00021

9. 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. https://doi.org/10.1016/j.neuroimage.2013.03.024

10. 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. https://doi.org/10.1016/j.neuron.2015.06.037

11. Engel SA, Rumelhart DE, Wandell BA, Lee AT, Glover GH, Chichilnisky E-J, et al. fMRI of human visual cortex. *Nature*. 1994; 369: 525. https://doi.org/10.1038/369525a0

12. 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. https://doi.org/10.1126/science.7754376

13. 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. https://doi.org/10.1093/cercor/bhm225

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

15. Van Essen DC. A Population-Average, Landmark- and Surface-based (PALS) atlas of human cerebral cortex. *NeuroImage*. 2005; 28: 635–662. https://doi.org/10.1016/j.neuroimage.2005.06.058

16. 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. https://doi.org/10.1093/cercor/3.2.79

17. 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. https://doi.org/10.1016/j.neuron.2011.08.026

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

19. 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

20. Bilenko NY, Gallant JL. Pyrcca: Regularized kernel canonical correlation analysis in Python and its applications to neuroimaging. *Front Neuroinform*. 2016; 10: 49. https://doi.org/10.3389/fninf.2016.00049

21. 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. https://doi.org/10.1093/cercor/bhw068

22. Van Uden CE, Nastase SA, Connolly AC, Feilong M, Hansen I, Gobbini MI and Haxby JV Modeling Semantic Encoding in a Common Neural Representational Space. *Front Neurosci*. 2018; 12: 437. https://doi.org/10.3389/fnins.2018.00437

23. Horikawa T, Aoki SC, Tsukamoto M, Kamitani Y. Characterization of deep neural network features by decodability from human brain activity. *Scientific Data*. 2019; 6: 190012. https://doi.org/10.1038/sdata.2019.12

24. 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. https://doi.org/10.1088/0954-898X_15_2_002

25. Lescroart MD, Gallant JL. Human Scene-Selective Areas Represent 3D Configurations of Surfaces. *Neuron*. 2019; 101: 178-192. https://doi.org/10.1016/j.neuron.2018.11.004

26. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition; 2014. Preprint. Available from: arXiv:1409.1556v1. Cited 8 Nov 2021.

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

28. Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*. 2009; 248–255. https://doi.org/10.1109/CVPR.2009.5206848

29. 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

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. https://doi.org/10.1093/scan/nsz037

31. 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. https://doi.org/10.1038/s41592-018-0235-4

32. 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. https://doi.org/10.1016/j.media.2007.06.004

33. Kourtzi Z, Kanwisher N. Cortical Regions Involved in Perceiving Object Shape. *J Neurosci*. 2000; 20: 3310–8. https://doi.org/10.1523/JNEUROSCI.20-09-03310.2000

34. 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. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997

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

36. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, et al. Caffe: Convolutional Architecture for Fast Feature Embedding; 2014. Preprint. Available from: arXiv:1408.5093. Cited 8 Nov 2021.

37. Mahendran A, Vedaldi A. Understanding deep image representations by inverting them. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*. 2015; 5188–5196. https://doi.org/10.1109/CVPR. 2015.7299155

38. 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

39. 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

40. Le QV, Ngiam J, Coates A, Lahiri A, Prochnow B, Ng AY. On optimization methods for deep learning. *Proc Int Conf Mach Learn* (Bellevue, Washington, USA). 2011; 265–272.

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

42. Gatys LA, Ecker AS, Bethge M. Image style transfer using convolutional neural networks. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*. 2016; 2414–2423. https://doi.org/10.1109/CVPR.2016.265

Supporting information captions

S1 Fig. Comparisons of neural code converter, alignment into shared space, and conversion via shared space.
(A) Illustration of neural code converter, alignment into shared space, and conversion via shared space. Both neural code converter and hyperalignment were trained with 2,400 training samples from a pair of source and target subjects. Shared space is estimated through the conventional hyperalignment algorithm (see Materials and Methods: "Hyperalignment analysis"). Neural code converter directly converted a source subject's responses to a target subject brain space (left). Alignment into shared space aligned both source and target subjects' responses into a shared space (middle). Conversion via shared space first mapped a source subject's responses into a shared space, followed by an inverse mapping into the target subject's brain space (right).
(B) Evaluation of alignment into shared space. There were no visual area labels because each dimension in the shared space did not represent a single voxel with a cortical location. Ten shared spaces were estimated for ten pairs of source-target subjects. Pearson's correlation calculations were computed using responses in the shared spaces and normalized by noise ceilings (pooled across ten pairs of alignments; each circle represents mean correlation coefficients for each pair of alignments).
(C) Evaluation of conversions via shared space (conventions follow Fig 2A).
(D) Inter-individual DNN feature decoding accuracies using neural code converter, alignment into shared space, and conversion via shared space. The inter-individual analysis using neural code converter and conversion via shared space shared the same DNN feature decoders. For fMRI responses aligned to a shared space, the DNN feature decoders were trained with one subject's responses and were tested with the other subject's responses (error bars, 95% confidence interval [C.I.] across 20 pairs of subjects).

S2 Fig. The cortical map of the effects of source area exclusion, target S2–S5. 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 test natural image dataset (performance drops were averaged across four source subjects for a single target subject). Only voxels that generate reliable responses were shown.

S3 Fig. Reconstructed natural images across subject pairs. For each image, the diagonal images in each block are the reconstructed images obtained in the "Within" condition; the off-diagonal images are those obtained in the "Across - functional [converter]" condition with the converters trained on 2,400 training samples.

S4 Fig. Reconstructed artificial images across subject pairs. For each image, the diagonal images in each block are the reconstructed images obtained in the "Within" condition, while the off-diagonal images are those obtained in the "Across - functional [converter]" condition with the

converters trained on 2,400 training samples.

S5 Fig. Reconstructed images using different repetitions of samples. A varying number of fMRI samples corresponding to an image label were averaged and were reconstructed into an image.

S6 Fig. Evaluations of subarea-wise neural code converters. Distributions of normalized pattern or profile correlation coefficients are shown for VC and visual subareas. Each horizontal black dash indicates the mean value; each circle represents the correlation coefficients of a subject pair.

S7 Fig. Identification accuracy based on pixel values and extracted DNN feature values of artificial images.
(A) Reconstructed artificial images via converters trained on varying amounts of data. All reconstructed images were produced from the same source subject (source: S2, target: S5)
(B) Identification accuracies via converters trained on varying amounts of data. The converters were trained on a varying amount of data. The identification analysis was performed using pixel values and extracted DNN feature values (AlexNet) of the reconstructed artificial images (error bars, 95% C.I. across 20 subject pairs for whole VC and subarea-wise conversion conditions; dotted lines, chance level, 50%).

S8 Fig. Evaluation of multiple subject feature decoders using artificial images dataset.
(A) DNN feature decoding accuracy obtained via multiple- and single-subject feature decoders. Multiple-subject feature decoders trained on data pooled from four subjects; single-subject feature decoders trained on only a single subject's data. The accuracies were obtained from fMRI data corresponding to the test artificial image session (error bars, 95% C.I. across four target brain spaces).
(B) Reconstructed artificial images obtained via multiple- and single-subject feature decoders.
(C) Identification accuracies obtained from reconstructed artificial images. The identification analysis was performed using pixel values and extracted DNN feature values of the reconstructed artificial images obtained via multiple- and single-subject feature decoders (error bar, 95% C.I. across four target brain spaces; dotted lines, chance level, 50%).

S9 Fig. Effect of pooling multiple subject data into different target brain spaces.
Identification accuracies were computed from reconstructed natural images obtained via feature decoders trained on pooled fMRI data in different target brain spaces (dotted lines, chance level, 50%). Multiple-subject feature decoders trained on data pooled from four subjects; single-subject feature decoders trained on only a single subject's data.