Self-Supervised Natural Image Reconstruction and Rich Semantic Classification from Brain Activity

Guy Gaziv†, 1, Roman Beliy†, 1, Niv Granot†, 1, Assaf Hoogi[1], Francesca Strappini[2], Tal Golan[3], Michal Irani[1]

(1) Dept. of Computer Science and Applied Math, Weizmann Institute of Science, Rehovot, Israel.
(2) Dept. of Neurobiology, Weizmann Institute of Science, Rehovot, Israel.
(3) Zuckerman Institute, Columbia University, New York, NY, USA.

*** michal.irani@weizmann.ac.il**
† The authors contributed equally to this work.

Abstract

Reconstructing natural images and decoding their semantic category from fMRI brain recordings is challenging. Acquiring sufficient pairs (image,fMRI) that span the huge space of natural images is prohibitive. We present a novel *self-supervised* approach for fMRI-to-image reconstruction and classification that goes well beyond the scarce paired data. By imposing cycle consistency, we train our image reconstruction deep neural network on many "unpaired" data: a plethora of natural images without fMRI recordings (from many novel categories), and fMRI recordings without images. Combining high-level perceptual objectives with self-supervision on unpaired data results in a leap improvement over top existing methods, achieving: (i) Unprecedented image-reconstruction from fMRI of never-before-seen images (evaluated by image metrics and human testing); (ii) Large-scale semantic classification (1000 diverse classes) of categories that are never-before-seen during network training. *Such large-scale (1000-way) semantic classification capabilities from fMRI recordings have never been demonstrated before.* Finally, we provide evidence for the biological plausibility of our learned model. [1]

Introduction

Natural images span a vastly rich visual and semantic space that humans are experts at processing and recognizing. The inverse problem addresses the task of decoding images seen by a person and their semantic categories, directly from brain activity (Fig 1a). This task is a cornerstone towards decoding the contents of dreams and mental imagery, as well as a potential basis for clinical communication prostheses. In the image reconstruction task, one attempts to decode natural images which were observed by a human subject from the induced brain activity captured by functional magnetic resonance imaging (fMRI). To learn the mapping between fMRI and image representation, typical fMRI datasets provide many pairs of images and their corresponding fMRI responses, henceforth "paired" data. The goal is to learn an fMRI-to-image decoder which generalizes well to reconstructing images from novel "test-fMRI", fMRI response induced by novel images from totally different semantic categories than those in the training data (referred to as "test-images"). Moreover, a complementary challenge to reconstructing the underlying image is also to decode its semantic category. *However, the shortage of "paired" training data limits the generalization power of today's fMRI decoders.* The number of obtainable image-fMRI pairs is bounded by the limited time a human can spend in an MRI scanner. This results also in a limited number of semantic categories associated with fMRI data. Accordingly, most datasets provide only up to a few thousands of such pairs. Such limited data cannot span the huge space of natural images and their semantic categories, nor the space of their fMRI recordings. Moreover, the poor spatio-temporal resolution of fMRI signals, as well as their low

[1] We will make our code publicly available upon publication.

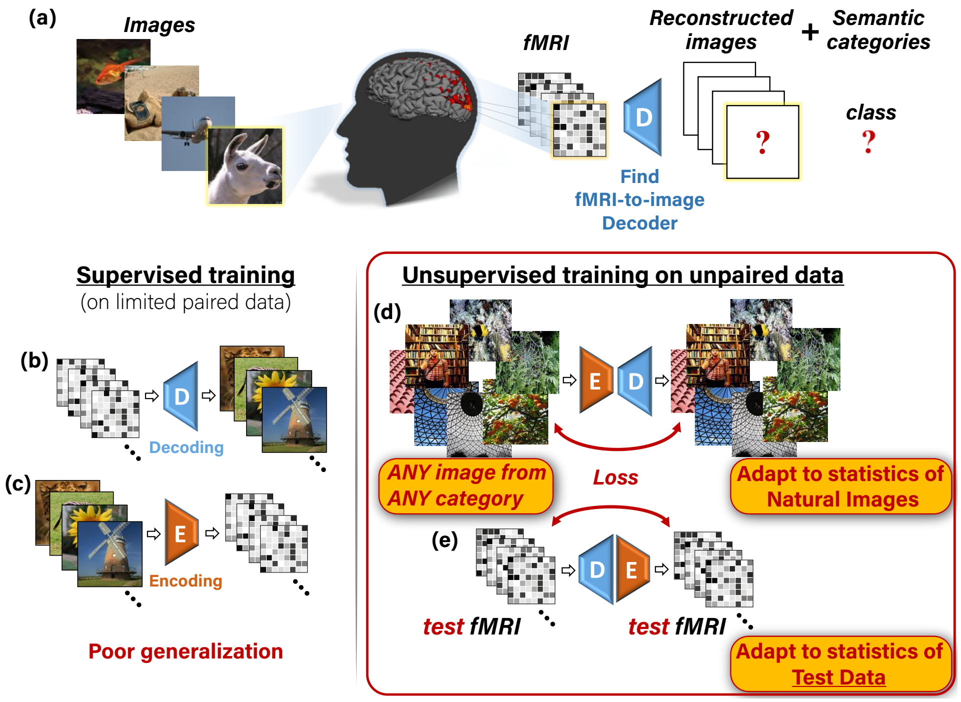

Figure 1. Our self-supervised approach. *(**a**) The task: reconstructing images and classifying their semantic category from evoked brain activity, recorded via fMRI. (**b**), (**c**) Supervised training for decoding (**b**) and encoding (**c**) using limited training pairs. This gives rise to poor generalization. (**d**), (**e**) Illustration of our added self-supervision, which enables training on "unpaired images" (any natural image with no fMRI recording – (**d**)), and on the "unpaired fMRI" (fMRI data without any corresponding images – (**e**)). This self-supervision allows adapting the decoder to the statistics of natural images and of the test-fMRI despite not having any information about the test images.*

Signal-to-Noise Ratio (SNR), reduce the reliability of the already scarce paired training data. Furthermore, the statistical properties of fMRI samples in the test-set are often *different* than those in the train-set, specifically in their SNR. The SNR discrepancy stems from averaging a different number of repeated recordings per image in the train-set and test-set (which is typical of many fMRI datasets). This disparity further challenges the generalization capacity of current decoding methods.

Reconstructing natural images from fMRI was approached by a number of methods, which can broadly be classified into three families: (i) Linear regression between fMRI data and handcrafted image-features (e.g., Gabor wavelets) [1–3], (ii) Linear regression between fMRI data and deep (CNN-based) image-features (e.g., using pretrained AlexNet) [4–7], or latent spaces of pretrained generative models [8–11], and (iii) End-to-end Deep Learning [12–15]. To our best knowledge, methods [6] and [13] are the current state-of-the-art in this field. All these methods inherently rely on the available "paired" data to train their decoder (pairs of images and their corresponding fMRI responses). Such purely supervised models, when trained on limited data, are prone to overfitting, which leads to poor generalization to new test-data (fMRI response evoked by new images).

Prior work on semantic classification of fMRI recordings induced by natural-images, can be characterized as two families: (i) Classifying new images from previously seen categories, and (ii) Classifying new images from novel never-before-seen categories. In the first, the categories to-be-decoded (of the test data) are represented in the training data [16–20]. This widely-explored family is limited to decode only the few and typically coarse classes which are represented in the limited "paired" data used for decoder training. The second family, which was introduced in [21, 22], addresses the much more challenging case, where the test-categories are *novel*, namely, not directly represented in the training data. Under this setting, decoding novel, rich, and fine-grained semantic

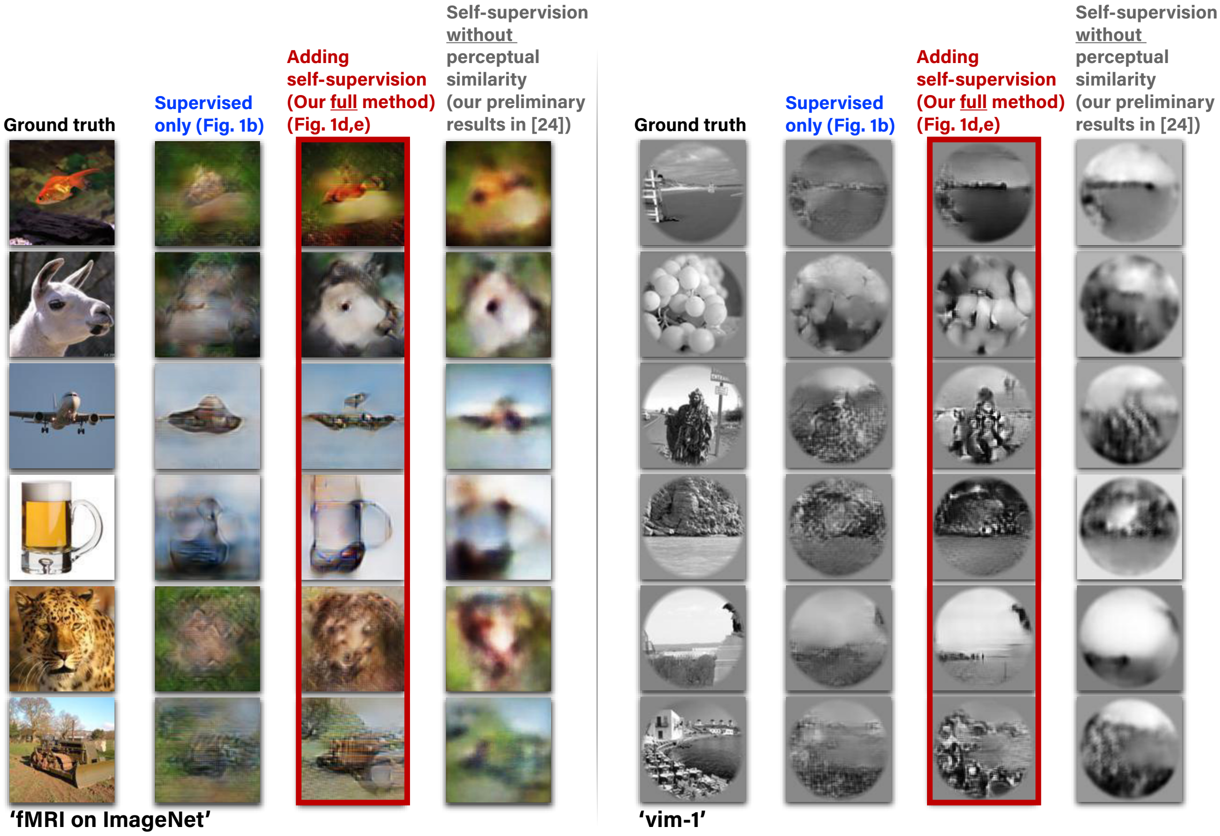
Ground truth
Supervised only (Fig. 1b)
Adding self-supervision (Our full method) (Fig. 1d,e)
Self-supervision without perceptual similarity (our preliminary results in [24])
Ground truth
Supervised only (Fig. 1b)
Adding self-supervision (Our full method) (Fig. 1d,e)
Self-supervision without perceptual similarity (our preliminary results in [24])
'fMRI on ImageNet'
'vim-1'

Figure 2. Adding unsupervised training on unpaired data together with perceptual criteria improves reconstruction.

(Left to Right): • *The images presented to the human subjects.* • *Reconstruction using the training pairs only (Fig 1b).* • *Reconstruction when adding self-supervised training on unpaired data (Fig 1d,e), and also adding high-level perceptual criteria to the decoder and other important improvements.* • *Our preliminary results [24] without using the perceptual criteria and other important improvements presented here. Example results are shown for two fMRI datasets: 'fMRI on ImageNet' [21] and 'vim-1' [1].*

categories (e.g., ImageNet [23]) remained a difficult task because of the narrow semantic coverage spanned by the limited paired training data.

To cope with this data limitation, recent approaches [4, 17, 19–22] harnessed a *pretrained* and semantically separable embedding. In this approach voxel responses are linearly mapped to *higher-level* feature representation of an image classification network. Once mapped, categorization is achieved by either (i) forward-propagating the decoded representation to the classification layer [4], or by (ii) nearest-neighbor classification against a gallery of category representatives, which are the mean feature representations of many natural images from that category [21]. While these methods benefited from the wide-coverage of their semantic representation (which stems from pretrained image features independently of the fMRI data), their method remained **supervised** in essence. This is because their training of the mapping from fMRI-to-feature representation relies solely on the limited "paired" training data. Consequently they are prone to poor generalization and are limited by the poor category coverage of the "paired" training data.

We present a new approach to overcome the above-mentioned limitations and inherent lack of training data, simultaneously for both tasks – image reconstruction and large-scale semantic classification. We achieve this by introducing *self-supervised training on unpaired data (images without fMRI recordings, and fMRI recordings without images)*. Our approach is illustrated in Fig 1. We train two types of networks: an Encoder E, to map natural images to their corresponding fMRI response, and a Decoder D, to map fMRI recordings to their corresponding images. Concatenating those two networks back-to-back, E-D, yields a combined network whose input and output are the same image (Fig 1d). ***This allows for unsupervised training on unpaired images*** (i.e., images without fMRI recordings, e.g., 50,000 natural images from 1000 semantic categories in our experiments). Such self-supervision adapts the network to the statistics of novel images and their novel categories (categories not represented in the "paired" training data or in the "unpaired" natural images). Moreover, concatenating our two networks the other way around, D-E, yields a combined network with the same shared weights as E-D, but whose input and output are now an fMRI recording (Fig 1e). ***This allows unsupervised training on unpaired fMRI samples***. Specifically, those unpaired fMRI samples can be legitimately drawn from the test-fMRI cohort, while their corresponding images ("test-images") are unknown and never used at any stage of the training (Fig 1e). Training on these unpaired test-fMRI (without their images) enables to adapt the network to the statistics of the new (unpaired) test-data. In particular, it addresses the discrepancy between the statistics of the training data, and that of the test data.

Unsupervised training on unpaired natural images was also recently proposed in [13], where they used these images to produce additional surrogate fMRI-data to train their model. However, this was never addressed in the context of semantic decoding, and does not help to adapt the network to the statistics of the new test-fMRI. To the best of our knowledge we are the first to present classification of semantic categories that are never-before-seen during training at a large-scale of 1000-way – detecting the correct class out of more than 1000 rich classes. We show that our self-supervised approach, combined with perceptual similarity criteria, gives rise to a dramatic improvement in both tasks: reconstruction of novel images from fMRI and decoding their semantic categories, despite the very scarce fMRI-based training data.

A preliminary version of our self-supervised approach and partial results (in the context of image reconstruction only) were previously presented in a conference proceeding [24]. However, here we present our complete and advanced reconstruction algorithm that has undergone major extensions, enabling a leap improvement in the image reconstruction quality over our previous method [24] (as demonstrated in Fig 2), and furthermore, enabling *new semantic classification capabilities.* As an aid for readers familiar with our previous report [24], we list the four major extensions introduced in the present paper: (i) We introduced two significant improvements to our algorithm: Adding high-level perceptual criteria [25] on the reconstructed (in contrast with optimizing Mean-Square-Error loss and on low level features alone), and increasing the expressiveness of the Encoder architecture to include higher-level "semantic" representation by using multiple levels of the pretrained VGG network (in contrast with a single readout layer before). Our present algorithm provides state-of-the-art reconstructions compared to leading existing methods to-date, as evident through image-metric-based as well as extensive human behavioral evaluations. (ii) We extended the self-supervised approach to allow also for semantic classification of reconstructed images. This classification relies on and demonstrates the fidelity of the reconstructions. (iii) We analyze the contributions of different visual cortex areas to the resulting image reconstructions, and (iv) We evaluate the biological plausibility of the learned models.

Our contributions are therefore several-fold:

- A self-supervised approach for simultaneous image reconstructing and semantic category decoding, which can handle the inherent lack of image-fMRI training data.
- Unprecedented state-of-the art image-reconstruction quality from fMRI of never-before-seen images (from never-before-seen semantic categories).
- Large-scale semantic classification (1000+ rich classes) of *never-before-seen semantic categories.* To the best of our knowledge, such large-scale semantic classification capabilities from fMRI data has never been demonstrated before.
- We provide analyses showing predominance of early visual areas in reconstruction quality, and biologically plausible receptive field formation in our models.

Results

In what follows we provide a high-level overview of our approach and its key components for (i) self-supervised image reconstruction, and (ii) classification to semantic categories. Following the overview we detail on our experimental results using our approach. The technical details are provided in the Methods section.

Self-supervised image reconstruction from brain activity – an Overview

The essence of our approach is to enrich the scarce paired image-fMRI training data with new types of easily accessible data *that are not paired*, from the image or the fMRI domains. These new data types include natural images for which there are no fMRI recordings, and fMRI recordings for which we do not have the underlying natural image ("unpaired" data). This type of training is enabled by imposing cycle-consistency on the unpaired data, using two networks, which learn two inverse mappings: from images to fMRI (encoding) and vice versa – from fMRI to images (decoding).

Our training consists of Encoder training followed by Decoder training, which we define in the two phases illustrated in Fig 3. In the first phase, we apply supervised training of the Encoder E alone. We train it to predict the fMRI responses of input images using the image-fMRI training pairs (Fig 3a). In the second phase, we use the pretrained Encoder (from the first phase) and train the Decoder D, keeping the weights of E fixed (Fig 3b). D is trained using both the paired and the unpaired data, simultaneously. Here, each training batch consists of three types of training data: (i) image-fMRI pairs from the training set (Fig 1b), (ii) unpaired natural images (with no fMRI, Fig 1d), and (iii) unpaired fMRI (with no images, Fig 1e).

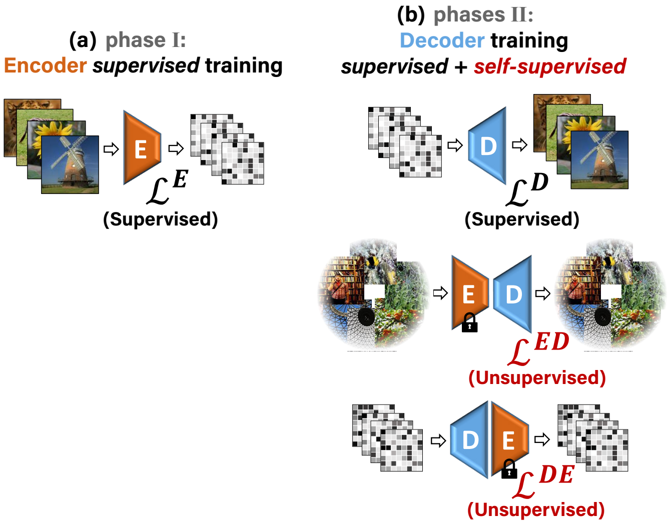

Figure 3. Training phases. ***(a)*** *The first training phase: Supervised training of the Encoder with {Image, fMRI} pairs.* ***(b)*** *Second phase: Training the Decoder with 3 types of data simultaneously: {Image, fMRI} pairs (supervised examples), unpaired natural images (self-supervision), and unpaired test-fMRI (self-supervision). Note that the test-images are never used for training. The pretrained Encoder from the first training phase is kept fixed in the second phase.*

🔒 = fixed weights

Training on *unpaired natural images* (without fMRI) allows to augment the training with data from a much richer semantic space than the one spanned by the paired training data alone. Specifically, we draw the unpaired images from a large *external* database of 49K images from 980 ImageNet ("ILSVRC") classes, which are mutually exclusive not only to the test-images contained the fMRI (paired) dataset, but also to their underlying test-classes. In principle, for optimal networks E and D, the combined $E - D$ network should yield an output image which is identical to its input image. This should hold for any natural image (regardless if an fMRI was ever recorded for it). Importantly, our image reconstruction losses (in $\mathcal{L}_{ED}$ and also in $\mathcal{L}_D$, Fig 3) require for the reconstructed images to be similar to the original image not only at a pixel level. We further require these two images to be ***perceptually similar***.

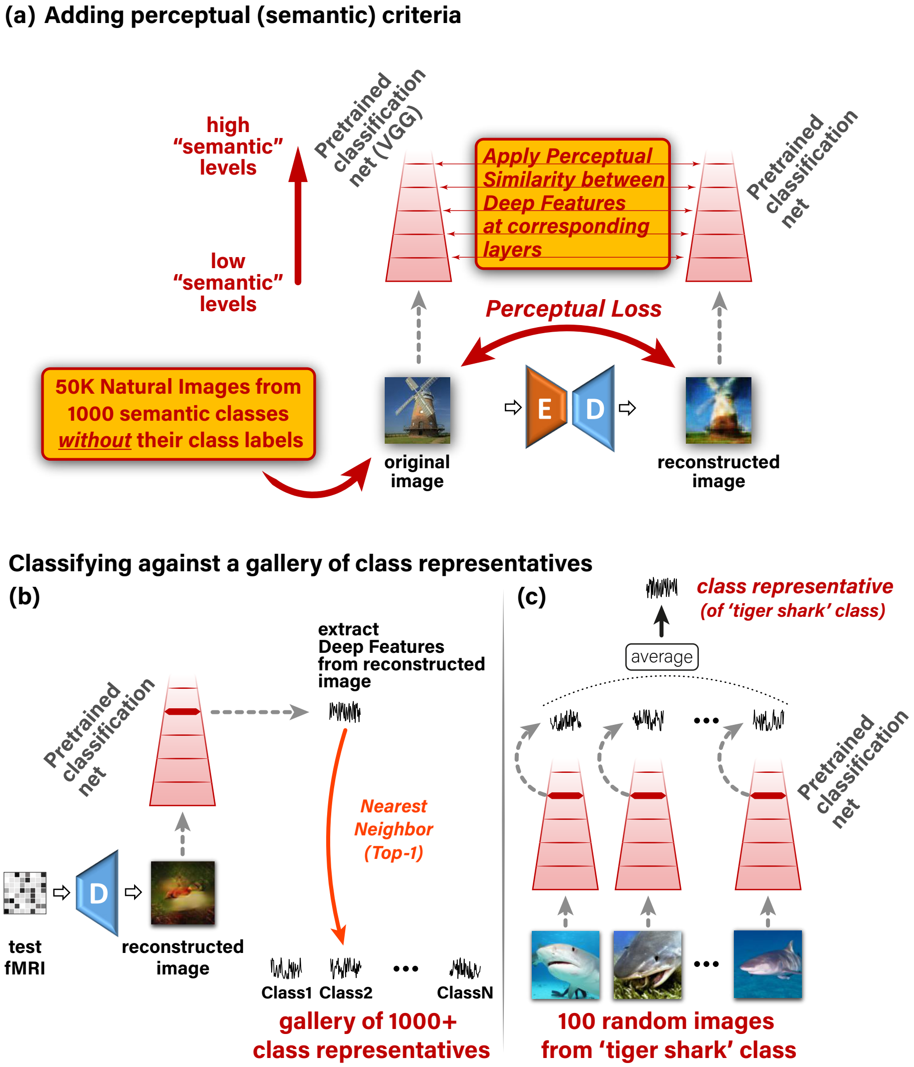
(a) Adding perceptual (semantic) criteria
high "semantic" levels
low "semantic" levels
Pretrained classification net (VGG)
Apply Perceptual Similarity between Deep Features at corresponding layers
Pretrained classification net
Perceptual Loss
50K Natural Images from 1000 semantic classes without their class labels
E
D
original image
reconstructed image
Classifying against a gallery of class representatives
(b)
Pretrained classification net
extract Deep Features from reconstructed image
Nearest Neighbor (Top-1)
test fMRI
D
reconstructed image
Class1 Class2 ••• ClassN
gallery of 1000+ class representatives
(c)
class representative (of 'tiger shark' class)
average
Pretrained classification net
100 random images from 'tiger shark' class

Figure 4. Adding high-level perceptual criteria improves reconstruction accuracy and enables large-scale semantic classification. *(a) Imposing Perceptual Similarity on the reconstructed image at the output of the Decoder when training on unpaired natural images (without fMRI and without any class labels) from many novel semantic classes. This adapts the Decoder to a significantly broader semantic space despite not having any explicit semantic supervision. **(b)** To classify a reconstructed image to its novel semantic class we extract Deep Features using a pretrained classification network, and follow a nearest-neighbor class-centroid approach against a large-scale gallery of 1000+ ImageNet classes. **(c)** We define class representatives as the mean-embedding of many same-class images [21].*

The Perceptual Similarity was first introduced in [25] as a metric which highly correlates with *human* image-similarity perception, and involves a broad range of visual feature representation levels. Our Perceptual Similarity loss is illustrated in Fig 4. We apply a pre-trained VGG classification network (a network which was trained for the task of object recognition from images [26]), on the reconstructed and the original images. We then impose similarity between their corresponding deep-image-features, extracted from multiple deep layers of VGG. *Using this metric enables to learn low-level to high-level "semantic" information from the broad semantic space, which is spanned by the external database of 'unpaired' images and their classes.* Importantly, the class labels of the unpaired images (or the paired images) are never used in our training process, hence may be unknown. Introducing the Perceptual Similarity gives rise to a leap improvement in the reconstruction quality compared to any previous method (including our own previous method [24] – see Fig 2, as well as others – see Fig 6). It provides a dramatic improvement in detail level and perceptual interpretability of the reconstructed images. Our new approach further enables large-scale image classification to rich novel semantic categories.

Self-supervised image classification to semantic categories – an Overview

Our new self-supervised perceptual approach extends well beyond the task of image reconstruction. It further allows for large scale semantic classification of fMRI data. We present classification of fMRI data against a gallery of more than 1000 rich image classes, in a challenging 1000-way classification task (see Fig 7). Scaling semantic classification of fMRI data to 1000-way, with promising results, has never been demonstrated before.

Our classification approach is based on our self-supervised perceptual reconstruction method described above. We use our perceptually trained Decoder to reconstruct the test-images from their test-fMRI. We then classify the reconstructed images against 1000+ rich ImageNet semantic classes. Fig 4bc shows our classification approach. To classify a reconstructed image to its novel semantic class we match a “Deep-Feature signature” extracted from the reconstructed image, against ”class-representative Deep-Feature signatures” (one per class), in a gallery of 1000+ semantic categories, which also include the 50 *novel* test-classes. More specifically: (i) We extract Deep Features from the reconstructed image at an intermediate level of a pretrained classification network (Fig 4b), (ii) Following [21], for each class in the gallery, we compute a single ”class representative” using 100 randomly sampled images from that class. The class representative is defined as the average Deep Features (centroid) of those 100 randomly sampled images from that class (see Fig 4c). (iii) We compute the correlation between the Deep Features extracted from the reconstructed image and each of the 1000+ class representatives. These yield 1000+ “semantic similarity” scores (Fig 4b). Ranking the gallery classes according to these similarity scores (for each reconstructed image) provides the basis for semantic classification at any desired ‘Top-X’ accuracy level (Fig 7). Specifically, the classification is marked ‘correct’ when the ground truth category matches the nearest neighbor gallery category (‘Top-1’) or when it is among the five nearest neighbors (‘Top-5’). The location of the ground-truth class within the sorted list of 1000+ classes further provides a ”rank score” for evaluating our classification accuracy (see Table 1).

This classification approach greatly benefits from our self-supervised perceptual approach, which enables to train on *additional unpaired images from arbitrarily many novel semantic categories* (Fig 1b). This allows to adapt the Decoder to a much richer (practically unlimited) semantic coverage in a **completely category-free way**, namely without any explicit semantic supervision. Therefore the key component which promotes capturing semantic similarities is the Perceptual Similarity metric, which involves higher-level ”semantic” criteria. This type of non-specific semantic supervision enables our method to generalize well to new never-before-seen semantic classes – classes which are neither contained in the paired training data, nor in the unpaired external images.

To test the feasibility of our approach we experimented with two publicly available (and very different) benchmark fMRI datasets: (i) fMRI on ImageNet [21], and (ii) vim-1 [1]. These datasets provide elicited fMRI recordings of human subjects paired with their corresponding underlying natural images. In both datasets, subjects were instructed to fixate at the center of the images.

Reconstructing images from fMRI – Detailed Results & Comparisons

Fig 2 shows our results with the proposed method, which includes the combined supervised and self-supervised training with perceptual criteria. These results (in red frames – 3rd column) are contrasted with the results obtainable when using supervised training only (e.g., the 1200 paired training examples of the ‘fMRI on ImageNet’ dataset – 2nd column), as well as when not using perceptual criteria (4th column). All the displayed images were reconstructed from the dataset’s test-fMRI cohort (fMRI of new images from never-before-seen semantic categories). The red-framed

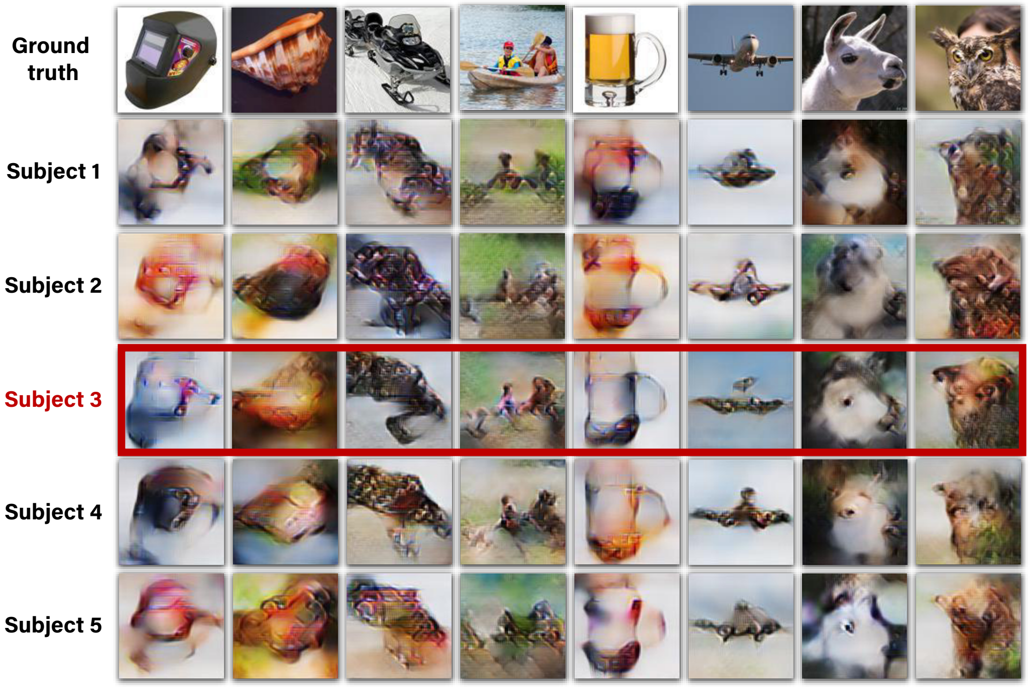

Figure 5. Reconstructions for all five subjects in 'fMRI on ImageNet' [21]. *Reconstructed images when using the full method which includes training on unpaired data (1d,e). Reconstruction quality varies across subjects, depending on noise-ceiling/SNR of subjects' data (voxel median noise ceiling for subjects 1-5:* $0.56, 0.57, 0.73, 0.68, 0.58$*). Subject 3 (in the dataset), which is framed above in red, is the subject of focus in the remaining parts of this paper unless remarked otherwise.*

images show many faithfully-reconstructed shapes, textures, and colors, which depict recognizable scenes and objects. In contrast, using the supervised objective alone led to reconstructions that were considerably less recognizable (2nd column), or less perceptually understandable (4th column). The reconstructions of the entire test cohort (50 images in the 'fMRI on ImageNet' dataset), and ablation studies analyzing the contribution of adding training on unpaired images and/or unpaired fMRI can be found in the Supplementary-Material.

To verify that our method can successfully be applied to different subjects, Fig 5 shows the reconstructions for all five subjects in the 'fMRI on ImageNet' dataset. Note that using fMRI data of different subjects give rise to varying quality of reconstruction as driven by the varied SNR in the subjects' fMRI data. Nevertheless, clear and common identifying markers of the ground truth image appear across all subjects. The red frame indicates the results for the best subject (Subject 3 in the dataset) which is the subject of focus in the remaining parts of this paper (unless mentioned otherwise). We compared our reconstruction results against the two leading methods: Shen et al. [6]) and St-Yves et al. [13] – each on its relevant dataset. Fig 6a,b compares the results of our method with those two methods (both of which are deep-learning GAN-based methods). Visual comparison of [6, 13] with our method (Fig 6a,b) highlights that despite their natural-like visual appearance, the reconstructed images of [6, 13] are often not faithful to the underlying ground truth image.

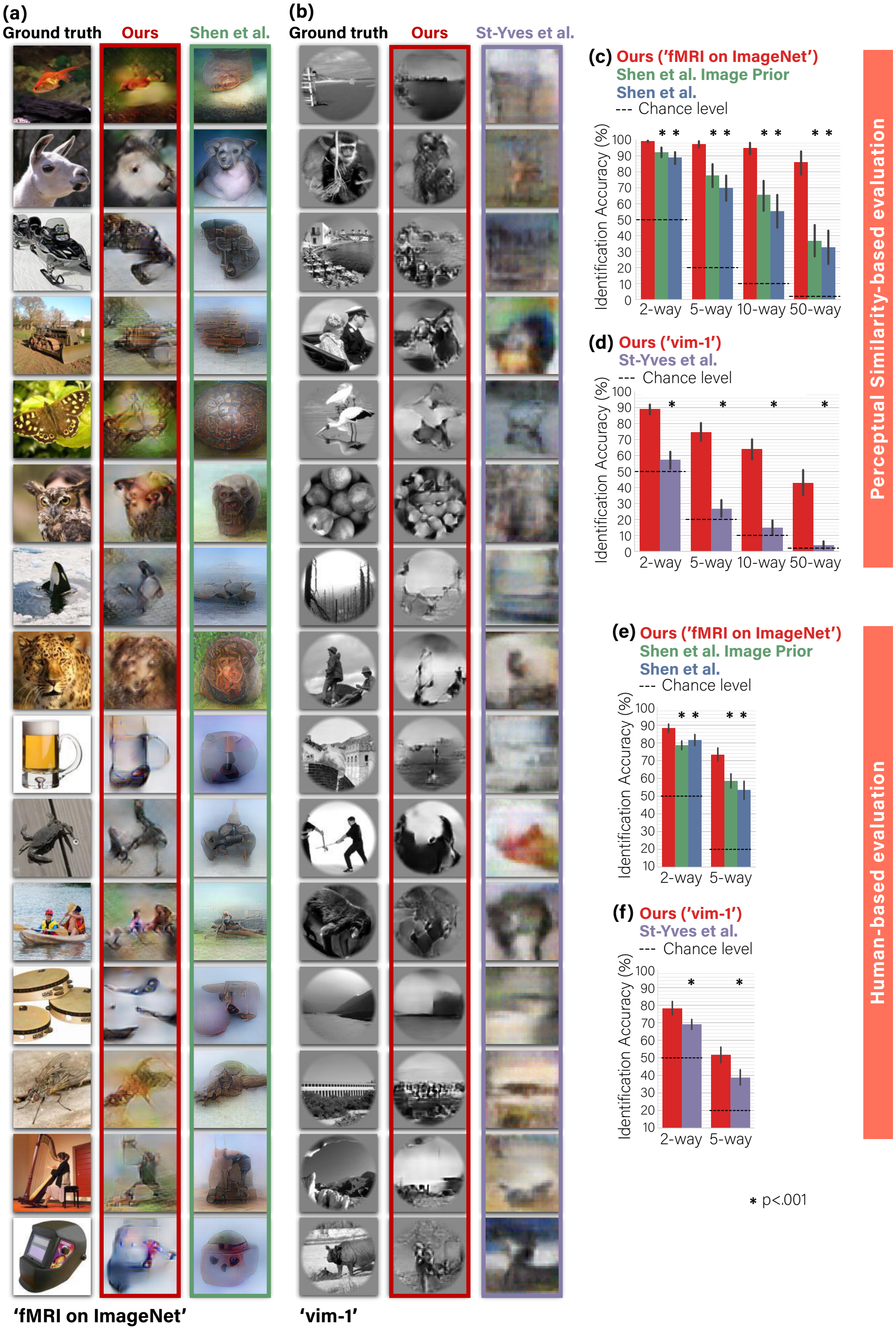

Figure 6. Comparison of image-reconstruction with state-of-the-art methods. *(a), (b) Visual comparison with [6, 13] – each compared on its relevant dataset. Our method reconstructs shapes, details and global layout in images better than the leading methods.* ***(c), (d)*** *Quantitative comparisons of identification accuracy (per method) in n-way identification task according to Perceptual Similarity metric (see text for details).* ***(e), (f)*** *n-way identification responses of human raters via Mechanical Turk. Our self-supervised approach significantly outperforms all baseline methods on two datasets and across n-way difficulty levels by both types of experiments – image-metric-based and behavioral human-based (Wilcoxon, $N = 50, 120$ for panels* ***(c), (d)****; Mann-Whitney, $N = 45$ for panels* ***(e), (f)****). 95% Confidence Intervals by bootstrap shown on charts.*

We further report quantitative comparisons, both by image-metric-based evaluation and by human behavioral evaluation. Our quantitative comparisons were performed on the top-SNR subject from each dataset (Subject 3 in 'fMRI on ImageNet' [21]; Subject 1 in [1]). Our quantitative evaluations are based on an n-way identification task [6, 7, 9, 14]. Namely, each reconstructed image is compared against n candidate images (the ground truth image, and $(n-1)$ other randomly selected images), and the goal is to identify its ground truth. We considered two identification methods under this task:

(i) **Image-metric-based** identification performed using the Perceptual Similarity metric [25] (between the reconstructed image and the candidate image). Each reconstructed image is compared against n images, and the nearest-neighbor candidate image under this metric was determined to be the identified 'correct' image. Panels 6c,d show the correct-identification rate (for each method separately) for n-way identification tasks for $n = 2, 5, 10, 50$. We evaluate our method and two variants of the method of [6] on the 'fMRI on ImageNet' benchmark dataset (Fig 6c). In 2-way identification task our method scored accuracy of 99.3% ($SEM^2 = 0.3\%$, $N = 50$), outperforming both variants of [6] by a margin of 7-49% across all task difficulty levels ($n = 2, 5, 10, 50$). We repeated the analysis for 'vim-1' fMRI dataset (Fig 6d), where our method scored accuracy of 89.2% ($SEM = 1.6\%$, $N = 120$) (2-way task), outperforming the method from [13] by a large margin of 32-49% across the same difficulty levels. Particularly in the challenging 50-way task our method achieved striking leaps: a relative improvement of at least 134% (more than x2 prediction accuracy) in 'fMRI on ImageNet', and 1051% improvement (more than x11 better prediction accuracy) in 'vim-1'. Importantly, the statistical power of these finds generalizes beyond the specific 50 test examples (image-fMRI) in 'fMRI on ImageNet', or the 120 test examples in 'vim-1'. (ii) **Human-based** identification. Panels 6e,f show reconstruction evaluation results when repeating the same quantitative comparison approach, but this time outsourcing the n-way identification task ($n = 2, 5$) to random human raters. We used Mechanical Turk to launch surveys to new 45 raters for each evaluated method. Our method scored 88.4% ($SEM = 1.1\%$, $N = 45$) and 78.2% ($SEM = 1.8\%$, $N = 45$) in a 2-way identification task on 'fMRI on ImageNet' and 'vim-1' respectively; Scaling the task difficulty to 5-way, our method scored 73.4% ($SEM = 1.9\%$, $N = 45$) and 51.8% ($SEM = 2.2\%$, $N = 45$). Overall our method significantly outperformed the previous methods, on both datasets, and across difficulty levels by a margin of at least 6.5% (Mann-Whitney test, $N = 45$, $p < .001$).

Notably, the identification accuracy of each reconstructed image *when using the image-metric for evaluation*, mostly depended on the reconstruction quality of the specific image, and was robust to randomizing the selection of the (non-ground-truth) candidate images. Furthermore, the choice of candidate images *in the human-based evaluation* was fixed across the raters and the methods we compared. Therefore, while the two types of evaluations (image-metric and behavioral) consider a seemingly similar n-way identification task, *they are not directly comparable.* Additionally, note that they suggest different statistical generalization insights – generalization beyond the specific set of test examples, when using the image-metric, and generalization beyond the specific pool of human raters, in the behavioral evaluations. Overall, our method significantly outperforms state-of-the-art methods by a large margin in both image-metric-based and human-based evaluations.

Decoding rich novel semantic categories of reconstructed images

The benefit of our self-supervised approach extends beyond the task of image reconstruction. It introduces significant gains in the task of semantic classification of the reconstructed images to their novel semantic categories (i.e., categories/classes never seen during training – were neither represented in the 'paired' training set not in the 'unpaired' external images). For the classification task, we consider the 'fMRI on ImageNet' dataset [21], whose train-set contains 1200 images from 150 ImageNet classes. Its test-set contains 50 images from *disjoint* ImageNet classes. The unpaired external images used for the self-supervised training of our network, are drawn from 1000 ImageNet classes. Notably,

[2] Standard Error of the Mean.

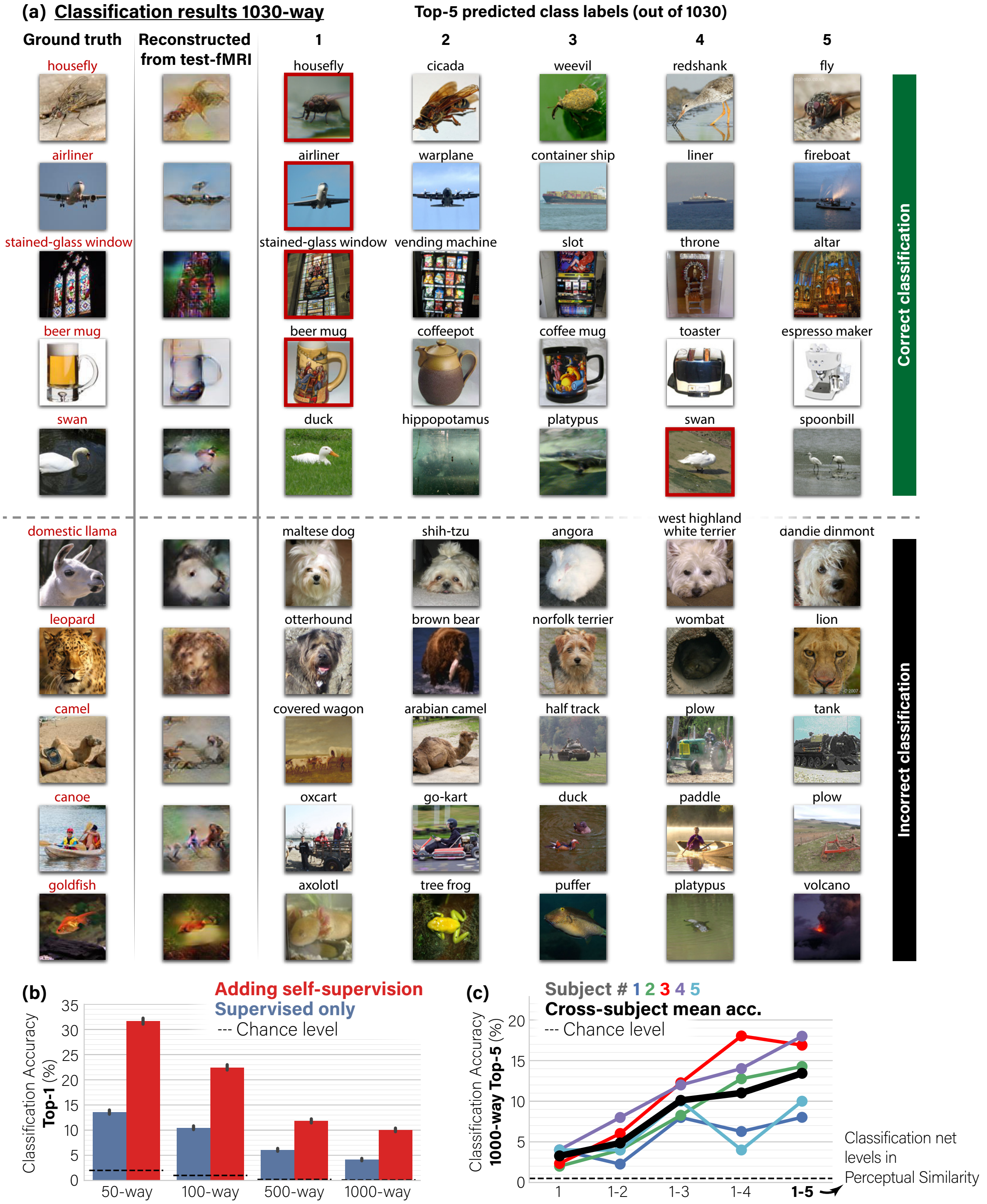

Figure 7. Self-supervision allows classification to rich and novel semantic categories. *(a) Visual classification results showing the Top-5 predictions out of 1030 classes for reconstructed test-image. We show examples where the ground-truth class (marked in red) is ranked among the Top-5 **(correct classification)** or excluded from it **(incorrect classification)**. For visualization, each class is represented by the nearest-neighbor image from the particular class. Note that "incorrect" predicted classes are often reasonable (e.g., "Leopard" wrongly predicted as "Lion"; "Camel" wrongly predicted as "Arabian camel"). **(b)** Top-1 Classification accuracy in an n-way classification task Adding unsupervised training on unpaired data (Fig 1d,e) dramatically outperforms the baseline of the supervised approach (Fig 1b). **(c)** Ablation study of the Classification accuracy as a function of the Perceptual Similarity criterion for decoder training: Applying perceptual similarity only on outputs of the first VGG16 block (low "semantic" layers), and up to all its 5 blocks (high "semantic" layers). Applying full Perceptual Similarity on higher-level VGG features substantially improves classification performance. Panels **a**,**b** show results for subject 3. 95% Confidence Intervals shown on charts.*

the unpaired images have no fMRI recordings whatsoever. The 50 test-classes (to be recovered from the 50 test-fMRIs), are neither included in the 150 'paired' train-classes, nor in the 1000 classes of the 'unpaired' external images. *These are totally novel classes* (details in Experimental datasets). When checking the classification results of the 50 test fMRI, we test the classification of their reconstructed images against the gallery of 1030 ImageNet classes[3]. As mentioned earlier, such classification requires no explicit class-labels in the training (see Fig 4). It is achieved by comparing the Deep-Features of the reconstructed image, with the Deep-Feature class-representative vector – one representative vector for each of the 1030 classes in our gallery (or any other gallery of novel classes).

Fig 7a exemplifies novel-class classification results by our method. We present each reconstructed image alongside the five top predicted ('Top-5') classes among the 1030 classes. For visualization purpose, each of the Top-5 classes are visually exemplified by the nearest-neighbor image (most similar Deep-Features) among 100 randomly selected images from that class label. The first five rows show correct-classification cases at Top-5 accuracy level. In these cases our method successfully includes the ground truth class among the nearest five classes (marked by a red frame). Interestingly, many of the non-ground-truth classes which are assigned among the Top-5, are also reasonable, frequently representing semantic and visual content, which is reminiscent of the ground truth class as well. For example, the 'housefly' is found similar also to other flies and comparable shape insects, and the 'beer mug' is found also similar to other types of mugs or pots. The bottom rows show incorrect-classification cases, where the ground truth class is not found among the Top-5 classes. Nonetheless, even in these allegedly "failure" cases, many of the Top-5 classes (and even Top-1) are considerably relevant both semantically and visually. For example, the reconstructed 'canoe' image was associated with the 'paddle' boat class; ; the 'camel' was wrongly classified as an 'arabian camel' (these are considered as 2 different classes in the 1030 class-gallery); the "leopard' was wrongly classified as a 'lion'; the 'domestic llama' was associated with several other white furry animal classes, etc.

We quantify the performance of our classification results using two different evaluation methods: (i) "Classification Rank" – the average rank of the correct class among all classes (Table 1), and (ii) the more familiar "n-way classification" accuracy ($n = 50, 100, 500, 1000$) – which is more intuitive (Fig 7b). Since "classification rank" does not binarize classifications into success or failure, it is more sensitive to differences among classifiers; hence we use it for statistical testing. Note that n-way classification performance can be derived from classification ranks, but not vice versa. In the "Classification Rank" method, for each reconstructed image we record the rank of its ground-truth class among the 1030 gallery classes according to its "semantic" (Deep-Feature) similarity. Table 1 summarizes novel-class classification rank results for all five subjects in 'fMRI on ImageNet'. To demonstrate the power of our self-supervised approach we compare its classification performance with a baseline of the purely supervised approach. This baseline uses reconstructions, which were produced by a Decoder trained using the scarce paired data alone (Fig 1b). This comparison shows a leap improvement in median classification rank in favor of our self-supervised approach in all five subjects. Notably, for Subjects 2-5 (excluding Subjects 1 who has the lowest median noise ceiling) the advantage of our self-supervised approach generalizes beyond the specifically chosen 50 test images and classes of the considered dataset.

In addition to the Ranking-score within the 1030 gallery classes (for each test-fMRI), we present another alternative way of evaluating the classification results – through classification accuracy in an n-way *classification* experiments (for $n = 50, 100, 500, 1000$), using our automated Deep-Features class-similarity criterion. Fig 7b shows Top-1 classification accuracy across a range of classification task difficulties (shown for Subject 3; the remaining subjects can be found in the Supplementary-Material). The tasks differ in the number of candidate classes (n-way) from which prediction is made (i.e., the

[3] ImageNet consists of 15K semantic classes, from which only 1000 classes participate in the ImageNet classification challenge (ILSVRC). In 'fMRI on ImageNet' [21] which we use, only 20 out of the 50 test-classes are included among the ILSVRC classes. The remaining 30 classes are taken from the larger collection of 15K ImageNet classes. Since our gallery is based on the 1000 ILSVRC classes, at train-time we omit the test-classes, resulting in 980 train-classes ($= 1000 - 20$). At classification test-time, we add the 50 test-labels to the gallery, resulting in 1030 class labels ($= 980 + 50$).

Classification rank (out of 1030) for 'fMRI on ImageNet' [21]

	fMRI data noise-ceiling median (SD)	Supervised only median rank (SE)	Adding self-supervision median rank (SE)
Subject 1	0.56 (0.28)	136.0 (48.3)	**106.0 (43.1)**
Subject 2	0.57 (0.30)	156.5 (48.1)	**∗75.5 (35.8)**
Subject 3	0.73 (0.28)	118.0 (37.0)	**∗38.5 (26.8)**
Subject 4	0.68 (0.30)	165.0 (49.9)	**∗71.5 (23.8)**
Subject 5	0.58 (0.29)	212.5 (40.0)	**∗87.0 (31.5)**

Table 1. Self-supervision allows classification to rich and novel semantic categories. *Median rank of the ground truth class among 1030 class representatives (Lower is better). Significant differences between the two methods are marked with asterisks (Wilcoxon test, $N = 50$, $p < .05$). Adding self-supervision leads to significant improvement in classification rank for the four (out of five) subjects with the highest fMRI median noise ceiling.*

percent of cases that the Top-1 predicted class out of n class labels is indeed the ground-truth class). Our full method scores 31.7% Top-1 accuracy ($SEM = 0.3\%$, $N = 25000$) in 50-way classification task. Even when scaling to 1000-way (as in ImageNet classification), our method scores 10.1% Top-1 accuracy ($SEM = 0.2\%$, $N = 25000$), which exceeds chance level accuracy by more than 100-fold. Contrasting this performance with the baseline of the supervised approach shows a striking leap improvement in classification-accuracy in favor of our self-supervised approach: between x2 and x3 accuracy improvement, in all the n-way experiments ($n = 50, 100, 500, 1000$).

We further performed an ablation study of the Perceptual Similarity for the task of semantic classification. Fig 7c shows 1000-way Top-5 classification accuracy by our self-supervised method, where the reconstructions used are produced by ablated versions of the Perceptual Similarity [25]. Specifically, we limit the Perceptual Similarity criterion, which is used in Decoder training, to a varying range of Deep VGG layers, starting from using only the outputs of the first block (low "semantic" features) of VGG16, and up to aggregating outputs from all five blocks (high "semantic" features) of the pretrained network as in the full method. We find that the classification accuracy of the reconstructed images shows an increasing trend with the number of higher-level features, which are used as reconstruction criteria. This highlights the significance of the Perceptual Similarity reconstruction criterion, which includes higher-level features, for semantic classification. Note that the increasing trend appears to various degree for different subjects, depending on experiment noise and subject-specific noise ceiling (e.g., Subject 1 having the lowest noise ceiling); Nevertheless, the trend is well illustrated by their cross-subject mean accuracy.

Our classification approach is inspired by [21]. Both methods use deep-feature embeddings to search for the nearest-neighbor class in a gallery of novel classes (our embedding is extracted from the reconstructed image (Fig 4b), whereas that of [21] employs an intermediate deep image-embedding decoded from fMRI). Notably, [21] presented classification of novel categories in a 2-way task (i.e., discriminating between the correct category and a single random category). Here we scale up this classification task to 1000-way (i.e., finding the correct category among 1000 rich categories). To the best of our knowledge, we are the first to demonstrate such large-scale semantic classification capabilities from fMRI data.

Predominance of early visual areas in reconstruction

To reconstruct the images of 'fMRI on ImageNet', we considered 4600 visual-cortex voxels provided and labeled in [21]. To study the contribution of different visual areas to our reconstruction performance, we selected subsets of voxels according to their marked brain areas, and restricted the training of our Encoder/Decoder to those voxels. Fig 8 shows reconstruction results when using voxels only from the following visual areas: (i) V1 (870 voxels), (ii) V1-V3, which refer to as Lower Visual Cortex (LVC, 2300 voxels), (iii) Fusiform Face Area (FFA), Parahippocampal Place Area (PPA), and

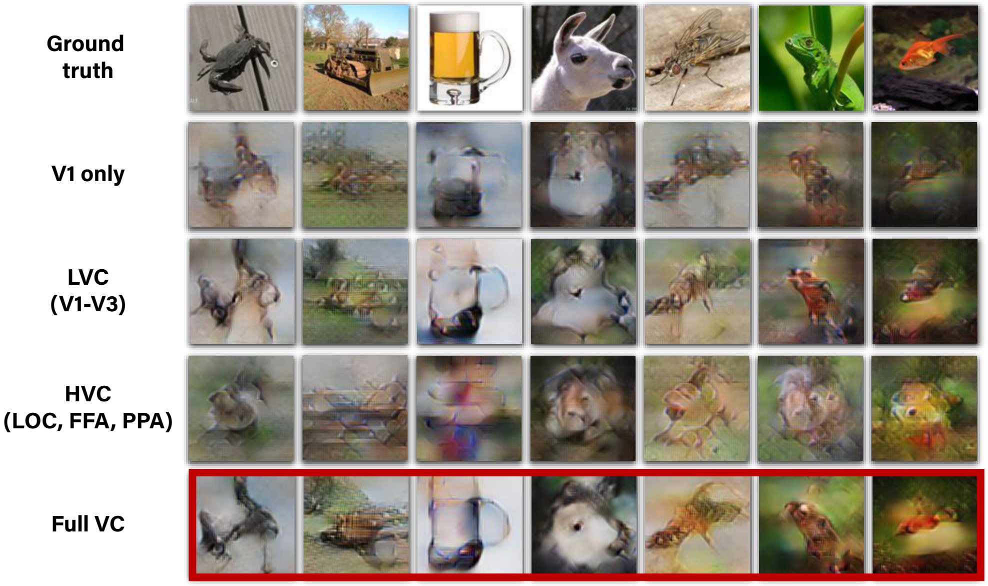

Figure 8. Decoding quality is dominated by early visual areas. *Columns show reconstructions using our method with fMRI data from various ROIs in the visual cortex including: • **Primary Visual Cortex** – V1 • **Lower Visual Cortex** – V1-V3 • **Higher Visual Cortex** – Fusiform Face Area (FFA), Parahippocampal Place Area (PPA), Lateral Occipital Cortex (LOC) • **Full Visual Cortex** – LVC + V4 + HVC (in red frame).*

Lateral Occipital Cortex (LOC), which we refer to as Higher Visual Cortex (HVC, 2150 voxels), or (iv) Full Visual Cortex (VC = LVC + V4 + HVC, 4600 voxels). These results show that the early visual areas, particularly V1-V3 (LVC), contain most of the information recoverable by our method, whereas considering voxels from HVC alone leads to substantial degradation in performance despite comprising approximately half of the complete visual cortex voxels. Nevertheless, the higher visual areas clearly add semantic interpretability to the reconstructed images (which is evident when comparing the reconstructions from the Full VC, to those from LVC only).

Importantly, we found that removing any single visual area from our dataset, including V1, does not degrade the results significantly, suggesting information redundancy across visual areas. The results are strongly affected, only when several regions, specifically the entire early visual cortex, are discarded. Furthermore adding V4 to either LVC or HVC did not change the results significantly.

Modeling biologically plausible population receptive fields

The human visual system is characterized by the well-known primate ‘retinotopic organization’ [27–35]. Retinotopy maps reflect the spatial tuning of cortical hypercolumns or of their aggregation into Population Receptive Field (pRF) as in the case of voxel data. Here we analyzed the biological meaning of our encoding model in terms of simulated retinotopy as previously proposed in [17,36].

To generate analogous spatial tuning maps for our *modeled* voxels, we estimated the voxel spatial tuning captured by our models. We visualized each voxel’s receptive field using the trained Encoder. Fig 9a shows receptive fields for several selected voxels, which indicates their spatial locality within the image. Next, we estimated the pRFs eccentricity and polar angle for each voxel. Lastly, we plot these data on the subject-specific cortical map, which corresponds to our analysis. Fig 9b,c shows the resulting tuning maps revealing the expected retinotopic organization. This includes the emergence of horizontal and vertical meridians and their transitions, contra-laterality and up-down inversion, and fovea-periphery gradual transition. We emphasize that this organization is purely driven by our optimization method involving natural stimuli and fMRI, where no biological atlas or other prior as

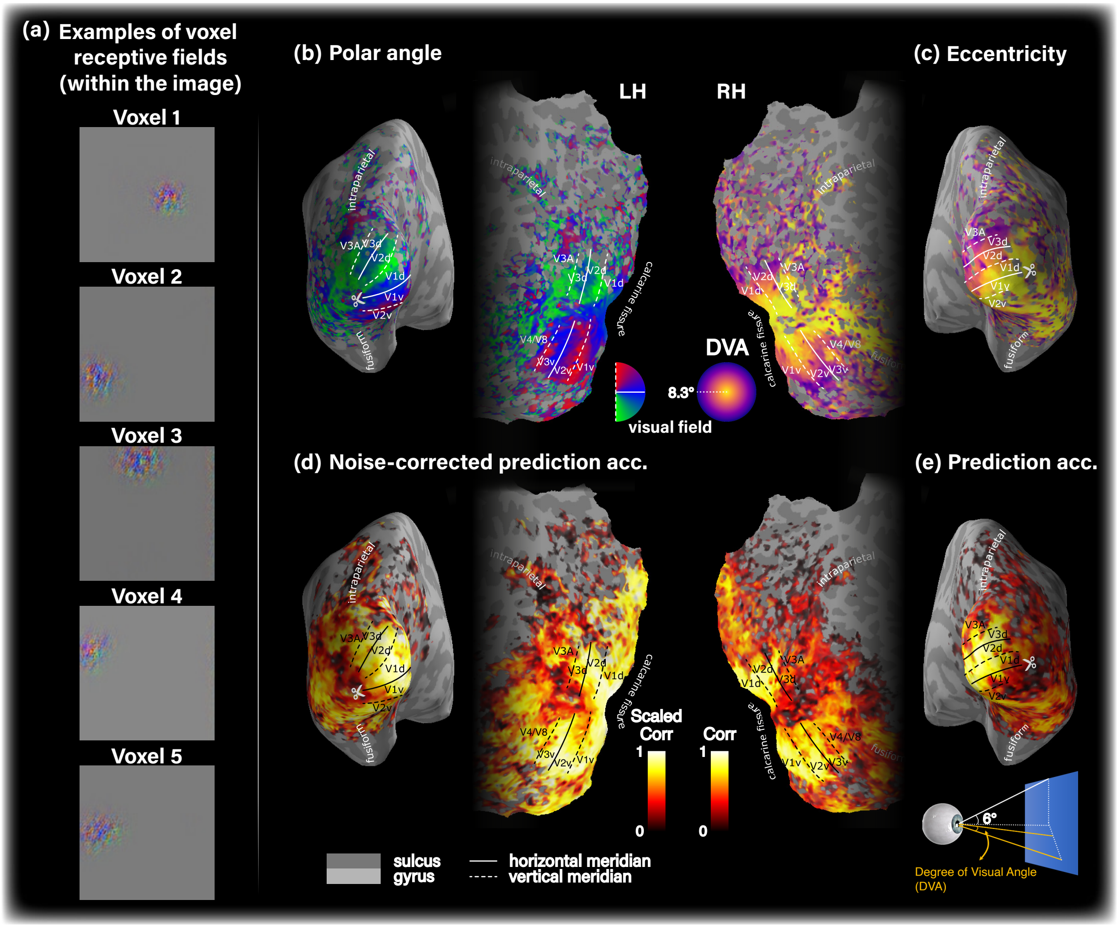

Figure 9. Our models capture biologically plausible voxel tuning properties. *(a) Receptive field of five selected voxels with high SNR from early visual cortex, which indicates their spatial locality in the image. Panels* ***(b)-(e)*** *show single subject data on the corresponding subject-specific cortical map.* ***(b)*** *Polar angle.* ***(c)*** *Eccentricity tuning, measured by degree of visual angle (DVA).* ***(d)*** *Noise-corrected prediction accuracy.* ***(e)*** *Prediction accuracy (non-scaled Pearson correlation). For simplicity we show the data on either left or right hemisphere. Voxel noise ceiling is coded by transparency level (alpha channel) in all cortical maps.*

such was imposed. To analyze the receptive field which is reflected through the Decoder, we considered the Decoder's fully-connected layer weights. When segregated per voxel, these weight maps reflect the voxel's receptive field. Importantly, these receptive field maps were well aligned (per voxel) with those of the Encoder. These results support the biological plausibility of our model's predictions.

We sought to analyze the prediction accuracy achieved by of our models. Fig 9d,e show the prediction accuracy distribution (Pearson correlation) of the modeled voxels when normalized by voxel noise ceiling (Fig 9d) and when not normalized (Fig 9e). The prediction noise ceiling is used to provide an estimate of the best possible prediction accuracy obtainable given infinite data [37]. These panels show high prediction accuracy in LVC, and low in HVC. Furthermore, they show that throughout the visual cortex our model markedly saturates the noise ceiling of the given data. This indicates sufficient expressive power to model, enabling it to capture the given data complexity.

Discussion

We presented state-of-the-art results in image-reconstruction and semantic categorization from fMRI data. To date, the performance in the task of natural image reconstruction and semantic categorization from human fMRI recordings is limited by the characteristics of fMRI datasets. In the typical case, the paired training data are scarce, represent a narrow semantic coverage, and have a different statistics than the test data. The statistics' discrepancy in our case, results from the differences between the train/test repeat-count (SNR) difference.

Our self-supervised training on tens of thousands of additional unpaired images from wide coverage adapts the decoding model to the statistics of natural images and novel categories. Furthermore, training on additional unpaired test-fMRI mitigates the impact of the discrepancy between the statistics of the train/test data. Thus our framework enables substantial improvement in image reconstruction quality and classification capabilities compared to methods that rely only on the scarce paired training data. This, together with high-level Perceptual Similarity constrains, leads to state-of-the-art image reconstructions from fMRI, of unprecedented quality, as supported by image-metric-based and extensive behavioral evaluations. We accomplish this for two substantially different fMRI datasets using a single method.

Our self-supervised training on tens of thousands of unpaired external images further leads to unprecedented capabilities in the semantic classification of fMRI data (and moreover, of classes never encountered during training). We consider the challenging 1000-way semantic classification task, and demonstrate a striking leap improvement (more than 2x) in classification performance when applying our self-supervised approach over a purely supervised approach. ***To the best of our knowledge, we are the first to demonstrate such large-scale semantic classification capabilities (1000-way) from fMRI data*** . We also show that incorporating the Perceptual Similarity criterion, with its reconstruction objectives over higher level feature representations, is a strong gain factor to achieving our high classification rates. Altogether we find that the Perceptual Similarity criterion, which is harnessed here for reconstruction and semantic classification, greatly benefits both tasks.

Our ablation studies indicate that reconstruction quality is dominated by data originating from Lower Visual Cortex (V1-V3). The extended architecture of the Encoder, which incorporates high-level features was designed to improve information-harnessing from the Higher Visual Cortex (HVC) as well. Indeed prediction accuracy maps show that the noise-ceiling is saturated throughout the visual cortex, including in higher visual areas. This findings suggests a reasonable representation of HVC by our model. Nevertheless, the SNR of the data arising from these areas renders them weaker contributors to overall reconstruction quality.

We provide evidence for the retinotopic organization implicitly learned (on its own) by our image-to-fMRI Encoder. This suggests that our models are biologically meaningful, as opposed to tailored and overfit to a limited dataset. Note that while we show data for the Encoder, we verified in our experiments that model voxels in the Decoder and the Encoder indeed agree (while not explicitly forced so).

The proposed method currently focuses on data from individual subjects. A natural extension of the present work is to combine information across multiple subjects. This is part of our future work.

Methods

Self-supervised Encoder/Decoder alternate training

The motivation behind training the Encoder (E) and Decoder (D) in separate phases (with a fixed Encoder during Decoder training) is designed to ensure that the middle junction's representation does not diverge from its physically meaningful entity by the unsupervised training objectives 1d,e. This middle junction represents fMRI responses in the combined E-D network, and natural images in the combined D-E network. Additionally, we start by supervised training of the Encoder in order to allow it to converge at the first phase, and then serve as strong guidance for the more severely ill-posed decoding task, which is the focus of the next phase. We next describe each phase in more detail.

Encoder supervised training (Phase I)

The supervised training of the Encoder is illustrated in Fig 3a. Let $\hat{r} = E\left(s\right)$ denote the encoded fMRI response from image, s, by Encoder E. We define fMRI loss by a convex combination of mean square error and cosine proximity with respect to the ground truth fMRI, r. The **fMRI loss** is defined as:

$$\mathcal{L}_r\left(\hat{r}, r\right) = \alpha \left\|\hat{r} - r\right\|_2 - \left(1 - \alpha\right) \cos\left(\angle\left(\hat{r}, r\right)\right), \tag{1}$$

where α is a hyperparameter set empirically (details in Implementation details). We use this loss for training the Encoder E. However, this loss is also used as the Decoder-Encoder loss (the self-supervised D-E loss on unpaired fMRI in Phase II), on which we detail in Decoder training.

Notably, in the considered fMRI datasets, the subjects who participated in the experiments were instructed to fixate at the center of the images. Nevertheless, involuntary eye movements were not recorded during the scans thus the fixation performance is not known. To accommodate the center-fixation uncertainty, we introduced small random shifts (+/- a few pixels) of the input images during Encoder training. This resulted in a substantial improvement in the Encoder performance and subsequently in the image reconstruction quality. Upon completion of Encoder training, we transition to training the Decoder together with the fixed Encoder.

Decoder training (Phases II)

The training loss of our Decoder consists of three main losses illustrated in Fig 3b:

$$\mathcal{L}^D + \mathcal{L}^{ED} + \mathcal{L}^{DE}. \tag{2}$$

$\mathcal{L}^D$ is a supervised loss on training pairs of image-fMRI. $\mathcal{L}^{ED}$ (Encoder-Decoder) and $\mathcal{L}^{DE}$ (Decoder-Encoder) are unsupervised losses on unpaired images (without fMRIs) and unpaired fMRIs (without images). All 3 components of the loss are normalized to have the same order of magnitude (all in the range $[0, 1]$, with equal weights), to guarantee that the total loss is not dominated by any individual component. We found our reconstruction results to be relatively insensitive to the exact balancing between the three components. We next detail each component of the loss.

$\mathcal{L}^D$: Decoder Supervised Training is illustrated in Fig 1b. Given {fMRI, Image} training pairs $\{(r, s)\}$, the supervised loss $\mathcal{L}^D$ is imposed on the decoded image, $\hat{s} = D\left(r\right)$, and is defined via the image reconstruction objective, $\mathcal{L}_s$, as

$$\mathcal{L}^D = \mathcal{L}_s\left(\hat{s}, s\right).$$

$\mathcal{L}_s$ consists of losses on image RGB values, $\mathcal{L}_{RGB}$, as well as losses on Deep Image Features extracted from the image using a pretrained VGG16 network [26] (a deep network tailored for the task of object recognition from images). We denote the deep features extracted from an image, s, by $\varphi\left(s\right)$, on which we apply a ***Perceptual Similarity*** criterion, $\mathcal{L}_{perceptual}$, which gave a significant performance leap.

Unlike our preliminary work [24], where we imposed only a Mean-Square-Error loss on the low level features alone (hence failed to capture or exploit any ”semantic” appearance or interpretation), here we impose Perceptual similarity [25] on the outputs of all the five feature-extractor blocks of VGG (from low to high VGG layers, i.e., lower-to-higher “semantic” levels), denoted as $\varphi_{vgg-blocks}\left(s\right)$. This metric is implemented by cosine proximity between channel-normalized ground-truth and predicted features at each block output. The complete criterion is then a sum of the block-wise contributions. The Image loss for a reconstructed image $\hat{s}$ reads:

$$\mathcal{L}_s\left(\hat{s},s\right)=\mathcal{L}_{RGB}\left(\hat{s},s\right)+\mathcal{L}_{perceptual}\left(\hat{s},s\right)+\mathcal{R}\left(\hat{s}\right) \tag{3}$$

$$\begin{cases}\mathcal{L}_{RGB}\left(\hat{s},s\right)\propto\left\|\hat{s}-s\right\|_1\\ \mathcal{L}_{perceptual}\left(\hat{s},s\right)\propto-\sum_{b=1}^{5}\cos\left(\angle\left(\varphi_{vgg-blocks}^{b}\left(\hat{s}\right),\varphi_{vgg-blocks}^{b}\left(s\right)\right)\right)\end{cases} \tag{4}$$

The last term, $\mathcal{R}\left(\hat{s}\right)$, corresponds to total variation (TV) regularization of the reconstructed (decoded) image, $\hat{s}=D\left(r\right)$. In addition to defining the Decoder supervised loss, the same Image loss is also used as the loss for the self-supervised Encoder-Decoder training on unpaired images (images without fMRI), explained next. We now detail on the main novelty of our method: Unsupervised training with unpaired data.

$\mathcal{L}^{ED}$: Encoder-Decoder training on unpaired Natural Images is illustrated in Fig 1d. This objective enables to train on any desired unpaired image (images for which no fMRI was ever recorded), well beyond the 1200 images included in the fMRI dataset. In particular, we used ∼50K additional natural images from ImageNet’s 1000-class data [23]. *This allows adaptation to the statistics of many more novel semantic categories, thus learning the common higher-level feature representation of various novel classes.* To train on images without corresponding fMRI responses, we map images to themselves through our Encoder-Decoder transformation,

$$s\mapsto\hat{s}_{ED}=D\left(E\left(s\right)\right).$$

The unsupervised component $\mathcal{L}^{ED}$ of the loss in Eq. 2 on unpaired images, s, reads:

$$\mathcal{L}^{ED}=\mathcal{L}_s\left(\hat{s}_{ED},s\right),$$

where $\mathcal{L}_s$ is the Image loss defined in Eq 3.

$\mathcal{L}^{DE}$: Decoder-Encoder training on unpaired test fMRI is illustrated in Fig 1e. Adding this objective greatly improved our reconstruction quality compared to training on paired samples only. To train on fMRI data without corresponding images, we map an fMRI response to itself through Decoder-Encoder transformation:

$$r\mapsto\hat{r}_{DE}=E\left(D\left(r\right)\right).$$

This yields the following unsupervised component $\mathcal{L}^{DE}$ of the loss in Eq. 2 on unpaired fMRI responses r:

$$\mathcal{L}^{DE}=\mathcal{L}_r\left(\hat{r}_{DE},r\right),$$

where $\mathcal{L}_r$ is the fMRI loss defined in Eq. 1.

Importantly, the fMRI samples which we used here were drawn from the test cohort (which is legitimate, since we never use nor know the test images). *This enables to adapt the Decoder to the statistics of the test-fMRI data* (which we wish to decode). Once the Decoder is trained using those 3 losses ($\mathcal{L}^{D}+\mathcal{L}^{ED}+\mathcal{L}^{DE}$), we apply it on the test-fMRI to decode it and reconstruct the test image.

Voxel receptive field visualization and estimation

To generate retinotopy maps (as in Fig 9), we start by visualizing each voxel's receptive field (pRF) using the trained Encoder. To this end we follow a gradient-based approach [38, 39]; Given a random input image, we compute the gradient of a particular voxel with respect to this input. This allows to visualize the image which would drive the maximum change in activity at the target voxel. To produce a heat-map, the values within the resulting gradient-image are squared, averaged across the color-channels, and normalized. Next, we define the pRF center as the center of mass of the *preprocessed* map. The preprocessing was designed to minimize noise effects. It included map smoothing with a Gaussian kernel, $\sigma = 3$, followed by raising the map values to the power of 10. About 15% of the voxels had pRF maps which were not confined spatially around a center of mass, and were thus discarded in subsequent analysis. The remaining 85% voxels were considered in the retinotopy maps.

Deep Architecture and Runtime details

An illustration of the Encoder and the Decoder architectures can be found in the Supplementary-Material. We focused on 112×112 RGB or grayscale image reconstruction (depending on the dataset), although our method works well also on other resolutions. The Encoder comprises four parallel branches of representation, built on top of features extracted from blocks 1-4 of VGG19. This enables to benefit from the hierarchy of "semantic" levels of the pretrained VGG network. The outputs of the four resulting branches (with their various resolutions) are then fed into branch-specific learned convolutional modules, which are designed to reduce the representation's dimensions to more compact representations of 28×28×32 or 14×14×32 (Height×Width×ConvolutionChannels). These modules consist of batch normalization, 3×3 convolution with 32 channels, ReLU, ×2 subsampling, and batch normalization. The first branch preceeds with an additional ×2 maxpooling while the fourth branch is not subsampled. Inspired by the feature-weighted receptive field [40], we designed a locally-connected layer which acts on the spatial and channel dimensions separately. This separation enables a dramatic decrease in the number of parameters that would be required to regress the voxel activations. In this space-feature locally-connected layer, for each spatial coordinate we stack along the channel dimension the values of the immediate 9 neighboring coordinates. Each resulting tensor (26×26×288 or 12×12×288, after eliminating the boundaries) is multiplied by a spatially locally-connected layer which learns the feature-to-voxel receptive field mapping. To encourage the locality of the receptive field mask (per voxel) we penalize the total variation of these spatial weights. The spatially-reduced tensors are followed by a cross-channel locally-connected layer, which weights the contribution of each feature/channel per voxel. Finally the outputs of the 4 branches are concatenated along the channel dimension and followed by a locally-connected layer, designed to weigh the contribution from each branch. We initialized all weights using Glorot normal initializer, except for the last layer which was 1-initialized (and forced to remain non-negative).

The Decoder architecture uses a locally-connected layer to transform and reshape the input vector-form fMRI input into 64 feature maps with spatial resolution 14×14. This representation is then followed by three blocks, each consists of: (i) ×2 up-sampling, (ii) 5×5 convolution with unity stride, 64 channels, and ReLU activation, and (iii) group normalization (16 groups). To yield the output image we finally performed an additional convolution, similar to the preceding ones, but with three channels to represent colors, and a sigmoid activation to keep the output values in the 0-1 range. We used Glorot-normal [41] to initialize the weights.

Hyperparameters. We trained the Encoder with using Adam optimizer for 50 epochs with an initial learning rate of 1e-3, with a 90% learning rate drop using milestones (20, 30, and 35 epochs). During Decoder training with supervised and unsupervised objectives, each training batch contained 16 pairs (supervised training), 16 unpaired natural images (randomly sampled from the external image database – images without fMRI), and 16 unpaired test-fMRI (fMRI without images). We trained the Decoder for 150 epochs using Adam optimizer with an initial learning rate of 1e-3, and 80% learning rate drop after every 30 epochs.

Runtime. Our system completes the two-stage training within approximately 1.5 hours using a single Tesla V100 GPU. Once trained, the inference itself (decoding of a new fMRI) is performed in real time.

Experimental datasets

We experimented with two publicly available benchmark fMRI datasets summarized in Table 2. The same architectures and hyperparameters were used for both datasets. The 1250 images in 'fMRI on ImageNet' were drawn from 200 selected ImageNet categories. 150 categories (classes) were used as training data (8 images per category – altogether 1200 training images). The 50 remaining image categories were designated as the novel test categories, represented by 50 test images (1 image from each test category).

fMRI Dataset	N train images (K repeats)	N test images (K repeats)	N voxels
fMRI on ImageNet [21]	1200 (1)	50 (35)	4500
vim-1 [1]	1750 (2)	120 (13)	8500

Table 2. Summary of fMRI datasets used in analyses. *Repeat count refers to the number of fMRI recordings per presented stimulus. Voxel count refers to approximated number of voxels used in analysis.*

External (unpaired) images database. For unsupervised training on unpaired images (Encoder-Decoder objective, Fig 1d) we used additional 49K natural images from 980 classes of ImageNet ("ILSVRC") train-data [23]. We verified that the images and categories in our additional unpaired external dataset do not overlap with the test-images and test-categories in the 'fMRI on ImageNet' (the inference target). Since the 50 test-classes of 'fMRI on ImageNet' [21] partially overlap with the 1000 original ILSVRC classes, we particularly discarded the 20 overlapping classes.

Behavioral experiments. The participants in the Mechanical Turk behavioral experiments gave their online informed consent to be recorded, and were granted financial incentives for every completed survey. The research protocol was reviewed and approved by the Bioethics and Embryonic Stem Cell Research Oversight (ESCRO) Committee at the Weizmann Institute of Science. In order to assure the validity of the behavioral data (e.g. bot observers, fatigue along the survey), we screened subjects according to their score in interleaved sanity check experiments. The sanity check experiments comprised adding to the actual experiments also 10% unexpected trivial identification tasks of mildly degraded versions of the ground truth images, instead of the reconstructed images. We further discarded subjects with MTurk success-score (reputation) lower than 97%. Each survey consisted of 50 or 20 trials corresponding to the number of test-images comparison in 'fMRI on ImageNet' [21] or 'vim-1' [1][4], all of which were reconstructed using a single particular method. In each trial subjects were presented with a reconstructed image and n candidate images, the ground-truth image and $n-1$ distractor images, and were prompted "Which image at the bottom row is most similar to the image at the top row?". To assure task difficulty agreement across subjects and reconstruction methods the set of distractor images was randomly selected for each test-image but remained fixed across surveys; Our results were insensitive to their re-selection.

Semantic category decoding. We defined the feature vector underlying the class representatives to be the outputs of block 4 in AlexNet, and used Pearson correlation as the distance metric for ranking class representatives. Our experiments showed that using this intermediate representation level as the embedding of choice yields optimal results for classification.

Noise-Ceiling. We estimated the fMRI prediction Noise-Ceiling by half-split over the test data repeats following [37].

[4] 'vim-1' originally contains 120 test-images, however in the behavioral evaluation we considered only the subset of 20 images that were defined in [13] as test-images

Statistics. We used Wilcoxon signed-rank (paired) test (two-tailed) for significance testing in the image-metric-based multi-image identification experiments, as well as in the rank-classification experiments. For the (unpaired) behavioral experiments we used Mann-Whitney rank test.

Acknowledgments

This project has received funding from the European Research Council (**ERC**) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 788535).

Author Contributions

G.G. and R.B. designed the experiments. N.G. added the concept of Perceptual Similarity to the framework. G.G. designed and wrote the paper, contributed the cortical maps, the statistical and the fMRI data analyses. R.B. and N.G. implemented the network and conducted the image-reconstruction and classification experiments. A.H. contributed the behavioral experiments. T.G. and F.S. provided guidance on fMRI preprocessing and brain visualization. T.G. also advised on statistical testing. M.I. conceived the original idea and supervised the project. All authors discussed the results and commented on the manuscript.

References

1. K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant, “Identifying natural images from human brain activity,” *Nature*, vol. 452, pp. 352–355, 3 2008.
2. T. Naselaris, R. J. Prenger, K. N. Kay, M. Oliver, and J. L. Gallant, “Bayesian Reconstruction of Natural Images from Human Brain Activity,” *Neuron*, vol. 63, pp. 902–915, 9 2009.
3. S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant, “Reconstructing visual experiences from brain activity evoked by natural movies.,” *Current biology : CB*, vol. 21, pp. 1641–6, 10 2011.
4. H. Wen, J. Shi, Y. Zhang, K.-H. Lu, J. Cao, and Z. Liu, “Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision,” *Cerebral Cortex*, vol. 28, pp. 4136–4160, 12 2018.
5. U. Guclu, M. A. J. van Gerven, U. Güçlü, and M. A. J. van Gerven, “Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream,” *Journal of Neuroscience*, vol. 35, pp. 10005–10014, 7 2015.
6. G. Shen, T. Horikawa, K. Majima, and Y. Kamitani, “Deep image reconstruction from human brain activity,” *PLOS Computational Biology*, vol. 15, p. e1006633, 1 2019.
7. C. Zhang, K. Qiao, L. Wang, L. Tong, Y. Zeng, and B. Yan, “Constraint-Free Natural Image Reconstruction From fMRI Signals Based on Convolutional Neural Network,” *Frontiers in Human Neuroscience*, vol. 12, p. 242, 6 2018.
8. K. Han, H. Wen, J. Shi, K.-H. Lu, Y. Zhang, D. Fu, and Z. Liu, “Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex,” *NeuroImage*, vol. 198, pp. 125–136, 9 2019.
9. Z. Ren, J. Li, X. Xue, X. Li, F. Yang, I. Zhicheng, J. Upenn, and X. Gao, “Reconstructing Seen Image from Brain Activity by Visually-guided Cognitive Representation and Adversarial Learning,” tech. rep., 2019.
10. M. Mozafari, L. Reddy, and R. Vanrullen, “Reconstructing Natural Scenes from fMRI Patterns using BigBiGAN,” tech. rep.
11. K. Qiao, J. Chen, L. Wang, C. Zhang, L. Tong, and B. Yan, “BigGAN-based Bayesian Reconstruction of Natural Images from Human Brain Activity,” *Neuroscience*, vol. 444, pp. 92–105, 7 2020.
12. G. Shen, K. Dwivedi, K. Majima, T. Horikawa, and Y. Kamitani, “End-to-end deep image reconstruction from human brain activity,” *bioRxiv*, p. 272518, 2018.
13. G. St-Yves and T. Naselaris, “Generative Adversarial Networks Conditioned on Brain Activity Reconstruct Seen Images,” in *Proceedings - 2018 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2018*, pp. 1054–1061, Institute of Electrical and Electronics Engineers Inc., 1 2019.
14. K. Seeliger, U. Güçlü, L. Ambrogioni, Y. Güçlütürk, and M. A. van Gerven, “Generative adversarial networks for reconstructing natural images from brain activity,” *NeuroImage*, vol. 181, pp. 775–785, 2018.
15. Y. Lin, J. Li, H. Wang, and S. Jiao, “DCNN-GAN: Reconstructing Realistic Image from fMRI,” tech. rep., 2019.
16. J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini, “Distributed and overlapping representations of faces and objects in ventral temporal cortex,” *Science*, vol. 293, pp. 2425–2430, 9 2001.
17. M. Eickenberg, A. Gramfort, G. Varoquaux, and B. Thirion, “Seeing it all: Convolutional network layers map the function of the human visual system,” *NeuroImage*, vol. 152, pp. 184–194, 2017.
18. T. Konkle and A. Caramazza, “Tripartite organization of the ventral stream by animacy and object size,” *Journal of Neuroscience*, vol. 33, pp. 10235–10242, 6 2013.
19. H. Wen, J. Shi, W. Chen, and Z. Liu, “Deep Residual Network Predicts Cortical Representation and Organization of Visual Features for Rapid Categorization,” *Scientific Reports*, vol. 8, p. 3752, 12 2018.
20. K. Qiao, J. Chen, L. Wang, C. Zhang, L. Zeng, L. Tong, and B. Yan, “Category decoding of visual stimuli from human brain activity using a bidirectional recurrent neural network to simulate bidirectional information flows in human visual cortices,” *Frontiers in Neuroscience*, vol. 13, no. JUL, 2019.

21. T. Horikawa and Y. Kamitani, "Generic decoding of seen and imagined objects using hierarchical visual features," *Nature Communications*, vol. 8, pp. 1–15, 5 2017.

22. Y. Akamatsu, R. Harakawa, T. Ogawa, and M. Haseyama, "Estimating Viewed Image Categories from fMRI Activity via Multi-view Bayesian Generative Model," in *2019 IEEE 8th Global Conference on Consumer Electronics (GCCE)*, pp. 127–128, IEEE, 10 2019.

23. J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, IEEE, 6 2009.

24. R. Beliy, G. Gaziv, A. Hoogi, F. Strappini, T. Golan, and M. Irani, "From voxels to pixels and back: Self-supervision in natural-image reconstruction from fMRI," in *Advances in Neural Information Processing Systems*, 2019.

25. R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," tech. rep.

26. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 9 2014.

27. G. St-Yves and T. Naselaris, "The feature-weighted receptive field: an interpretable encoding model for complex feature spaces," *NeuroImage*, vol. 180, pp. 188–202, 10 2018.

28. V. B. Mountcastle, "Modality and topographic properties of single neurons of cat's somatic sensory cortex.," *Journal of neurophysiology*, vol. 20, pp. 408–34, 7 1957.

29. D. Y. Ts'o, M. Zarella, and G. Burkitt, "Whither the hypercolumn?," *The Journal of physiology*, vol. 587, pp. 2791–805, 6 2009.

30. T. Bonhoeffer and A. Grinvald, "Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns," *Nature*, vol. 353, pp. 429–431, 10 1991.

31. B. M. Dow, "Orientation and Color Columns in Monkey Visual Cortex," *Cerebral Cortex*, vol. 12, pp. 1005–1015, 10 2002.

32. B. A. Wandell and J. Winawer, "Computational neuroimaging and population receptive fields," *Trends in Cognitive Sciences*, vol. 19, pp. 349–357, 6 2015.

33. J. Gomez, V. Natu, B. Jeska, M. Barnett, and K. Grill-Spector, "Development differentially sculpts receptive fields across early and high-level human visual cortex," *Nature Communications*, vol. 9, p. 788, 12 2018.

34. R. B. Tootell, M. S. Silverman, E. Switkes, and R. L. De Valois, "Deoxyglucose analysis of retinotopic organization in primate striate cortex," *Science*, vol. 218, pp. 900–901, 11 1982.

35. M. I. Sereno, A. M. Dale, J. B. Reppas, K. K. Kwong, J. W. Belliveau, T. J. Brady, B. R. Rosen, and R. B. Tootell, "Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging," *Science*, vol. 268, no. 5212, pp. 889–893, 1995.

36. H. Wen, K. Han, J. Shi, Y. Zhang, E. Culurciello, and Z. Liu, "Deep Predictive Coding Network for Object Recognition," 2018.

37. A. Lage-Castellanos, G. Valente, E. Formisano, and F. De Martino, "Methods for computing the maximum performance of computational models of fMRI responses," *PLOS Computational Biology*, vol. 15, p. e1006397, 3 2019.

38. K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," tech. rep.

39. S. Grossman, G. Gaziv, E. M. Yeagle, M. Harel, P. Mégevand, D. M. Groppe, S. Khuvis, J. L. Herrero, M. Irani, A. D. Mehta, and R. Malach, "Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks," *Nature Communications*, vol. 10, pp. 1–13, 12 2019.

40. G. St-Yves and T. Naselaris, "The feature-weighted receptive field: an interpretable encoding model for complex feature spaces," 2017.

41. X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," 3 2010.