From voxels to pixels and back: Self-supervision in natural-image reconstruction from fMRI

We thank the reviewers for their comments and endorsements. Below are our answers to the main questions/concerns.

R1: Training on test-fMRI samples – not convinced the approach is valid. We understand the reviewer's concern. Note however that our "training on test data" refers only to training on ***unlabeled samples from the Decoder's input space*** *(test-fMRI)*, whereas the *test-images* (the "labels") are never used at any stage of the training. Thus, such training is valid. We will better clarify the distinction between training on the "test-fMRI" (which is the input to the network, hence totally valid to train on), vs. training on the the "test-images" (which is the desired output of the network, hence illegal/invalid to train on, and indeed we do not). We realize that this distinction is confusing, and will clarify it.

R1: A general-purpose Decoder is the ultimate aim. We respectfully beg to differ. We believe that a 'universal' fixed Decoder for arbitrary input fMRI samples is inconceivable given the limited training data, and the large variability in fMRI acquisition parameters and subjects. Instead, we propose an adaptive decoder that learns to adapt to new fMRI samples (of never-before-seen images, and with new statistics), as they come along. This adaptation underlies the high performance of our Decoder on new fMRI data. Since this holds for any set of test-fMRI (which can be incorporated in training) our self-supervised Decoder has significant capability to generalize well to new/held-out data.

R1: Different statistics of fMRI in the Train & Test datasets. Indeed, the train/test fMRI SNR discrepancy results from averaging a different number of repeated recordings per image (typical of many fMRI datasets). This statistical discrepancy introduces an additional challenge of 'domain transfer/adaptation', affecting the performance of current decoding methods. Our $\mathcal{L}^{DE}$ objective directly addresses this issue: It enables to learn the *different* statistics of the test-fMRI, and specifically of the target-fMRI. Fig. A shows the impact of a different number of repeat-counts of the test-fMRI (averaging 1, 5, 10, 20, or 35 randomly selected repeats). Reconstruction improves as the number of test repeats (SNR) increases *in the Test-set only* (Train-set remains the same). It shows that our D-E architecture exploits the better SNR of the test-fMRI and adapts D to the statistics of the test-data. We will add this explanation/figure to the paper/supp-material. As suggested by R1, we will add more background regarding SNR/repeat-count.

R1+R4: Loss magnitudes & Loss ablation experiments. The 3 components of the loss in Eq.2 ($\mathcal{L}^{D}$, $\mathcal{L}^{ED}$, $\mathcal{L}^{DE}$) are *normalized* to have the same order of magnitude (all in the range $[0,1]$). This guarantees that no loss is dominated by the other two. We therefore assigned equal weights to all losses. We will add these clarifications. There are two simple ways to examine the relative effect of the different loss terms: (i) change the relative weights of the image and voxel reconstruction losses, or (ii) alter the ratio between the unlabeled and labeled examples in each batch. Fig. B shows the effect of both of these – altering the number of unlabeled examples per batch (8, 16, or 32 samples out of 64), and assigning different weights $\times 2$ to the image/voxel losses. We will add this explanation+figure to paper/Supp-Material.

R4: Choice of hyperparameters; potential risk of overfitting the test dataset – Maybe try another dataset? This is an important point which we failed to stress: Our method was applied on two separate (and very different) datasets using *the exact same* hyperparameters. We will add this important clarification to the paper. This is strong evidence to the general applicability of our method to very different datasets with very different statistics. We are not aware of any other publicly available fMRI dataset of *natural images* for which there are reported reconstructions. We found our reconstructions to be relatively robust to variations in the hyperparameters (as can also be seen in Fig. B). We further performed a cross-validation procedure for selecting the 'optimal' set of hyperparameters based on the train-set alone. This yields very similar hyperparameters to those we used, and similar reconstructions.

R4: Neuroscientific insights. While we believe that image-reconstruction can eventually become a strong neuroscientific tool, this is not the focus of the current paper. This paper highlights a new learning method/approach, which is exemplified on a difficult neuroscientific problem, but is not limited to it. We believe this approach may apply to other domains and problems, which are characterized by scarce labeled data.

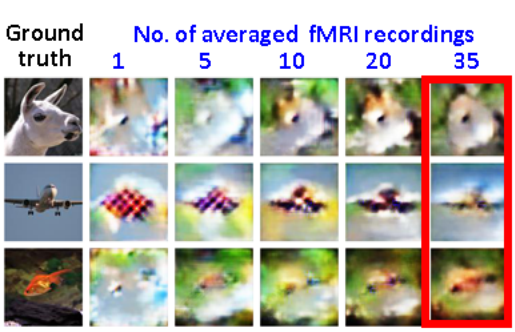

Figure A: Effect of averaging repeated fMRI recordings.

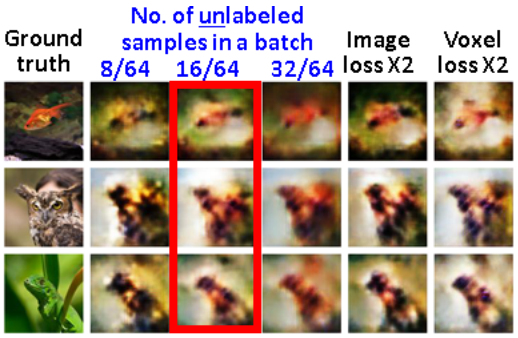

Figure B: Varying loss-weights & labelled/unlabeled ratio.

Red frames mark the current configuration in the paper