Contrastive-Equivariant Self-Supervised Learning Improves Alignment with Primate Visual Area IT

Thomas Yerxa [1 *] **Jenelle Feather** [1,2] **Eero P. Simoncelli** [1,2] **SueYeon Chung**[1,2]
[1]Center for Neural Science, New York University
[2]Center for Computational Neuroscience, Flatiron Institute, Simons Foundation

Abstract

Models trained with self-supervised learning objectives have recently matched or surpassed models trained with traditional supervised object recognition in their ability to predict neural responses of object-selective neurons in the primate visual system. A self-supervised learning objective is arguably a more biologically plausible organizing principle, as the optimization does not require a large number of labeled examples. However, typical self-supervised objectives may result in network representations that are overly invariant to changes in the input. Here, we show that a representation with structured variability to input transformations is better aligned with known features of visual perception and neural computation. We introduce a novel framework for converting standard invariant SSL losses into "contrastive-equivariant" versions that encourage preservation of input transformations without supervised access to the transformation parameters. We demonstrate that our proposed method systematically increases the ability of models to predict responses in macaque inferior temporal cortex. Our results demonstrate the promise of incorporating known features of neural computation into task-optimization for building better models of visual cortex.

1 Introduction

In the past decade, task-optimized deep neural networks (DNNs) have been used to predict responses of object-selective neurons in primates to natural image stimuli [Yamins et al., 2014, Schrimpf et al., 2020, Willeke et al., 2023]. Such networks have a pronounced advantage over more traditional models for explaining responses in deeper areas with more abstract representations, such as inferior temporal cortex (IT). This observation naturally leads to the hypothesis that task optimization can provide a normative account for IT neuron tuning properties: late-stage visual representations are shaped by the need to perform ecologically relevant tasks.

However, the task that initially led to these advances was that of supervised object classification, a specific task that relies on an implausibly large number of labeled examples [Lindsay, 2021]. More recently, computer vision has undergone a "self-supervised learning" (SSL) revolution. A variety of methods have been proposed to learn representations that match or surpass supervised training on multiple tasks by deriving sources of supervision from the data itself rather than relying on human annotations. For example, many popular SSL strategies aim to unify representations of different transformations of the same image (commonly referred to as "views"), while enforcing diversity among representations of distinct images. Additionally, self-supervised representations can predict primate neural responses with fidelity comparable to supervised representations [Zhuang et al., 2021, Konkle and Alvarez, 2022, Parthasarathy et al., 2024].

*Corresponding Author: tey214@nyu.edu

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-3043

Both of these training objectives are forms of invariance learning: responses of an ideal object classification model should be invariant across different objects from the same class, and self-supervised learning strives to achieve invariance to the transformations used to generate different views. However biological visual representations are not fully invariant across views [DiCarlo and Cox, 2007, Kuoch et al., 2024]. Indeed it has been demonstrated that training according to either of these two objectives leads to representations that are invariant to stimulus perturbations that are salient to human observers [Feather et al., 2023]. Additionally, even in Area IT, which is thought to subserve invariant object recognition, neural populations encode a significant amount of "category orthogonal" information (e.g., object pose or viewing conditions that are unrelated to semantic category) [Hong et al., 2016]. Furthermore, such selectivity for object-orthogonal attributes is meaningfully organized within Area IT [Hong et al., 2016] (i.e. object orthogonal attributes are linearly decodable from population responses). Whether such structured variability emerges in invariance-trained networks is likely determined by the uncontrolled inductive biases of the network architecture [Alleman et al., 2024].

Here, we develop an equivariant learning framework that encourages such structured variability in network representations. Our contributions are:

- We propose a novel framework that converts standard invariance-based self-supervised learning methods into "contrastive-equivariant" versions that produce structured, transformation-related variability. Unlike previous approaches, our method does not require supervised access to transformation parameters or costly modifications to the training procedure.
- We examine the tradeoff between invariance and structured variability through a series of representational analyses. We find that, relative to networks trained for invariance alone, our contrastive-equivariant network learns structured transformation variability that is shared across images and factorized with respect to variability related to changes in image content.
- We explore the impact of including an equivariant loss for predicting neural activity in IT, showing for the first time that explicitly encouraging structured variability via optimization leads to an improved ability to predict cortical responses to natural images.

2 Method

2.1 Transformation-Invariant Self-Supervised Learning (iSSL)

The influential work of [Chen et al., 2020] showed that applying two random transformations (often called "augmentations") to a batch of images, then training a network to identify which pairs of transformed images originated from the same sample with a cross-entropy style loss (the InfoNCE loss, first formulated in [Gutmann and Hyvärinen, 2010]) yields representations that are competitive with supervised training for object classification. Many subsequent studies have developed alternative objective functions that produce similar results: Barlow Twins[Zbontar et al., 2021], VICReg [Bardes et al., 2021], and W-MSE [Ermolov et al., 2021] enforce augmentation invariance along with a constraint that the global covariance matrix is the identity; SimSiam [Chen and He, 2021] and BYOL [Grill et al., 2020] employ architectural constraints that regularize towards uniform representations and simply optimize for transformation invariance. Other studies have formalized the problem in terms of maximizing information [Ozsoy et al., 2022] or capacity [Yerxa et al., 2024], subject to an invariance constraint, which has enabled connections to normative theories of coding efficiency and manifold capacity [Barlow et al., 1961, Chung et al., 2018].

To formalize the definition of iSSL, we denote a dataset of images (e.g., ImageNet) by $X \in \mathbb{R}^{N \times D}$, where N is the number of images and D is their dimensionality (number of pixels). Let $\tau(\cdot; \rho) : \mathbb{R}^D \to \mathbb{R}^D$ be a function parameterized by ρ that maps images to images (for example, for τ a random crop operation, ρ specifies the region to be cropped). The goal of iSSL algorithms is to learn the parameters W of some function $f(\cdot; W) : \mathbb{R}^D \to \mathbb{R}^d$ such that the variability over ρ is minimal while preserving variability over X (to avoid trivial solutions such as $f(\cdot; W) = 0$ for all inputs). Many methods achieve this by observing pairs of randomly augmented views of a batch of images: $X^A = \tau(X; \rho_1)$, $X^B = \tau(X; \rho_2)$, with $\rho_1, \rho_2 \sim p(\rho)$ where $p(\rho)$ is a pre-chosen probability distribution over augmentation parameters. Generally iSSL frameworks employ an objective function that operates on the outputs of f, $Z^A = f(X^A; W)$, $Z^B = f(X^B; W)$. One popular framework is

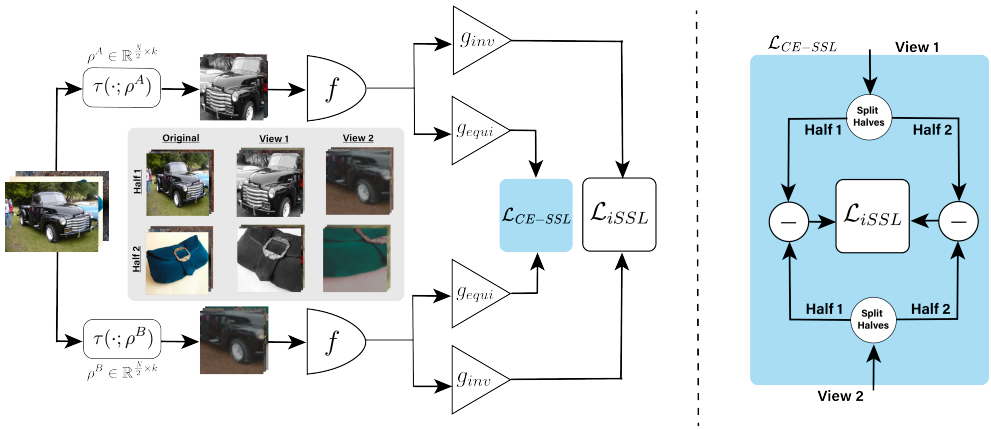

Figure 1: Diagram of the proposed training method. At the beginning of each training epoch the dataset is randomly split into two non-overlapping halves. Left Gray Panel: corresponding images in each subset are augmented using the same set of two random transformations (so the total number of random transformations is halved relative to a standard iSSL training scheme). Every view is passed through a representation network f (ResNet-50 in this work) and the outputs are projected into two embedding spaces by different projector networks, g_{inv} and g_{equi}. In the invariant embedding space a standard iSSL loss is applied, while in the equivariant embedding space the same iSSL loss is applied to the difference vectors between transformation-positive pairs (visualized on the right).

"Barlow Twins" [Zbontar et al., 2021], which uses the objective: $\mathcal{L}_{BT} = \Sigma_i(1-\mathcal{C}_{ii})^2 + \lambda\Sigma_{i,i\neq j}(\mathcal{C}_{ij})^2$ where $\mathcal{C}$ is the cross-correlation matrix between Z^A and Z^B. The first term encourages the outputs in response to the same image subject to different augmentations to be correlated, while the second encourages the outputs in response to distinct images to be uncorrelated.

Because complete invariance to the transformations employed in iSSL is harmful for downstream tasks, most frameworks employ a learnable "projector network" that maps the outputs of the representation network to an embedding space before applying the loss. The nearly ubiquitous use of this "guillotine regularization" [Bordes et al., 2022], means that most iSSL methods aim to learn a function *from which an augmentation invariant subspace can be extracted*. While this approach does permit some transformation-related variability in the representation, there is no explicit control or encouragement of that variability, and no incentive for that variability to be usefully structured.

2.2 Contrastive-Equivariant Self-Supervised Learning (CE-SSL)

To induce structured variability in learned representations, we require that an equivariant subspace can be extracted from f alongside the invariant subspace described above. A function is equivariant to a set of input transformations if there exists a corresponding set of output transformations that induce the same changes. In the self-supervised learning setting this property can be expressed as:

$$\forall \tau_\rho \in P,\ \ \forall x \in X,\ \ \exists T_\rho : f(\tau_\rho(x)) = T_\rho(f(x)), \tag{1}$$

where $\tau_\rho = \tau(\cdot;\rho)$ and P is the set of possible values of transformation parameters. Note that invariance is a special case of equivariance, in which T_ρ is an identity transformation for all ρ. To avoid this degenerate solution, we will require both that similarly transformed inputs be related by the same transformation in the output space, and that differently transformed inputs are related to each other by different transformations.

Our training methodlogy, summarized in Fig. 1, follows the principle first proposed by [Gupta et al., 2023]: "Equivariance should be learned from pairs of data, as in invariant contrastive learning." First we split our dataset of images into two random non-overlapping equal-sized partitions $X_1, X_2 \in \mathbb{R}^{\frac{N}{2} \times D}$. Next we apply a randomly selected augmentation to both X_1 and X_2, so that corresponding rows of $X_1^{A/B}$ and $X_2^{A/B}$ contain distinct images that have been subjected to the same augmentations.

https://doi.org/10.52202/079017-3043

Note that this reduces the total number of random samples of ρ by a factor of two relative to standard iSSL methods. Finally the resulting representation vectors are each fed through two distinct projector networks, $z_i^{A/B} = g_{inv}(r_i^{A/B})$ and $\tilde{z}_i^{A/B} = g_{equi}(r_i^{A/B})$. These two embeddings are optimized to be invariant to transformations and discriminative across base images, or invariant to base images and discriminative across transformations, respectively. The overall objective (loss) functions is:

$$\begin{aligned} \mathcal{L}_{\text{overall}} &= (1-\lambda)\mathcal{L}_{iSSL} + \lambda\mathcal{L}_{CE-SSL} \\ \text{where} \quad \mathcal{L}_{iSSL} &= \mathcal{L}([z_1^A, z_2^A], [z_1^B, z_2^B]), \\ \text{and} \quad \mathcal{L}_{CE-SSL} &= \mathcal{L}(z_1^A - z_1^B, z_2^A - z_2^B), \end{aligned} \tag{2}$$

where both terms are written in terms of $\mathcal{L}$, a self-supervised learning loss function that encourages invariance and uniformity [Wang and Isola, 2020] (e.g., L_{BT}) and λ is a hyperparameter that determines the relative importance of extracting an invariant or equivariant subspace from the shared representation. In the notation of Eq. (1), by designing $\mathcal{L}_{CE-SSL}$ to encourage similar transformations to induce similar displacements in the output space, we are implicitly specifying that our output transformations are of the form $T_\rho(z) = z + z_\rho$. Thus we leverage the principles underpinning contrastive invariance learning to encourage representations that contain useful transformation-related information; this choice differentiates this formulation from previous equivariant self-supervised learning approaches.

3 Results

3.1 Implementation Details

Architecture and invariance ojective. For all experiments we use a ResNet-50 architecture [He et al., 2016] as the backbone representation network f. Our training scheme is compatible with any choice of iSSL framework, as specified by the choice of $\mathcal{L}_{iSSL}$. We experimented with three different base methods chosen to span the range from "instance contrastive" to "dimension contrastive" [Garrido et al., 2023a]: SimCLR [Chen et al., 2020], MMCR [Yerxa et al., 2024], and Barlow Twins [Zbontar et al., 2021]. In each case, we define g_{inv} using the projector network architecture proposed in the original work. To retain the synergy between the normalization scheme, loss function, and projector architecture achieved by each framework we use the same architecture for both g_{inv} and g_{equi}.

Pretraining dataset and augmentations. We train using the ImageNet-1k dataset and the standard set of augmentations first introduced in [Grill et al., 2020], which includes random resized cropping, color jittering, Gaussian blurring, solarization, and horizontal flips. See Appendix A.2 for exact training details.

Invariance-equivariance tradeoff. For each of the three choices of $\mathcal{L}_{iSSL}$ we trained networks with hyperparameter values $\lambda \in \{0.0, 0.001, 0.1, 0.2, 0.3, 0.4, 0.5\}$, yielding a total of 21 learned representations (note: $\lambda = 0$ corresponds to standard iSSL). We found that classification performance becomes severely degraded for values of λ larger than 0.5 (see Appendix A.4).

3.2 Representational Analyses

Bures metric comparisons. We conducted a series of experiments to determine the extent to which various sources of variability in our dataset were meaningfully organized. The experiments utilized the Bures metric, which is the Wasserstein ("Earth Mover's") distance between mean-centered Gaussian distributions with covariance matrices C_1 and C_2:

$$D_B(C_1, C_2) = \text{trace}\left(C_1 + C_2 - 2\left(C_2^{1/2} C_1 C_2^{1/2}\right)^{1/2}\right). \tag{3}$$

When C_1 and C_2 are normalized to have a trace of 1, the maximal distance of 2.0 occurs when the variabilities lie in orthogonal subspaces (or are completely "factorized" from each other) and the minimum distance of 0.0 occurs when the covariances are equal. More generally, a large Bures distance indicates two sources of variability are factorized from each other and a low distance

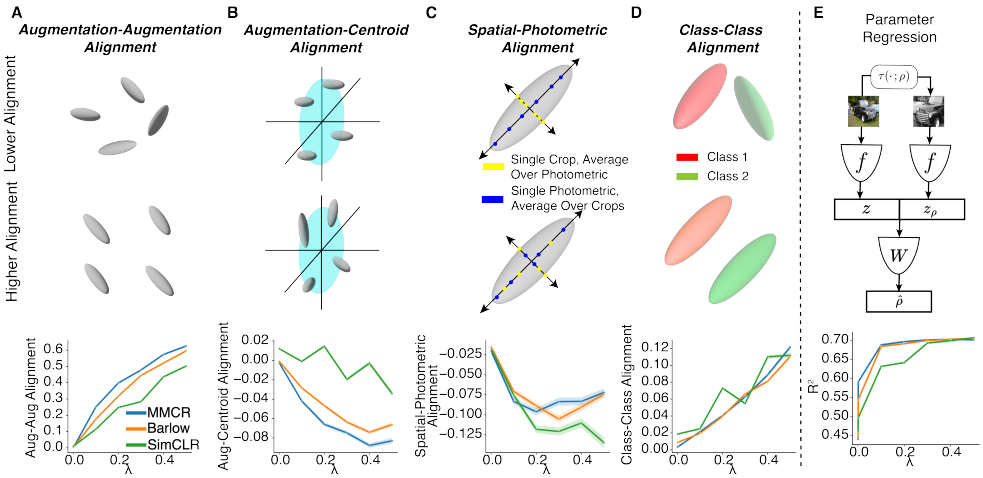

Figure 2: Effects of equivariance on representational geometry. **A**: Alignment between augmentation manifolds (gray ellipsoids). **B**: Alignment between augmentation manifolds (gray ellipsoids) and the centroid manifold (blue disk). **C**: Alignment between spatial and photometric manifolds. Gray ellipsoids represent single augmentation manifolds, and blue/yellow points indicate the mean over the outputs from many transformations of a single view obtained via a photometric/spatial transformation, respectively. Expected distance is larger when the two sources of variability are factorized. **D**: Same as A., but for class manifolds. **E**: A schematic of the parameter regression experiment. In each panel, the bottom row depicts the results of each analysis described in the text of Sections 3.2 and 3.2. Shaded regions indicate 95% confidence intervals (estimated over the same comparisons the expected distance is estimated over for A-D and over 5 independent runs of the regression experiments for E). A summary of the sources of variability used to compute C_1 and C_2, and the ensemble used to estimate the expected alignment is measured can be found in Table A.5

indicates shared structure. We first estimate the trace-normalized covariance of the outputs of some network f over two sources of variability and compute the Bures metric between the two.

Because we are mainly interested in the impact of the equivariance loss relative to the invariant baseline, each analysis below is carefully controlled to expose any structural differences. In particular, we estimate C_1 and C_2 over identical inputs for an invariant network and an equivariant network trained using the same base objective but a non-zero value of λ. We then can directly compare the measured Bures distance for the invariant network and each equivariant network ($\lambda \neq 0$): $\Delta D_B = D_B(C_1^{\lambda=0}, C_2^{\lambda=0}) - D_B(C_1^{\lambda\neq 0}, C_2^{\lambda\neq 0})$; this measure quantifies the amount of alignment between two sources of variability in an equivariant network relative to the invariant network baseline. In each of the panels in the bottom row of Fig. 2 we show how $\mathbb{E}[\Delta D_B]$ (denoted simply "Alignment") evolves as a function of λ when C_1 and C_2 are estimated over different sources of variability (y-axis labels indicate the ensembles over which the variabilities were estimated). We show the joint distribution of $(D_B(C_1^{\lambda=0}, C_2^{\lambda=0}), D_B(C_1^{\lambda\neq 0}, C_2^{\lambda\neq 0}))$ and summarize the sources of variability in each experiment described below in Appendix A.5.

Augmentation-Augmentation alignment. First we determine the extent to which augmentation variability is shared across base images in each network (Fig. 2A). For these experiments, both C_1 and C_2 are estimated over many random transformations of single images in the validation set (we will refer to the responses to such a group as an "augmentation manifold"). The expectation is then over randomly sampled pairs of augmentation manifolds. The positive expected difference of distance indicates the equivariant networks consistently produce lower distance between augmentation manifolds, indicating more shared augmentation variability across base images. This structure is closely related to what is encouraged by the equivariance loss term and the orderly increase as a function of λ suggests that we are optimizing effectively.

Augmentation-Centroid factorization. We next investigate the extent to which variability over augmentations is factorized from variability over base images (Fig. 2B). We use the "centroid

https://doi.org/10.52202/079017-3043

manifold," [Yerxa et al., 2024] to characterize the variability over base images, by measuring the covariance over base images of the means over augmentations. That is, for these experiments (for each network respectively) C_1 is the covariance of the centroids of all augmentation manifolds and C_2 is the covariance of a randomly selected augmentation manifold. We observe that equivariant networks generally exhibit a larger distance between centroid and augmentation manifolds indicating increased factorization (or lower alignment) of image-content variability and image-augmentation variability. This structure was not explicitly encouraged by the objective and can be considered an emergent property of the equivariant learning procedure.

Spatial-Photometric factorization. Next we ask whether our equivariant training procedure induced increased factorization of variability to different types of input transformations (Fig. 2C). The standard augmentation procedure involves first taking a random crop (spatial variability) of a given image and then applying a series of pixel-level transformations (color-jittering, gaussian blurring, etc.) (photometric variability). To assess the impact of these two distinct classes of image transformations we first chose 20 random crops a given image, then applied the same set of 20 random photometric transformations to each individual crop, yielding 400 different views of each base image. C_1 and C_2 are then estimated over network responses that are averaged over different crops or different photometric transformations respectively, and the expectation is taken over different (single) base images. We observe the equivariant networks consistently exhibit increased factorization (i.e. larger Bures distances relative to the invariant trained network). This again is an emergent property of the equivariant learning procedure, and is particularly interesting in light of recent work that discovered that this form of transformation-factorization is more correlated with neural predictivity than transformation invariance [Lindsey and Issa, 2024].

Class-Class factorization. Finally we asked whether within-class variability was more or less shared between distinct classes in equivariant networks by estimating C_1 and C_2 over responses to all images in distinct classes in the validation set (the expectation is then taken over different random pairs of classes) (Fig. 2D). Increased sharing of variability between class manifolds has been demonstrated to increase manifold capacity, and can make representations better suited for multi-task evaluations [Wakhloo et al., 2023, 2024]. We observe higher alingment (lower expected pairwise Bures distances) in the equivariant networks indicating that the "class manifolds" relative to the invariant networks.

Linear embedding of augmentation-related information. While the above experiments demonstrate that equivariant training induces increased alignment of transformation-related variability between images, this does not necessarily imply that this variability is coherently organized. To assess this more directly, we measure the extent to which augmentation parameters can be linearly decoded from the networks' representations. Specificlly, we regress the concatenated outputs of a clean and transformed image onto the parameters of the applied augmentation. We report the resulting coefficient of determination (R^2) on a heldout set of validation images (Fig. 2). The equivariant training is seen to increase the amount of linearly accessible augmentation information relative to invariant training (the leftmost points plotted in Fig 2E). We further analyzed a set of equivariant models trained with weaker augmentation parameters (see A.6 for details). In these networks, we again observe that equivariant training increased the amount of linearly accessible augmentation information compared to invariant training. This holds not only for augmentation parameters within the training range (left panel 7) but also for parameter values beyond the training range (right panel 7). Thus, the equivariance properties of the models generalize beyond the training distribution. Future work could examine generalization to unseen types of augmentations.

3.3 Neural Predictivity

We utilized the BrainScore evaluation pipeline Schrimpf et al. [2018] to measure the extent to which each learned representation can linearly predict neural responses measured in macaque area IT, for four different experimental datasets. At the time of testing, our highest performing model (Barlow Twins objective, $\lambda = 0.2$) had the 10th highest average predictivity for area IT out of approximately 250 publicly available models on the Brain-Score leaderboard. Across a reasonably large range of values of λ, the equivariant models improved the neural predictivity relative to the invariant baseline ($\lambda = 0$) for all four datasets (Fig. 3). Many previous publications have noted that changes in training objective function have a small effect on neural predictivity, relative to other factors

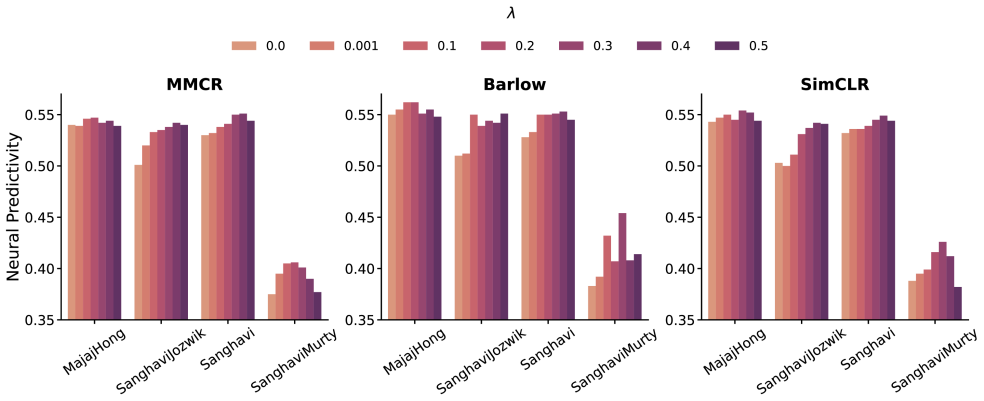

Figure 3: Brain-Score (noise-ceiled predictivity evaluated via ridge regression) for each value of λ (different colored bars) for each IT dataset (groups of columns) and base objective functions (different figure panels). For all datasets and base objectives the invariant network ($\lambda = 0$, lightest bars) is outperformed by at least one equivariant network, and the spread in predictivity over values of λ is significantly larger than the spread in predictivity over base objective functions for invariant networks.

such as training dataset [Tuckute et al., 2022, Conwell et al., 2023, Yerxa et al., 2024]. In contrast, encouraging equivariance produced much larger gains than choosing between different base invariant objectives: the range of predictivities over the sweep of λ was around 4 times larger than the range of predictivities over objective functions for the invariant baseline. We further contextualize the scale of predictivity improvements in Fig 4 by comparing models to all public submissions on the BrainScore leaderboard; our equivariant training procedure improves performance of the already-strong invariant models to nearly state-of-the-art levels of IT predictivity. By training the most predictive model for 1000 epochs (rather than 100), we achieved 0.5355 mean fraction of explained variance, which makes this the top IT brain prediction model. To ensure that the observed alignment increases are not architecture specific, we trained a smaller set of models using different backbone architectures and observed similar trends when using both smaller and larger networks (see Appendix A.7 for details).

We quantified the correlation between our various representational measurements and the neural predictivity for each of the four electrophysiology datasets in Table 1. We observed that the only representational metric with a correlation greater than 0.4 across all four neural datasets was the Spatial-Photometric distance, which is the metric most closely related to the factorization score described in [Lindsey and Issa, 2024]. While this previous study described a correlation between structured variability and neural predictivity measured from a large set of pre-trained models, our results demonstrate that explicitly encouraging such structures can improve alignment between artificial and biological representations. In addition to the previously described representational measurements, we also looked at the linear decoding of the hue modulation parameter in isolation. Hue modulation is one of 12 augmentation parameters that are linearly decoded in the parameter regression measures described in Section 3.2. We observed a strong correlation between neural predictivity and hue modulation, particularly with the Sanghavi-Jozwik dataset, which is the only response dataset that included color image stimuli (last column of Table 1).

3.4 Transfer Learning

Several previous studies that aim to reduce augmentation-invariance of self-supervised features have reported that the resulting representations generalize better to out-of-distribution classification datasets [Gupta et al., 2023, Xiao et al., 2020, Suau et al., 2023, Chavhan et al., 2022]. However most of these studies focused on using the smaller ImageNet-100 dataset for training, in one case reporting that the transfer learning gains diminish or disappear when using ImageNet-1k [Chavhan et al., 2022]. We tested our set of networks on 6 different downstream tasks and found limited evidence that the equivariant features confer an advantage in terms of out-of-distribution generalization when training on a sufficiently large and diverse dataset (see Table 2). To address this discrepancy with the literature we conducted additional experiments on networks trained using the ImageNet-100 dataset, and in this case observed improvement in generalization to diverse downstream tasks.

https://doi.org/10.52202/079017-3043

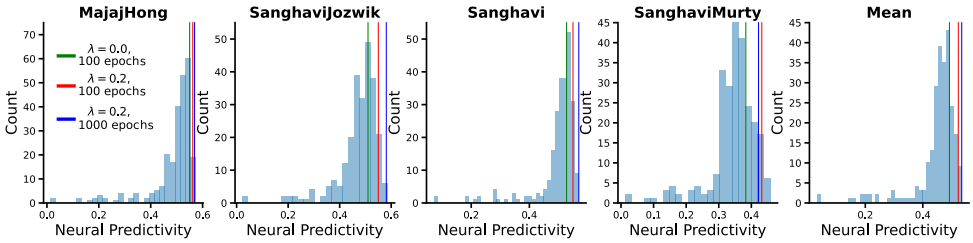

Figure 4: Histogram of neural predictivity scores for the 249 models on the public Brain-Score leaderboard at the time of testing, for each of the four considered IT datasets, as well as the mean over the four datasets. In each plot the vertical green line shows the score of the invariant Barlow Twins model, the red line shows the score for the equivariant Barlow Twins model with $\lambda = 0.2$, and the blue line shows a new model trained for 1000 epochs, also using Barlow Twins for the base loss and $\lambda = 0.2$.

Table 1: Absolute values of Pearson correlation coefficients (R^2) between various representational measurements and the neural predictivity across each of the four IT datasets. The correlation was measured over each value of λ and base objective function for a total of 21 networks. Each column corresponds to a panel in Fig. 2, except for hue, which is the regression score obtained for the random hue modulation parameter in isolation.

Neural Dataset	Aug-Aug	Aug-Centroid	Spatial-Photometric	Class-Class	Param Regression	Hue Regression
Majaj-Hong	0.03	0.02	0.42	0.10	0.38	0.28
Sanghavi-Jozwik	0.86	0.7	0.83	0.77	0.91	0.91
Sanghavi	0.84	0.84	0.68	0.63	0.85	0.86
Sanghavi-Murty	0.33	0.41	0.42	0.14	0.56	0.48

It is also worth noting that CE-SSL trained networks do not outperform their invariant counterparts on in-distribution generalization (see A.2.1 and Fig. 5). This is not surprising in light of the fact that the suite of augmentations and architectures employed in SSL have been in some sense optimized by the community in order to improve performance on this task (by aligning the transformation invariance task with the standard in-distribution classification task). However, for out-of-distribution classification tasks where the task-alignment is worse, the equivariance task could mitigate this mismatch. A concrete example is the Flowers-102 dataset, where the color of petals is a much stronger predictor of class than color is in, say, the ImageNet-1k dataset (so the color insensitivity induced by the standard augmentations could be detrimental). For this dataset we do see marginal improvements, but note that the improvements are much more pronounced when pretraining on smaller datasets (ImageNet-100). There are at least 2 possible explanations for this: (1) for ImageNet-1k pretraining the performance of the networks is already quite high, and the task is saturated, or (2) there is a more fundamental reason that the improvements in transfer learning induced by equivariance decrease as the size and diversity of the pretraining dataset grows. Future work could explicitly disambiguate between these hypotheses to determine why the benefits of transformation-related variability for out of distribution generalization are outweighed by the gains of scaling the dataset. Furthermore this result shows that the increased neural predictivity we observe in ImageNet-1k trained networks cannot be explained by a need to perform better on a variety of invariant-classification tasks.

4 Relationship to Existing Augmentation-Sensitive SSL Methods

A key feature that differentiates our approach is that it encourages equivariant structure without explicit access to augmentation parameters. This is enabled by the "paired augmentation" data generation procedure, and to the best of our knowledge CARE [Gupta et al., 2023] is the only existing work that shares this feature. Our method has two advantages over CARE: (1) in CARE the equivariance loss is applied in the same space as the invariance constraint, and because there are no "negative equivariant pairs" in the CARE framework, learning an invariant representation would perfectly satisfy the equivariant constraint; and (2) in CARE the standard augmentation pipeline is

Table 2: Frozen-Linear Evaluation for invariant and equivariant trained networks on 6 different downstream datasets: Cifar-10/100 [Krizhevsky et al., 2009], Oxford-Pets Parkhi et al. [2012], Describable Textures Database [Cimpoi et al., 2014], Flowers-102 [Nilsback and Zisserman, 2008], and Food-101 [Bossard et al., 2014]. We closely follow the evaluation procedure from [Lee et al., 2021] (see Appendix A.4 for details) and report top1 accuracy for each objective/dataset. In all cases we report the mean over 5 runs of the evaluation procedure, we observed very little variability (maximum of .2%, over all evaluations, we report the standard deviation over runs in Appendix A.4. The equivariant networks are denoted by prepending a "CE" before the objective and were trained using $\lambda = 0.1$, which enabled a substantial amount of structure variability without significantly impacting frozen-linear classification on the SSL training dataset (see Appendix A.2). For ImageNet-1k trained networks out of distribution performance decreased for most evaluations, while for ImageNet-100 trained networks performance was improved in 15 of 18 cases.

ImageNet-100 Training						
Objective	Cifar-10	Cifar-100	Pets	DTD	Flowers-102	Food-101
MMCR	84.3	63.3	67.0	66.1	83.1	60.9
Barlow	87.7	68.7	74.8	67.0	88.3	63.4
SimCLR	87.8	68.8	74.3	66.6	88.5	64.8
CE-MMCR	87.3	69.4	68.9	65.7	87.5	64.1
CE-Barlow	88.0	69.1	73.6	67.3	89.5	65.5
CE-SimCLR	87.9	68.2	72.6	67.5	88.6	65.2
ImageNet-1k Training						
Objective	Cifar-10	Cifar-100	Pets	DTD	Flowers-102	Food-101
MMCR	92.2	76.9	85.3	75.4	93.9	73.8
Barlow	91.8	75.8	86.5	73.0	93.8	72.2
SimCLR	91.8	74.7	85.1	74.5	92.7	70.5
CE-MMCR	92.2	76.6	84.3	75.7	94.0	73.8
CE-Barlow	91.8	75.3	85.7	75.6	94.2	73.0
CE-SimCLR	91.0	73.6	82.3	73.9	92.2	70.5

used to optimize the base $\mathcal{L}_{iSSL}$ loss and paired augmentation are used in parallel to optimize the equivariance term (so an increased number of passes through the network is necessary relative to standard training).

In EquiMod [Devillers and Lefort, 2023] the projector network is conditioned on the augmentation parameters by appending the parameters to the output of f. The authors theorize that knowledge of the augmentation parameters could allow the projector network to better extract an invariant subspace tailored to each transformation, thus allowing for more structured variability in the representation space. Alternatively, the projector network could simply ignore the augmentation parameters, resulting in a structure that is identical to invariant SSL. In practice [Garrido et al., 2023b] have found this to be the case. Split-Invariant-Equivariant (SIE) and Amortised Invariance (AI) learning [Garrido et al., 2023b] each improve on this principle by using a hypernetwork approach: a separate network takes as inputs the augmentation parameters and outputs the parameters of either g or both f and g respectively. While the collapse issue of EquiMod is avoided, this comes at the expense of significantly complicating the network computation, and in the case of AI introduces new parameters that need to be tuned for every downstream task (when augmentation information is not available). Still other methods supplement the standard invariant SSL loss with an auxillary term that involves predicting the parameters of the input transformation [Lee et al., 2021, Dangovski et al., 2021]. The relationship of our method to these is analogous to the relationship of transformation-invariant self-supervised learning to supervised classification.

https://doi.org/10.52202/079017-3043

5 Discussion

We've developed a new self-supervised objective that explicitly encourages structured variability in networks, and demonstrated that it can produce increased alignment with responses of neurons in primate visual area IT. While we are not the first to incorporate a notion of equivariance to self-supervised learning, our method improves on existing work in several ways: it require no extra passes through the network relative to invariance-based learning, it encourages diversity in the representation of transformation-related information by leveraging advances in invariance-based learning, and it does not rely on supervised access to transformation parameters. The parsimony of our approach (applying the same objective to both outputs of individual images and to displacements between similarly transformed images) allows our technique to be easily adapted to other settings such as temporal self-supervised learning (discussed below). Although in this work we focused on the visual domain, similar equivariant and invariant objectives could be investigated for other domains such as audio and langauge representation learning.

Our approach induced several interesting features in the learned representations: transformation variability is shared across base images and factorized with respect to variability over base images, the variability induced by distinct types of transformations are factorized from each other, there is increased alignment between class manifolds, and transformation related information is linearly encoded. Some of these properties are closely related to the imposed objective and some are emergent. We also confirmed that several of these representational properties are correlated with increased neural predictivity. Future work can extend these correlative observations to better understand how increasing transformation sensitivity improves neural alignment. For example, one could analyze the residuals of predicted neural firing rates of distinct models to determine how "overlapping" the variance predicted by each is (or alternatively, attempt to fit the residual variance of one model with another). Such analyses are becoming more feasible with the collection and release of larger scale datasets of neural responses to natural images (e.g., [Madan et al., 2024]). We view this result as demonstrating the promise of incorporating knowledge gained from experimental observations and large scale comparative studies into optimization procedures to produce better models.

Although our experimnents reveal both induced and emergent benefits, the inclusion of an additive equivariance term in the objective does lead to fewer guarantees regarding the learned structure. For example in schemes where the output transformations (T_ρ's) are explicitly represented or learned, the resulting representation is "steerable" by default. It is of interest to investigate whether the output transformations could be reliably recovered from our learned representations. Additionally it would be interesting to consider other types of output transformations (CARE [Gupta et al., 2023] focuses on orthogonal transformations, and in the case where output transformations are learned they can be computed with nonlinear neural networks).

Finally, it is of interest to explore the use of more ecologically relevant sources of training data, e.g., by replacing synthetically transformed views of images with temporally adjacent frames of natural videos. This approach is particularly appealing from the perspective of biological plausibility, as the pairing of such training examples is readily available from natural visual experience. Several recent publications have shown that such a strategy can produce representations with competitive neural predictivity and performance on computer vision tasks [Zhuang et al., 2021, Parthasarathy et al., 2023, Venkataramanan et al., 2023]. In this context, the typical invariance loss can be thought of as incentivizing representational slowness [Földiák, 1991, Wiskott and Sejnowski, 2002]. The equivariance mechanism described in this work could be implemented by applying the same invariance-based loss function to the first temporal derivative of the responses, i.e. by encouraging the displacement between successive pairs of frames to be constant. Such a temporal-equivariance objective would incentivize representational straightness, which has been used to describe features of both human perception and neural activity in the ventral stream [Hénaff et al., 2021, 2019]. Straightness in artificial representations has been found to be correlated with both neural predictivity and adversarial robustness [Lindsey and Issa, 2024, Harrington et al., 2022, Niu et al., 2024]. These connections provide an array of promising research directions.

Acknowledgments and Disclosure of Funding

This work was funded by the Center for Computational Neuroscience at the Flatiron Institute of the Simons Foundation. S.C. is supported by the Klingenstein-Simons Award, a Sloan Research Fellowship, NIH award R01DA059220, and the Samsung Advanced Institute of Technology (under the project "Next Generation Deep Learning: From Pattern Recognition to AI"). All experiments were performed on the Flatiron Institute's high-performance computing cluster.

References

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

Martin Schrimpf, Jonas Kubilius, Michael J Lee, N Apurva Ratan Murty, Robert Ajemian, and James J DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 2020. URL https://www.cell.com/neuron/fulltext/S0896-6273(20)30605-X.

Konstantin F. Willeke, Kelli Restivo, Katrin Franke, Arne F. Nix, Santiago A. Cadena, Tori Shinn, Cate Nealley, Gabrielle Rodriguez, Saumil Patel, Alexander S. Ecker, Fabian H. Sinz, and Andreas S. Tolias. Deep learning-driven characterization of single cell tuning in primate visual area V4 unveils topological organization. Technical Report 2023.05.12.540591, bioRxiv, May 2023. URL https://doi.org/10.1101/2023.05.12.540591.

Grace W Lindsay. Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of cognitive neuroscience*, 33(10):2017–2031, 2021.

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.

Talia Konkle and George A Alvarez. A self-supervised domain-general learning framework for human ventral stream representation. *Nature communications*, 13(1):491, 2022.

N Parthasarathy, O J Hénaff, and E P Simoncelli. Layerwise complexity-matched learning yields an improved model of cortical area V2. *Trans. Machine Learning Research*, Jun 2024. URL https://openreview.net/forum?id=lQBsLfAWhj. Featured Certification.

James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007.

Michael Kuoch, Chi-Ning Chou, Nikhil Parthasarathy, Joel Dapello, James J DiCarlo, Haim Sompolinsky, and SueYeon Chung. Probing biological and artificial neural networks with task-dependent neural manifolds. In *Conference on Parsimony and Learning*, pages 395–418. PMLR, 2024.

Jenelle Feather, Guillaume Leclerc, Aleksander Mądry, and Josh H McDermott. Model metamers reveal divergent invariances between biological and artificial neural networks. *Nature Neuroscience*, 26(11):2017–2034, 2023.

Ha Hong, Daniel LK Yamins, Najib J Majaj, and James J DiCarlo. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature neuroscience*, 19(4): 613–622, 2016.

Matteo Alleman, Jack Lindsey, and Stefano Fusi. Task structure and nonlinearity jointly determine learned representational geometry. In *The Twelfth International Conference on Learning Representations*, 2024.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

 https://doi.org/10.52202/079017-3043

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *International Conference on Machine Learning*, pages 3015–3024. PMLR, 2021.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Serdar Ozsoy, Shadi Hamdan, Sercan Ö Arik, Deniz Yuret, and Alper T Erdogan. Self-supervised learning with an information maximization criterion. *arXiv preprint arXiv:2209.07999*, 2022.

Thomas Yerxa, Yilun Kuang, Eero Simoncelli, and SueYeon Chung. Learning efficient coding of natural images with maximum manifold capacity representations. *Advances in Neural Information Processing Systems*, 36, 2024.

Horace B Barlow et al. Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1(01), 1961.

SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. Classification and geometry of general perceptual manifolds. *Physical Review X*, 8(3):031003, 2018.

Florian Bordes, Randall Balestriero, Quentin Garrido, Adrien Bardes, and Pascal Vincent. Guillotine regularization: Why removing layers is needed to improve generalization in self-supervised learning. *arXiv preprint arXiv:2206.13378*, 2022.

Sharut Gupta, Joshua Robinson, Derek Lim, Soledad Villar, and Stefanie Jegelka. Structuring representation geometry with rotationally equivariant contrastive learning. *arXiv preprint arXiv:2306.13924*, 2023.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann Lecun. On the duality between contrastive and non-contrastive self-supervised learning. In *ICLR 2023-Eleventh International Conference on Learning Representations*, 2023a.

Jack W Lindsey and Elias B Issa. Factorized visual representations in the primate visual system and deep neural networks. *eLife*, 13, 2024.

Albert J Wakhloo, Tamara J Sussman, and SueYeon Chung. Linear classification of neural manifolds with correlated variability. *Physical Review Letters*, 131(2):027301, 2023.

Albert J Wakhloo, Will Slatton, and SueYeon Chung. Neural population geometry and optimal coding of tasks with shared latent structure. *arXiv preprint arXiv:2402.16770*, 2024.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*, 2018. URL `https://www.biorxiv.org/content/10.1101/407007v2`.

Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H McDermott. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *bioRxiv*, pages 2022–09, 2022.

Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *bioRxiv*, 2023. doi: 10.1101/2022.03.28.485868. URL `https://www.biorxiv.org/content/early/2023/07/01/2022.03.28.485868`.

Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. In *International Conference on Learning Representations*, 2020.

Xavier Suau, Federico Danieli, T Anderson Keller, Arno Blaas, Chen Huang, Jason Ramapuram, Dan Busbridge, and Luca Zappella. Duet: 2d structured and approximately equivariant representations. In *Proceedings of the 40th International Conference on Machine Learning*, pages 32749–32769, 2023.

Ruchika Chavhan, Jan Stuehmer, Calum Heggan, Mehrdad Yaghoobi, and Timothy Hospedales. Amortised invariance learning for contrastive self-supervision. In *The Eleventh International Conference on Learning Representations*, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Hankook Lee, Kibok Lee, Kimin Lee, Honglak Lee, and Jinwoo Shin. Improving transferability of representations via augmentation-aware self-supervision. *Advances in Neural Information Processing Systems*, 34:17710–17722, 2021.

Alexandre Devillers and Mathieu Lefort. Equimod: An equivariance module to improve visual instance discrimination. In *International Conference on Learning Representations*, 2023.

Quentin Garrido, Laurent Najman, and Yann Lecun. Self-supervised learning of split invariant equivariant representations. *arXiv preprint arXiv:2302.10283*, 2023b.

Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljačić. Equivariant contrastive learning. *arXiv preprint arXiv:2111.00899*, 2021.

Spandan Madan, Will Xiao, Mingran Cao, Hanspeter Pfister, Margaret Livingstone, and Gabriel Kreiman. Benchmarking out-of-distribution generalization capabilities of dnn-based encoding models for the ventral visual cortex. *arXiv preprint arXiv:2406.16935*, 2024.

 https://doi.org/10.52202/079017-3043

Nikhil Parthasarathy, SM Eslami, Joao Carreira, and Olivier Henaff. Self-supervised video pretraining yields robust and more human-aligned visual representations. *Advances in Neural Information Processing Systems*, 36:65743–65765, 2023.

Shashanka Venkataramanan, Mamshad Nayeem Rizve, João Carreira, Yuki M Asano, and Yannis Avrithis. Is imagenet worth 1 video? learning strong image encoders from 1 long unlabelled video. *arXiv preprint arXiv:2310.08584*, 2023.

Peter Földiák. Learning invariance from transformation sequences. *Neural computation*, 3(2): 194–200, 1991.

Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002.

Olivier J Hénaff, Yoon Bai, Julie A Charlton, Ian Nauhaus, Eero P Simoncelli, and Robbe LT Goris. Primary visual cortex straightens natural video trajectories. *Nature communications*, 12(1):5982, 2021.

Olivier J Hénaff, Robbe LT Goris, and Eero P Simoncelli. Perceptual straightening of natural videos. *Nature neuroscience*, 22(6):984–991, 2019.

Anne Harrington, Vasha DuTell, Ayush Tewari, Mark Hamilton, Simon Stent, Ruth Rosenholtz, and William T Freeman. Exploring perceptual straightness in learned visual representations. In *The Eleventh International Conference on Learning Representations*, 2022.

X Niu, C Savin, and E P Simoncelli. Learning predictable and robust neural representations by straightening image sequences. In *Adv. Neural Information Processing (NeurIPS)*, volume 37, Vancouver, Dec 2024.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

A Appendix

A.1 Reproducibility

All code used for pretraining, evaluation, and analyses, will be made available in a public github repository upon publication.

A.2 Additional Pretraining Details

Here we report some additional hyperparameters not included in the main text.

Optimization: For all experiments we trained for 100 epochs using a batch size of 2048 and used the LARS optimizer [You et al., 2017] with weight decay of 1e-6 and momentum of 0.9. Note that the $\mathcal{L}_{CE-SSL}$ loss is evaluated on pairs of augmented views and thus had an effective batch size of 1024. We use a base learning rate of 4.8 and a learning rate schedule consisting of linear warm-up for the first 10 epochs followed by cosine decay throughout training.

Projector Architectures Each trained network uses two projectors with matching architectures. For Barlow Twins we used the architecture proposed in the original work [Zbontar et al., 2021] (3 layer MLP with hidden layer and output layer widths of 8192). For MMCR we also used a 3 layer MLP with 8192 hidden width but 512 output units (also in line with the original work [Yerxa et al., 2024]). For SimCLR we used the same projector architecture as MMCR, which is larger than the MLP described originally because subsequent work [Garrido et al., 2023a] has found that SimCLR benefits from a more expressive projector.

A.2.1 ImageNet-1k

For Barlow Twins we set the λ_{BT}, which balances the on and off diagonal loss terms, hyperparameter to $5e-3$. For SimCLR we used a temperature of $\tau = 0.15$.

A.2.2 ImageNet-100

Besides the change of dataset, the only hyperparameter change in this setting is that we increased the number of pretraining epochs from 100 to 200 to be more in line with previous work.

A.3 Online-Linear Evaluation for the Pretraining Dataset

Because frozen-linear evaluation on large datasets is computationally intensive we instead opt for online-linear classification. During pretraining the representation network outputs are detached from the gradient propogation graph and fed through a linear layer that is optimized with the standard supervised cross entropy loss. Previous work has shown that online-evaluation is very strongly correlated with frozen-linear evaluation and incurs only a minimal cost on top of self-supervised pretraining. We report the accuracies for ImageNet-1k trained networks in Fig. 5 and the smaller set of ImageNet-100 trained models in Table 3.

Model	Accuracy
MMCR	79.0%
Barlow	81.4%
SimCLR	83.5%
CE-MMCR	79.6%
CE-Barlow	81.6%
CE-SimCLR	82.5%

Table 3: In distribution accuracy on the validation set of ImageNet-100 evaluated by online-linear classification.

https://doi.org/10.52202/079017-3043

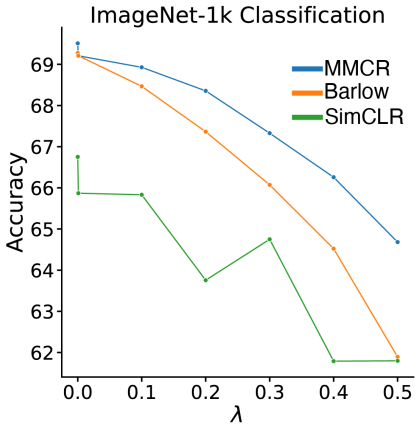

Figure 5: In distribution accuracy on the validation set of ImageNet-1k evaluated by online-linear classification, for each objective and as a function of λ.

A.4 Transfer Learning Evaluation Procedure

We closely follow the evaluation procedure from [Lee et al., 2021], we repeat the details here for completeness. First images are resized such that the shortest edge is 224 pixels, then center cropped to 224x224 resolution. Then features are extracted from train, validation, and test splits of each dataset. L-BFGS is used to optimize the standard cross entropy loss with $\mathcal{L}_2$ regularization, the value of the ridge parameter is swept over selected via performance on the validation set. Subsequently the linear classifier is retrained using both the train and validation sets, and we report the final accuracy on the held out test set.

This classification procedure was run 5 times with different random initializations, we reported the mean performance in 3.4, and report the standard deviation over runs below.

Table 4: Standard deviation of top 1 accuracies over 5 independent runs of the transfer learning evaluation procedure.

ImageNet-100 Training						
Objective	Cifar-10	Cifar-100	Pets	DTD	Flowers-102	Food-101
MMCR	1e-2	7e-3	4e-2	1e-5	7e-2	2e-1
Barlow	1e-4	2e-4	2e-2	3e-1	3e-2	9e-3
SimCLR	1e-2	7e-3	1e-2	4e-2	7e-2	5e-3
CE-MMCR	1e-2	2e-2	4e-2	3e-2	1e-1	1e-2
CE-Barlow	5e-2	2e-2	3e-2	3e-2	8e-2	1e-2
CE-SimCLR	5e-3	2e-2	2e-2	2e-2	5e-2	1e-2
ImageNet-1k Training						
Objective	Cifar-10	Cifar-100	Pets	DTD	Flowers-102	Food-101
MMCR	1e-2	2e-2	8e-2	4e-2	2e-2	7e-3
Barlow	1e-2	1e-1	1e-1	4e-2	3e-2	3e-3
SimCLR	2e-2	7e-3	1e-4	4e-2	2e-2	8e-3
CE-MMCR	9e-3	2e-2	2e-1	2e-1	6e-2	2e-2
CE-Barlow	1e-5	1e-2	9e-2	1e-5	5e-2	9e-2
CE-SimCLR	1e-2	1e-2	2e-2	2e-2	7e-2	8e-3

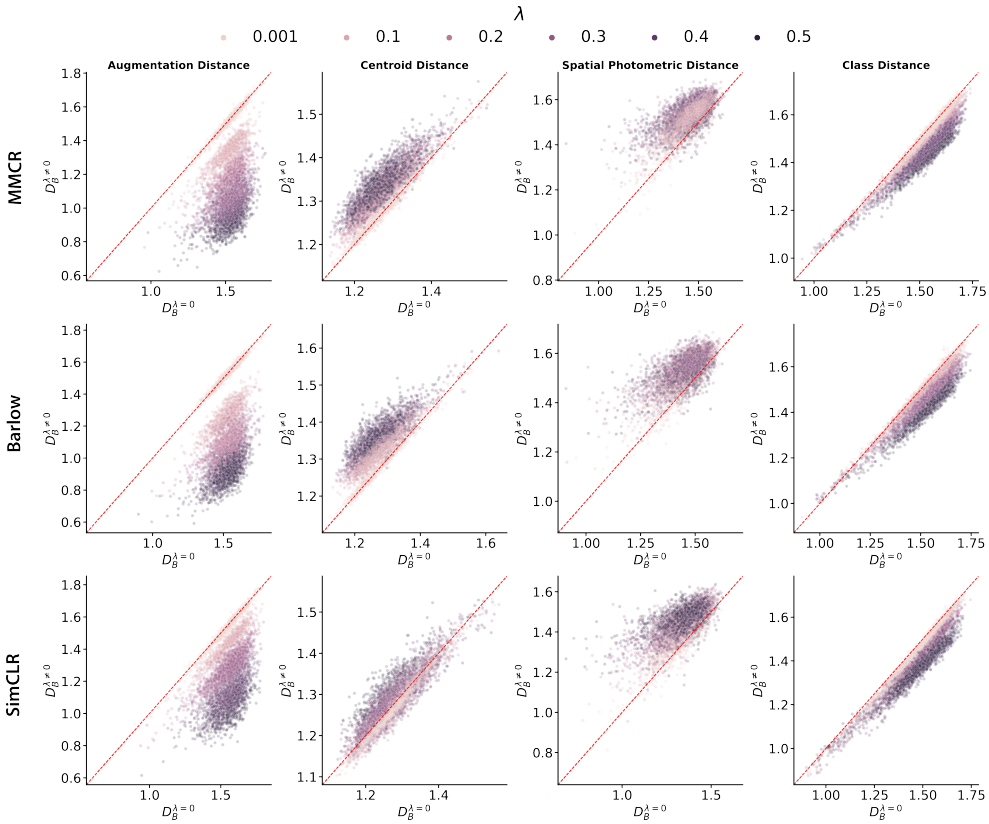

Figure 6: Joint distributions of invariant and equivariant networks (increasing λ increases the importance of the equivariant loss) for all of the Bures metric comparisons detailed in 3.2. The mean value curves in Fig. 2 are generated by taking the mean of $x - y$ for each of these plots, separately for each objective and value of λ. The confidence intervals are estimated from the distribution of $x - y$ values as well. Columns depict Bures distances between covariances estimated over different sources of variability, and columns index different base invariant objective functions.

A.5 Additional Details for Representational Analyses

Below we depict the joint distributions of the distances described in 3.2. For each setting (unique objective function and value of λ there are 800 unique measurements (i.e. points of a unique hue on an individual plot). Meaning, for example, there are 800 random pairs of augmentation manifolds compared in both invariant representation space and equivariant representation space for each equivariant network in the left most column. We summarize describe the sources of variability over which covariance matrices are estimated and the variables over which the expected Bures distance is calculated for the experiments in Fig. 2 A-D.

A.6 Out-of-distribution Equivariance

We aim to test whether learning equivariances using weak augmentations induces structured variability in response to stronger augmentations (the extent to which learned equivariances generalize beyond the range of transformations seen during training). We trained models using the Barlow Twins objective and the same sweep over values of λ using “weak" augmentations of (1) double the minimum crop size, (2) half the maximum value of color jittering, and (3) half the maximum size of Gaussian blurring kernel. We then repeat the parameter decoding experiments from the main paper Fig. 2E on these weak augmentations (left panel Fig. 7), and on the non-overlapping part of the parameter space between the weak and strong augmentation distributions, i.e only for augmentations whose parameters are in distribution for the models trained as in the main text but out of distribution for the new models trained using weaker augmentations (right panel Fig. 7). In the first panel,

https://doi.org/10.52202/079017-3043

Table 5: Table defining the sources of variability producing covariance matrices and the random variables that the expected Bures distance is computed over for the experiments in Fig. 2.

Column	**Source of Variability C_1**	**Source of Variability C_2**	**Ensemble for estimating expectation**
Augmentation-Augmentation Distance (A)	Augmentations of single image.	Augmentations of single image.	Random pairs of distinct images
Augmentation-Centroid Distance (B)	Augmentation manifold centroids over all images	Augmentations of single image.	Random images (those used to compute C_2)
Spatial-Photometric Distance (C)	Photometric augmentations (of a single image) after averaging over many random crops.	Random crops (of a single image) after averaging over many photometric augmentations.	Unique base images.
Class-Class Distance (D)	(Unaugmented) exemplars from one class	(Unaugmented) exemplars from one class	Random pairs of distinct classes

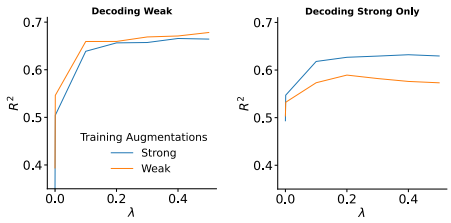

Figure 7: Augmentation parameter decoding performance on held-out test images for networks trained using either strong or weak transformations. In the left panel we plot the decoding performance for weak transformations only, and in the right panel the performance for strong transformations only (which are not seen by the weak-trained networks during pretraining).

we can see that the best parameter decoding performance occurs when the pretraining distribution of transformations is matched to the evaluation transformations (i.e. weak-trained models slightly outperform the strong-trained models at decoding the parameters of weak transformations). In the right panel we see that models trained with strong augmentations have higher decoding performance, but models trained only on weak augmentations still demonstrate significantly increased ability to linearly decode strong augmentation parameters relative to the invariant trained models ($\lambda = 0$ models), indicating a degree of generalization in the learned equivariances.

A.7 Different Backbone Architectures

To verify that the effect of equivariance neural predictivity observed in the main text is not limited to a specific choice of architecture, we trained invariant ($\lambda = 0.0$) and equivariant ($\lambda \in [0.1, 0.2]$) networks using $\mathcal{L}_{BT}$ with smaller (ResNet-34) and larger (ResNet-50) backbones. As shown in Fig. 8, we observe the same trend across different architectures: contrastive-equivariant training increases alignment to area IT as measured by linear predictivity.

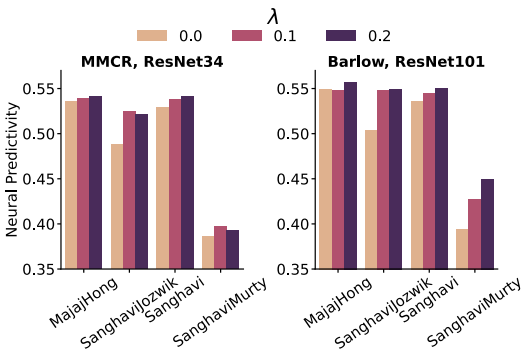

Figure 8: Neural predictivity results for models trained using the MMCR base loss and a ResNet-34 backbone (left panel), and the Barlow Twins base loss and a ResNet-101 backbone (left panel). We see similar trends in terms of increased neural predictivity for equivariant trained models as observed using the ResNet-50 backbone in the main text.

A.8 Compute Resources

All pretraining runs used 8 A100 Nvidia GPUs with 40GB of memory each. In our setting pretraining run times were around 15 hours, and we note that CE-SSL training generally increased training time by approximately 10% relative to standard self-supervised training. Subsequent evaluations ran on a single A100.

A.9 Limitations

We discuss some limitations not addressed in the discussion here. Our current training setup requires selection of the hyperparameter λ to balance between the equivariant and invariant loss functions. Future work could investigate methods to balance the two losses without explicitly training an individual network for each choice of λ. It is worth noting that some invariant SSL methods may be more or less sensitive to this additional hyperparameter. For example, some of the non-monotonicity of SimCLR curves in Fig. 2 may suggest that SimCLR representations are more difficult to smoothly shape within the CE-SSL paradigm.

Additionally, computational limitations prevent us from doing an extensive architecture search over the two projector networks employed in contrastive equivariant training. In addition to varying the depth and width of each projector, it would be of interest to "split" the representation space and have each projector operate on a subset of dimensions as input (as in SIE [Garrido et al., 2023b]). Additonally computational limitations prevented us from extensively evaluating the variability in neural predictiivity over independent runs of contrastive equivariant training (the BrainScore framework recently stopped providing estimates of the error of neural predicitivity, but in previous studies the reported error for IT predicitivity was $\approx$ 3e-3, which is small relative to the variability we observed across the parameter of interest λ). Pilot experiments indicated to us that the variability over training runs was small relative to the variability over values of lambda (we trained two invariant models and two with $\lambda = 0.1$ to get a rough estimate of this, in both cases we kept the more predictive model for inclusion in all analyses that appear in this paper).

A.10 Broader Impacts

In this work we propose one strategy for inducing increased alignment between artificial and biological visual representations. Better understanding the computational principles underlying visual representations has the potential to benefit the quality of computer vision applications, to offer insights into the structure of the primate visual system, and to improve clinical treatment of disorders related to visual perception.

https://doi.org/10.52202/079017-3043

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and precede the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction are directly related to the each of the results in the results section of the paper

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: The limitations are discussed in an appropriately labelled section of the appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper only presents empirical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Each experiment is described in the main text and exact details needed to reproduce the results are described in relevant appendices. Furthermore, source code will be released upon publication.

https://doi.org/10.52202/079017-3043

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Data sources are clearly documented and source code will be released in the form of a public repository upon publication.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Most experiments are described adequately in the main text and exhaustive details are given in relevant appendices.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All experiments we conducted have clearly defined measures of significance. We relied on a prominent 3rd party software to perform neural predictivity measurements (BrainScore), and no measurement of significance was provided to us. We discuss this limitation in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Computing resources are described in the appendix.

https://doi.org/10.52202/079017-3043

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: No special circumstances required any deviation from the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader impacts of the work are described in an appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All creators of assets (in this paper, datasets) are properly cited and the terms of use are properly respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No new assets are released by the paper.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

https://doi.org/10.52202/079017-3043

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: There are no participants involved in this study.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.