

# STRETCHING BEYOND THE OBVIOUS: A GRADIENT-FREE FRAMEWORK TO UNVEIL THE HIDDEN LANDSCAPE OF VISUAL INVARIANCE

**Lorenzo Tausani<sup>1</sup>, Paolo Muratore<sup>1</sup>, Morgan B. Talbot<sup>2,3,4</sup>, Giacomo Amerio<sup>1</sup>,  
Gabriel Kreiman<sup>2,3</sup>, Davide Zoccolan<sup>1</sup>**

<sup>1</sup>Neuroscience Area, International School for Advanced Studies (SISSA), Trieste (Italy)

<sup>2</sup>Boston Children’s Hospital, Harvard Medical School, Boston (USA)

<sup>3</sup>Center for Brains, Minds, and Machines, MIT, Cambridge (USA)

<sup>4</sup>Harvard-MIT Program in Health Sciences and Technology, MIT, Cambridge (USA)

## ABSTRACT

Uncovering which feature combinations are encoded by visual units is critical to understanding how images are transformed into representations that support recognition. While existing feature visualization approaches typically infer a unit’s most exciting images, this is insufficient to reveal the manifold of transformations under which responses remain invariant, which is critical to generalization in vision. Here we introduce Stretch-and-Squeeze (SnS), an unbiased, model-agnostic, and gradient-free framework to systematically characterize a unit’s maximally invariant stimuli, and its vulnerability to adversarial perturbations, in both biological and artificial visual systems. SnS frames these transformations as bi-objective optimization problems. To probe invariance, SnS seeks image perturbations that maximally alter (stretch) the representation of a reference stimulus in a given processing stage while preserving unit activation downstream (squeeze). To probe adversarial sensitivity, stretching and squeezing are reversed to maximally perturb unit activation while minimizing changes to the upstream representation. Applied to CNNs, SnS revealed invariant transformations that were farther from a reference image in pixel-space than those produced by affine transformations, while more strongly preserving the target unit’s response. The discovered invariant images differed depending on the stage of the image representation used for optimization: pixel-level changes primarily affected luminance and contrast, while stretching mid- and late-layer representations mainly altered texture and pose. By measuring how well the hierarchical invariant images obtained for  $L_2$ -robust (i.e., adversarially trained) networks were classified by humans and other observer networks, we discovered a substantial drop in their interpretability when the representation was stretched in deep layers, while the opposite trend was found for standard (i.e., not robustified) models. This indicates that  $L_2$  adversarial training fails to increase the interpretability of high-level invariances, despite good perceptual alignment between humans and robustified models at the pixel level. This demonstrates how SnS can be used as a powerful new tool to measure the alignment between artificial and biological vision.

## 1 INTRODUCTION

Both visual neuroscience and deep learning seek to understand image processing systems composed of millions of interacting functional units, whose activity patterns are shaped by their experience with natural image statistics (Matteucci et al., 2024; Barrett et al., 2019). This common goal raises a fundamental question in both fields: which combination of image features do visual neurons become tuned for? Traditionally, this question has been addressed by developing feature visualization approaches that discover the “preferred” stimuli that maximally activate a given unit - often referred to as the unit’s most exciting images (MEIs) (Olah et al., 2017; Nguyen et al., 2019; Fel et al., 2023; Xiao & Kreiman, 2019; Walker et al., 2019). MEIs, however, only reveal a few instances within the

vast set of images that strongly activate a unit (Nguyen et al., 2016; Cadena et al., 2018), offering poor insight into the manifold of transformations under which the unit’s activity remains invariant.

To overcome this limitation, we developed Stretch-and-Squeeze (SnS), an unbiased, model-agnostic method to probe visual invariance in both artificial and biological neurons. Building upon an existing framework for finding MEIs (Xiao & Kreiman, 2020), SnS optimization exploits evolutionary algorithms, but with a different objective: to find invariant or adversarial images. SnS integrates the search for invariant images and adversarial examples within a unified optimization objective, generating image perturbations starting from a reference image (e.g., a MEI of the unit). To probe a unit’s invariance, SnS explicitly looks for images that are maximally distinct from this reference stimulus in the representation of a chosen visual processing stage, while preserving the response of a target unit downstream. This allows SnS to sample the invariance manifolds of the selected unit when the representation of an effective stimulus (either a MEI or a natural image) is stretched at different processing stages. As a result, our approach reveals the actual image variation axes the unit is able to tolerate, providing a richer, more veridical description of its invariance landscape, as compared to traditional tests based on predefined (e.g., affine) transformations (Goodfellow et al., 2009; Kheradpisheh et al., 2016; Engstrom et al., 2019c). In our study, we carried out comprehensive tests of the effectiveness of SnS to achieve the goals listed above, using a popular CNN (ResNet50) as a benchmark. To study invariances of the whole model with respect to specific image classes, we applied SnS to the network’s readout units, as they are selective for the object categories the model has learned to classify (e.g., a ‘cup’ neuron). We found dramatic differences among the invariance landscapes of these units when stretching the representations at different depths of the processing hierarchy. Stretching early, middle, and late representations yielded invariant images that differed from the reference (and among themselves) in terms of luminance/contrast, texture, and pose, respectively. We also discovered important differences in these hierarchical invariances between the standard and an adversarially trained version of the network. Like previous studies (Feather et al., 2023) we found that invariant images from the  $L_2$ -robust network were more recognizable by human subjects and other *observer* networks. However, we identify a striking point of divergence: while robust CNN invariances become less interpretable when stretching at deeper layers, those of standard networks become more interpretable, causing the robust network’s advantage to erode in later layers. Importantly, being gradient-free and model-agnostic, SnS can uncover the invariance fields of units in “black-box” image processing systems, where access to hidden units is either absent or very limited. Neurophysiologists face this scenario when they want to infer the tuning properties of biological visual neurons. To test the applicability of SnS in neurophysiology experiments, we verified that the method works even if the experimenter can record the activity of just a small fraction of the units in the processing stage where the stretching is applied.

## 2 RELATED WORKS

**Probing CNN representations: feature visualization, invariance, and adversarial examples.** The question of functional interpretability and feature visualization in deep learning has predominantly been approached by employing gradient-based optimization for image synthesis, taking advantage of the complete analytical description and differentiability of the network (Olah et al., 2017; Nguyen et al., 2019; Fel et al., 2023). While most studies have focused on finding the MEIs of CNNs’ units, recent efforts have started to also explore their invariance landscapes. Most of them have focused on providing a local measure of model invariance around a given input, probing the shape of the representation manifold in the vicinity of a chosen image (Berardino et al., 2017; Hénaff & Simoncelli, 2015). Another prominent approach relies on discovering *model metamers* - i.e., synthesized stimuli that match the internal representation of a reference (natural) image in a specific layer of a network (Mahendran & Vedaldi, 2015). Using metamers, Feather et al. (2023) were able to demonstrate that standard CNNs display highly idiosyncratic invariances at the top layers of processing, with metamers being unintelligible to human observers or other neural networks. In contrast, CNNs trained to be robust to adversarial images (i.e., imperceptibly modified inputs capable of altering CNN object classification (Szegedy et al., 2013)) yielded metamers that were substantially more interpretable by human observers. This shows how invariance and adversarial vulnerability are conceptually related (Jacobsen et al., 2020). For instance, adversarially trained (“robust”) networks can craft subtle image perturbations that not only fool the network but can also impair (Gaziv et al., 2023) or enhance (Talbot et al., 2025) human object recognition, thereby reinforcing the perceptual parallels between robust network representations and invariance in human vision.

**Applications to visual neuroscience.** Advances in CNN interpretability have also impacted neuroscience, which has long suffered from a lack of effective methods to investigate the functional tuning of visual neurons in higher cortical areas. By leveraging the strong functional analogy between CNNs trained for image classification and the primate object recognition pathway (known as the *ventral stream* (DiCarlo et al., 2012)), it is possible to create *digital twins* of the ventral stream using CNNs (Yamins et al., 2014; Yamins & DiCarlo, 2016; Schrimpf et al., 2018). This allows employing gradient-based optimization to synthesize images that modulate the activity of biological neurons (Bashivan et al., 2019) or investigate properties like adversarial robustness in visual neurons (Guo et al., 2022). The digital twin paradigm has also been extended beyond primate vision, informing models of the auditory cortex (Kell et al., 2018) and visual processing in rodents (Nayebi et al., 2023; Walker et al., 2019; Tong et al., 2023). Importantly, related gradient-based techniques (Cadena et al., 2018) have also been used to map invariance in individual neurons, for instance, by looking for images that maximally activate a mouse primary visual neuron while being maximally distinct in pixel space (Ding et al., 2023). Other recent approaches (Baroni et al., 2023; Bashiri et al., 2025) used Compositional Pattern-Producing Networks (CPPNs) to map low-dimensional (1 and 2d) invariance manifolds of digital twin neurons of macaque primary visual cortex, successfully recovering visual invariances established by previous neurophysiological and computational work (Schwartz et al., 2006; Sharpee, 2013).

Obviously, these approaches have an intrinsic limitation: they are only as good as the fidelity of the digital twin at capturing the selectivity of visual neurons. To overcome this constraint, gradientless feature visualization methods like XDREAM have successfully synthesized effective MEIs for units in both the primate ventral stream (Ponce et al., 2019) and artificial neural networks (Xiao & Kreiman, 2019) by relying on evolutionary algorithms. While recent work has also begun to explore the feature landscape around the MEIs obtained with XDREAM (Wang & Ponce, 2022b), current gradientless approaches have not been systematically applied to characterize the invariance of visual tuning in artificial and biological architectures. Hence the novelty of SnS, which is, to the best of our knowledge, the first gradientless approach to systematically infer the invariance manifolds of visual units.

### 3 METHODS

#### 3.1 THE STRETCH-AND-SQUEEZE ALGORITHM (SNS)

SnS consists of three key components: a generative model  $\psi$ , a test (or *subject*) network  $\phi$  and a gradient-free optimizer (Fig. 1a). The generative model is a pretrained deep neural network (Dosovitskiy & Brox, 2016) that maps  $n$ -dimensional vectors  $\xi^t \in \mathbb{R}^n$  (referred to as *codes*) to RGB images:  $\mathbf{x} = \psi(\xi) \in \mathcal{X} \subseteq \mathbb{R}^{C \times H \times W}$ . In all the experiments,  $n = 4096$ . Crucially, the generative model  $\psi$  was trained on naturalistic stimuli and embodies a powerful prior over the distribution of possible images. We use the Covariance Matrix Adaptation Evolutionary Strategy (CMA-ES) optimizer (Hansen et al., 2003) to adjust the codes and iteratively improve on our objective. At each iteration  $t$  the optimizer yields a new set of codes  $\xi^{t+1} \in \mathbb{R}^n$ , which in turn is used to generate a new batch of images. A core tenet of the SnS algorithm is the relational construction of its fitness function. We start by introducing two layer indices  $\kappa$  and  $\ell$  for our test network  $\phi$ , where for convention we include the input stage as  $\kappa = 0$ , and the measuring function  $\Gamma(\mathbf{x}, \phi^\ell) \in \mathbb{R}^d$ , which returns the activations  $\mathbf{a}^\ell$  of all the  $d$  units in layer  $\ell$ , when the network  $\phi$  is presented with stimulus  $\mathbf{x}$ . We then identify a reference stimulus  $\mathbf{x}_{\text{ref}}$  from which we construct the pair of reference states  $(\mathbf{a}_{\text{ref}}^\kappa, \mathbf{a}_{\text{ref}}^\ell)$  as  $\mathbf{a}_{\text{ref}}^{\kappa, \ell} = \Gamma(\mathbf{x}_{\text{ref}}, \phi^{\kappa, \ell})$ . Lastly, we introduce two optimization objectives as either the minimization  $\mathcal{L}_{\text{squeeze}}$  or maximization (i.e. minimization of the negative)  $\mathcal{L}_{\text{stretch}}$  of the euclidean distance of a given state  $\mathbf{a}^\kappa = \Gamma(\mathbf{x}, \phi^\kappa)$  from the corresponding reference state:

$$\mathcal{L}_{\text{stretch}}^\kappa(\mathbf{a}^\kappa, \mathbf{a}_{\text{ref}}^\kappa) = -\|\mathbf{a}^\kappa - \mathbf{a}_{\text{ref}}^\kappa\|_2, \quad \mathcal{L}_{\text{squeeze}}^\kappa(\mathbf{a}^\kappa, \mathbf{a}_{\text{ref}}^\kappa) = +\|\mathbf{a}^\kappa - \mathbf{a}_{\text{ref}}^\kappa\|_2. \quad (1)$$

Then, for a given choice of layer indices  $\kappa, \ell$  and reference states  $\mathbf{a}_{\text{ref}}^{\kappa, \ell}$ , indicating with  $\mathcal{X} \subseteq \mathbb{R}^{C \times H \times W}$  the set of input images, we define  $f: \mathcal{X} \rightarrow \mathbb{R}^2$  as:

$$f : \mathbf{x} \mapsto \begin{pmatrix} \mathcal{L}_{\text{stretch}}^{\kappa}(\Gamma(\mathbf{x}, \phi^{\kappa}), \mathbf{a}_{\text{ref}}^{\kappa}) \\ \mathcal{L}_{\text{squeeze}}^{\ell}(\Gamma(\mathbf{x}, \phi^{\ell}), \mathbf{a}_{\text{ref}}^{\ell}) \end{pmatrix}, \quad (2)$$

and introduce the bi-objective minimization problem:  $\min_{\xi} (f_1(\mathbf{x}), f_2(\mathbf{x}))$ , where  $\mathbf{x} = \psi(\xi)$ .

We define  $\Xi_{\text{sns}}$  as the formal solution to such bi-objective optimization problem in the Pareto sense, i.e.  $\Xi_{\text{sns}}$  is the collection of Pareto-optimal solutions  $\mathbf{x}^* \in \mathcal{X}$ . In particular, denoting a solution  $\mathbf{x}_1$   $f$ -dominant with respect to  $\mathbf{x}_2$  by  $\mathbf{x}_1 \succ_f \mathbf{x}_2$ , where we write  $\mathbf{x}_1 \succ_f \mathbf{x}_2 \iff f_i(\mathbf{x}_1) \leq f_i(\mathbf{x}_2) \forall i$  and  $\exists i$  for which the inequality holds strictly, we can write the  $\Xi_{\text{sns}}$  collection of solutions as:

$$\Xi_{\text{sns}} \equiv \{\mathbf{x}^* \in \mathcal{X} : \{\mathbf{x} \in \mathcal{X} : \mathbf{x} \succ_f \mathbf{x}^*\} = \emptyset\}. \quad (3)$$

In practice, for each round of optimization, we sorted the fitness scores by organizing them in Pareto fronts (Deb, 2011) (see Supplementary Material, Section A.3 for further details). In order to make our formulation more transparent to different components, we introduce the following shorthand notation for  $\Xi_{\text{sns}}$  and refer to equation 3 for its exact definition:

$$\Xi_{\text{sns}} \equiv \arg \min_{\mathbf{x} \in \mathcal{X}} \left[ \mathcal{L}_{\text{stretch}}^{\kappa}(\Gamma(\mathbf{x}, \phi^{\kappa}), \mathbf{a}_{\text{ref}}^{\kappa}), \mathcal{L}_{\text{squeeze}}^{\ell}(\Gamma(\mathbf{x}, \phi^{\ell}), \mathbf{a}_{\text{ref}}^{\ell}) \right]. \quad (4)$$

While this formulation is general and in principle supports arbitrary choices for the layer indices  $\kappa$  and  $\ell$ , as well for the corresponding reference states  $\mathbf{a}_{\text{ref}}^{\kappa, \ell}$ , in our experiments we restricted the scope of the optimization problem to investigate specific properties of the test network  $\phi$  (e.g. robustness, invariance) by making the following design choices. First, we singled out a target unit  $u_{\text{targ}}^{\ell}$  in layer  $\ell$  (see below for details) and considered  $\mathbf{x}^* \in \mathcal{X}$  the maximally exciting stimulus (i.e., the MEI) for our target unit, computed via 500 iterations of the XDREAM algorithm (Ponce et al., 2019). Our reference state for layer  $\ell$  was then  $a_{\text{ref}}^{\ell} = \Gamma(\mathbf{x}^*, \phi_u^{\ell}) \in \mathbb{R}$ , i.e. the scalar activation of the target readout unit for its MEI. We then explored the set of solutions  $\Xi_{\text{sns}}$  as we varied  $\kappa < \ell$ . In particular, we can express both the search for adversarial attacks or invariant stimulus for our target unit  $u_{\text{targ}}^{\ell}$  as the following two optimization problems:

$$\Xi_{\text{inv}} \equiv \arg \min_{\mathbf{x} \in \mathcal{X}} \left[ \mathcal{L}_{\text{stretch}}^{\kappa}(\Gamma(\mathbf{x}, \phi^{\kappa}), \Gamma(\mathbf{x}^*, \phi^{\kappa})), + |a_u^{\ell} - a_{\text{ref}}^{\ell}| \right] \quad (5)$$

$$\Xi_{\text{adv}} \equiv \arg \min_{\mathbf{x} \in \mathcal{X}} \left[ \mathcal{L}_{\text{squeeze}}^{\kappa}(\Gamma(\mathbf{x}, \phi^{\kappa}), \Gamma(\mathbf{x}^*, \phi^{\kappa})), - |a_u^{\ell} - a_{\text{ref}}^{\ell}| \right], \quad (6)$$

where we have used  $\Gamma(\mathbf{x}, \phi_u^{\ell}) = a_u^{\ell}$  and wrote  $\mathcal{L}^{\ell} = \pm |a_u^{\ell} - a_{\text{ref}}^{\ell}|$ .

This dual formalization reflects the fact that both invariance and robustness relate changes of high order representations to changes at the input level (Fig. 1b). However, we remark that a key flexibility of SnS is that its general formalization for the *stretch* and *squeeze* objectives allows computation not only in the input pixel space, but also in any intermediate representation stages  $\kappa$  and  $\ell$  of the network. This allows probing invariance (or adversarial attacks) at different levels of feature abstraction.

For the characterization of invariance, we set this representation space at three distinct hierarchical levels within ResNet50: (i)  $\kappa = 0$ , i.e. the input pixel space (denoted `low_level`), (ii) a mid-level convolutional layer (1<sup>st</sup> convolutional stage in layer 3, `mid_level`), and (iii) a deep convolutional layer (7<sup>th</sup> convolutional stage of layer 4, `high_level`). Our primary target units  $u_{\text{targ}}^{\ell}$  were chosen in the final readout layer (fully connected), as these units are selective for specific object categories, allowing us to study the invariance of the whole network to transformations of the image classes those neurons are tuned to. To demonstrate SnS’s generalizability beyond category-selective units, we also performed optimizations targeting units within `mid_level` and `high_level`, which are potentially more analogous to biological neurons in the visual system. In these cases, the input-side distance was computed in pixel space (i.e.  $k = 0$ ). For adversarial attack generation, we restricted our analysis to the configuration targeting readout layer units and defining input distance in pixel space. SnS optimizations were conducted on the units of two specific instances of ResNet50, our subject network: the standard ImageNet-pretrained model available in PyTorch (Paszke et al., 2019), and a robust counterpart generated via adversarial training with an  $L_2$  perturbation norm constraint

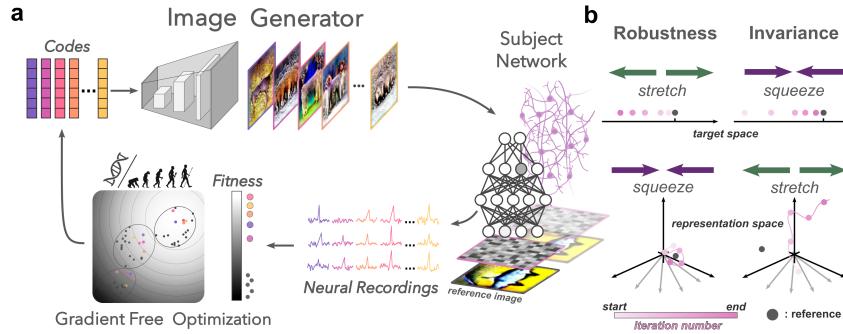


Figure 1: **The Stretch-and-Squeeze (SnS) algorithm.** (a) Overview of the SnS algorithm: candidate stimuli are synthesized from latent codes via the image generator, and the activation response of target units is recorded and used as fitness score by the optimizer, that adjusts a new set of codes. (b) Dual fitness objectives in SnS. To probe *invariance* (right), SnS maximizes stimulus distance from a reference in the representation space (stretch) and minimizes the variation in the activation of a target unit (squeeze). Conversely, to synthesize *adversarial examples* (left), SnS minimizes changes in the representation space, while maximizing the variation in the activation of the target unit.

of  $\epsilon = 3$  (Engstrom et al., 2019a). SnS invariances were studied also in ResNet18 and VGG16\_bn (see Section 3.2) and Vision Transformers (ViT, see Supplementary Material, Section I).

Initialization strategies were adapted based on the optimization goal: for adversarial attacks, the search was initialized from the MEI  $x^*$ , for invariance experiments, initialization began from random normally distributed vectors. For additional details regarding the computational experiments we refer to Supplementary Material, Section A.

### 3.2 INTERPRETABILITY OF THE INVARIANT IMAGES BY HUMANS AND OBSERVER NETWORKS

To assess whether the invariant images produced by SnS retain sufficient information for correct classification by other visual systems, we tested the ability of humans or other *observer* neural networks to classify them, using a 12-alternative forced choice task (AFC). We used several distinct ImageNet-pretrained architectures as observer networks (ResNet18, VGG16\_bn, Wide-ResNet50-2, DenseNet161, ResNeXt50\_32x4d, Shufflenet\_v2\_x1\_0), both standard obtained from PyTorch (Paszke et al., 2019), and robust (all  $L_2$ ,  $\epsilon = 3$ ), obtained from Salman et al. (2020). To assess human recognizability of invariant images, we recruited 25 human participants using the online Prolific platform. Details on the procedures to run these interpretability experiments are reported in Supplementary Material, Section B.

## 4 RESULTS

### 4.1 SNS GENERATES EFFECTIVE ADVERSARIAL AND INVARIANT IMAGES

We first validated the SnS framework by generating invariant and adversarial images for a sample of 77 readout units of a  $L_2$ -robust ResNet50 using their MEIs  $x^*$  as reference images, and applying the stretching (to achieve invariance) or the squeezing (to achieve adversarial images) to the pixel representation (see Section 3.1). For each unit, 10 independent invariant and adversarial images were synthesized, each with a different initialization seed, and distance metrics with respect to the reference MEI were aggregated as the mean over the synthesized images. These metrics were then further aggregated across units as mean  $\pm$  SD and reported in Fig. 2. As shown in the figure, SnS successfully generated effective adversarial examples (top left). These stimuli substantially suppressed the activation of the readout units relative to their MEIs (mean reduction of  $111\% \pm 7\%$ ), being displaced from the MEIs by a mean  $L_2$  distance of  $72 \pm 12$  pixels. This relatively large pixel budget reflects both the network robustness and our stringent unit-silencing objective - a stricter criterion than mere misclassification. Consistent with Tsipras et al. (2018) and Gaziv et al. (2023), perturbations were semantically relevant, not noise-like (Fig. 2).

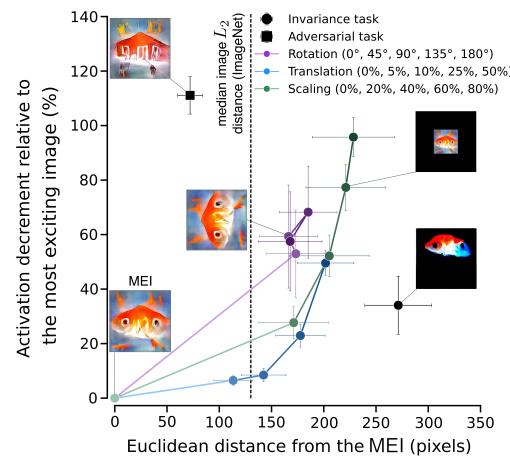
The invariant images generated by SnS (bottom right of Fig. 2) were also very effective, yielding only a minor drop of the units’ activation relative to the MEIs (mean reduction of  $34\% \pm 11\%$ ). At the same time, they substantially departed from the MEIs, with a mean  $L_2$  pixel distance ( $271 \pm 32$  pixels) that considerably exceeded the median distance between ImageNet images, as reported by Gaziv et al. (2023). More impressively, SnS discovered image transformations that were more extreme (in terms of pixel distance) than those achieved by applying to the MEI standard data augmentations (colored dots/lines), while altering the activation of the readout units substantially less than the most extreme augmentations (darker dots). This indicates that SnS can discover the actual axes of image variation a unit learned to tolerate, exploring invariance manifolds more precisely than traditional tests with predetermined (e.g., affine) transformations. The same result was obtained when we applied SnS to discover invariant and adversarial images for a subset of readout units, but taking highly effective natural images as references instead of the MEIs (see Fig. S5a). See also Fig. S5b for additional examples of invariant and adversarial images obtained using as references either MEIs or natural images. The ability to obtain invariant and adversarial images is enabled intrinsically by the dual loss formulation of the SnS optimization. Indeed, when images are synthesized using only one of the two loss components at a time, results diverge drastically from those of SnS (see Fig. S6 in Section D of the Supplementary Materials).

Finally, we note that SnS successfully generated invariant images also for units in intermediate convolutional layers (Supplementary Material, Section E). This demonstrates its applicability to the analysis of latent representations that are more analogous to those of biological neurons.

#### 4.2 SNS REVEALS LAYER-SPECIFIC INVARIANT MANIFOLDS

Next, we compared the kinds of invariant images that SnS discovered by stretching the MEI representations of the 77 readout units at different processing stages along a  $L_2$ -robust ResNet50 - i.e. in the pixel space (as done in Fig. 2), as well as in an intermediate (Layer3\_conv1) and a deep (Layer4\_conv7) convolutional layer. Qualitatively, this yielded different sorts of invariances (Fig. 3a). Stretching in the pixel-space produced mainly luminance and contrast changes; stretching mid-level representations mainly affected texture and color; stretching high-level representations tended to produce abstract variations like viewpoint changes or multiple object instances. The same qualitative differences were observed when SnS was used to produce invariant images for 77 readout units of *standard* ResNet50 (Fig. 3b), although, in this case, the images appeared to be less interpretable than those obtained for the robust counterpart. Also, the images lacked the high-frequency patterns that are typically found in the MEIs of standard networks units (Engstrom et al., 2019b), suggesting that the SnS generator strongly regularizes the image search towards natural statistics.

The qualitative descriptions above are just approximate semantic labels to concisely describe image transformations that are, in fact, too rich and complex to be readily captured by a textual description (e.g., stretching in pixel space also produced size changes). To quantify these observations more objectively, we applied PCA to the invariant images yielded by SnS. We then used the first  $k$  principal components as feature vectors to train a Support Vector Classifier (SVC) to classify the invariant images according to the layer where the SnS stretching was applied (for details see Supplementary Material, Section A.4). As shown in Fig. 3c, a few components were enough to yield above



**Figure 2: SnS generates effective adversarial and invariant examples.** The average activation reduction ( $\pm$  SD) of  $n = 77$  readout units of a  $L_2$ -robust ResNet50 (relative to their MEIs) is plotted against the  $L_2$  pixel distance from the MEIs, when the latter were transformed with SnS to yield either adversarial or invariant images, or were subjected to affine transformations (rotation, translation and scaling). The pixel distance refers to  $224 \times 224$  RGB images with values between 0 and 1.

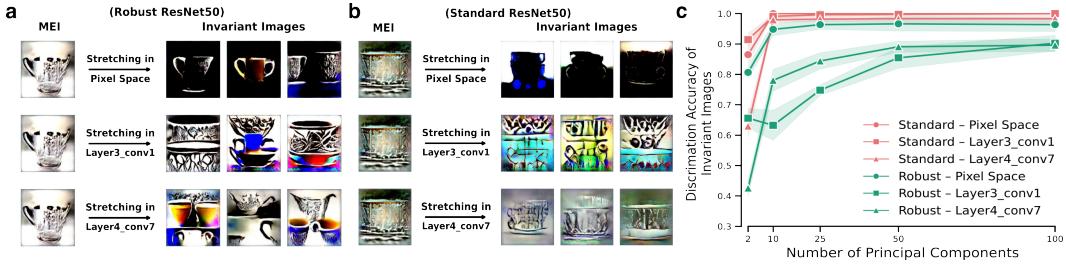


Figure 3: **SnS discovers layer-specific invariances.** (a) Example MEI and associated invariant images obtained for a readout neuron (“cup”) in  $L_2$ -robust ResNet50 and computed for three different choices of the stretching representational stages (rows). Multiple results are shown for different random initialization seeds (columns) (b) Same as (a) but for a standard ResNet50. (c) Accuracy of a SVC in discriminating the three classes of invariant images produced by stretching pixel-, mid-, and high-level representations as a function of the number of principal components fed to the classifier.

chance performance (i.e.,  $> 33.3\%$  correct), and a few tens of components enabled near-perfect discrimination for the standard network (red) and above 80% correct for the robust network (green).

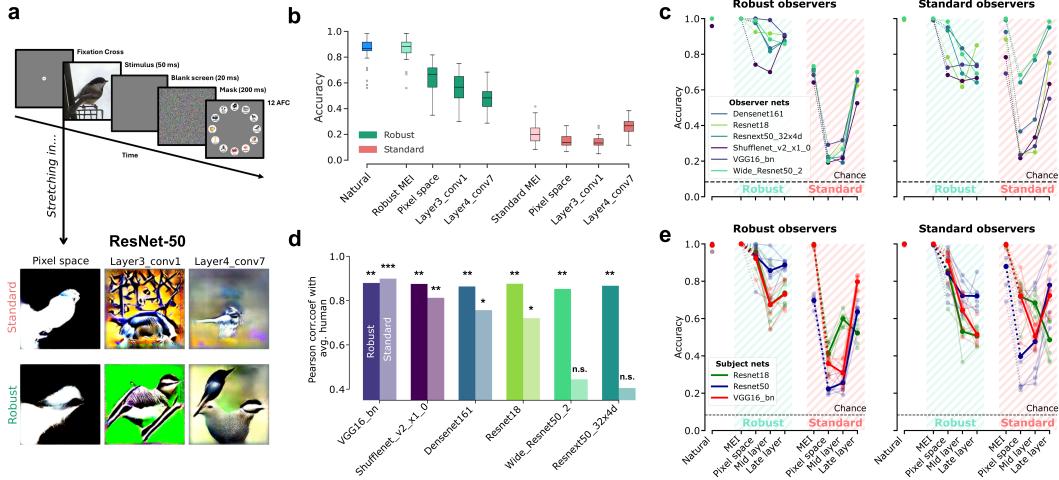
To characterize how the invariant images generated by stretching at a given processing stage were represented across the network, we measured their  $L_2$  distance from their reference MEIs in the representation provided by each layer of the network (see Supplementary Material, Section A.5.2 for detail). In addition to demonstrating the effectiveness of SnS to explore invariances far away from their reference MEI and heterogeneous across multiple runs (see Section A.5.2 for detail), the representational distance between the invariant images and their MEIs was preserved in the layers that were adjacent to the one where the stretching was applied, revealing a clear hierarchical pattern (Fig. S3 a,b). Stretching in the pixel-space yielded invariant images that remained dissimilar from the MEIs in the first layers of the network; stretching in mid and deep convolutional layers yielded images that were more different from the MEIs in the central and final portions of the network.

Yet another way to characterize the invariance manifolds is to estimate their intrinsic dimension (ID). We did so for an example readout unit of  $L_2$ -robust ResNet50, using both PCA and 2NN-ID, a state-of-the-art nonlinear estimator (Facco et al., 2017) (see Section F in the Supplementary Material). 2NN-ID yielded ID estimates one order of magnitude lower than PCA, indicating that the geometry of the three manifolds was highly nonlinear (Fig. S8). Both methods, however, revealed a consistent ID ranking across stretching layers: lowest for the pixel space, highest for mid layers, and lower again for deep layers - a pattern mirroring previously reported trends in the ID of image representations across deep CNNs and visual cortex (Ansini et al., 2019; Muratore et al., 2022a)

**Hierarchical invariances are robust to representation subsampling.** To assess SnS’s applicability to neuroscience, where only a small neuronal subpopulation is typically recorded from a visual area, we generated the invariant images using heavily subsampled hidden layer representations. Both qualitative inspection and representational distance analysis revealed that the resulting invariances closely resembled those obtained using the full layer representation, supporting SnS’s potential for neuroscience experiments (see Supplementary Material, Section G for further details).

#### 4.3 INVARIANT IMAGES FOR ROBUST AND STANDARD NETWORKS ARE PERCEPTUALLY DIFFERENT

We next evaluated how recognizable the invariant images generated by SnS were by humans and other CNNs. In our tests, invariant images for both kinds of networks (i.e., standard and  $L_2$ -robust) were generated by stretching the low- (pixel), mid- and high-level representations of the same set of natural images sampled from 12 different Imagenet categories (Fig. 4a). Human subjects performed a 12-AFC classification task on these invariant images, their reference images, and MEI controls (Fig. 4a, top; for details see Section 3.2 and Supplementary Section B). Invariant images from the robust network (all generation stages) were significantly more recognizable by humans than their standard network counterparts (Fig. 4b, compare the green to the red boxes;  $p < 0.001$ , Supplementary



**Figure 4: Interpretability of the invariant images by humans and other networks.** (a) Illustration of the human classification task with the invariant images generated by SnS for standard and  $L_2$ -robust networks. (b) Classification accuracies of humans (averaged across subjects and categories) displayed for each experimental condition. (c) Classification accuracies across multiple  $L_2$ -robust (left) and standard (right) networks. (d) Correlation coefficient between average human performances and the performances of each observer model. For each architecture, the darker color indicates the robust model and the lighter color indicates the standard version. (e) Analysis from (c) extended to multiple architectures: ResNet50, ResNet18, and VGG16\_bn. Solid lines indicate the average between different observer networks (translucent lines).

Material Table S2). Invariant images from the  $L_2$ -robust network were more recognizable when stretching was applied to earlier-layer representations, with pixel-space stretching yielding the highest discrimination accuracy. Interestingly, an opposite trend was observed for the standard network: invariant images from a late layer were more recognizable than those from a middle layer or from pixel space ( $p < 0.001$  for each of the above comparisons: see Supplementary Material Table S2). These findings provide a more granular understanding of the perceptual misalignment between humans and CNNs, extending beyond what classic feature visualization tools can reveal with a direct comparison of MEIs. In fact, human participants classified the MEIs obtained for the robust network (Fig. 4b, cyan box) with an accuracy comparable to that achieved on the natural images encoded by the corresponding readout units (Fig. 4b, blue box). By comparison, human classification of the MEIs obtained for the standard network was relatively poor (Fig. 4b, pink box). Therefore, considering only MEIs' interpretability, the  $L_2$  robustification process would seem very effective in increasing the alignment of image classification between ResNet50 and human observers. SnS reveals instead that this alignment gradually deteriorates when considering invariant images obtained by stretching the representation in progressively deeper layers, with the interpretability gap between the robust and standard networks progressively narrowing. Interestingly, this also shows how SnS can uncover trends that are complementary to those found by previous approaches aimed at assessing invariance, such as computation of metamers (Feather et al., 2023). Similarly, SnS provides a complementary tool to test model alignment with human perception with respect to previous approaches which study alignment through model disagreement (Wang & Simoncelli, 2008; Golan et al., 2020), as well as to other approaches that have examined human interpretability of parametric image transformations at the pixel or patch level (Ullman et al., 2016; Malik et al., 2023).

The same classification task was administered to various *observer* CNNs (both standard and  $L_2$ -robust). The alignment with the accuracy trends found in humans was relatively high for all robust observer networks (compare Fig. 4b to 4c, left), while their standard counterparts displayed a more heterogeneous degree of consistency with human choices (compare Fig. 4b to 4c, right). As a result, correlations with human perception were very high and statistically significant for all robust networks, but only for a fraction of the standard ones (Fig. 4d). Similarly to humans, all  $L_2$ -robust observer networks better classified the invariant images obtained for  $L_2$ -robust ResNet50 than for its standard

counterpart, with this difference narrowing for higher-order invariances. This demonstrates that, despite the  $L_2$  robustification process, the invariance landscape of ResNet50 readout units remained idiosyncratic and not fully generalizable to other human and artificial observers. This conclusion held true also when standard networks were used as observers, although, for some models, the interpretability gap between invariant images obtained for robust and standard networks was narrower (Fig. 4c, right). Similar trends were found when SnS was used to uncover the hierarchical invariances of two different subject networks (ResNet18 and VGG16\_bn, again  $L_2$ -robust and standard; Fig. 4e).

We also applied SnS to produce hierarchical invariant images for the readout units of  $L_\infty$ -robust ResNet50. In this case as well, we obtained layer-specific invariant manifolds that were easily separable in pixel space (Fig. S10a and b). However, unlike  $L_2$ -robustification,  $L_\infty$ -robustification yielded invariant images whose interpretability by  $L_\infty$ -robust observer networks remained consistently high or even increased with stretching depth (Fig. S10c). This indicates that  $L_\infty$  adversarial training yields invariance manifolds that are, overall, more interpretable than those produced by  $L_2$  robustification.

To demonstrate the flexibility of the SnS framework, we performed the same interpretability experiment on invariant images evolved for readout units of a vision transformer (ViT). Similar to CNNs, the nature and interpretability of ViT invariances depended on the hierarchical level at which the representation was stretched. Invariances generated from pixel space were substantially less interpretable than those from mid- and high-level representations, which were remarkably similar and more interpretable. This observation aligns with the view that ViTs learn less strictly hierarchical and more globally-integrated features than CNNs (Dosovitskiy et al. (2020); see Fig. S11 in Section I of the Supplementary Material). Taken together, the results of Fig. 4 and Fig. S11 indicate that the level of transferability of the invariances obtained for the readout units of a generator network depends on three key factors: 1) the architecture of the generator (i.e., CNNs vs. vision transformers); 2) its training procedure (robustified vs. standard, in the case of CNNs); and 3) the hierarchical stage at which the representation is stretched to produce the invariant images.

## 5 DISCUSSION

We introduced Stretch-and-Squeeze, a novel, model-agnostic framework that reconceptualizes how we probe invariance of visual representations. Instead of testing pre-defined transformations (e.g., affine) or imposing strict representational identity (as with metamers (Feather et al., 2023)), SnS discovers the actual manifold of transformations a unit tolerates by maximizing stimulus dissimilarity in a chosen representational space while preserving the target unit’s activation. This exploration of functionally equivalent yet representationally distinct inputs allows SnS to map a broader, potentially more “ecologically” relevant stretch of a unit’s response manifold and its boundaries. Compared to other approaches aimed at uncovering invariant manifolds (Cadena et al., 2018; Baroni et al., 2023), SnS is distinguished by its gradient-free and model-agnostic formulation, a critical advantage for applications in neuroscience, especially in cases where reliable digital twin models are unavailable. Moreover, differently from these previous approaches, which have been applied to explore only the low-dimensional invariance fields of visual neurons in early processing stages (i.e., primary visual cortex), SnS operates in a highly expressive search space (i.e., the latent code space with dimension  $d = 4096$ ), which imposes only a soft constraint on the dimensionality of the invariant manifolds. This allows sufficient expressivity to effectively map the complex invariant fields of high-order units, such as CNNs’ readout neurons, thus characterizing the invariances of the entire model with respect to the object categories it has learned to classify.

The hierarchical application of SnS shows that CNN units in readout layers are not necessarily highly robust to the standard transformations (scaling, translation, etc) often applied to probe invariance (DiCarlo et al., 2012; Goodfellow et al., 2009; Kheradpisheh et al., 2016; Engstrom et al., 2019c) (Fig. 2). Rather, they strongly tolerate image changes along radically different axes of abstraction (Fig. 3): from low-level properties (e.g., luminance) to mid-level features (e.g., texture) and, finally, to high-level semantic variations (e.g., object pose). This illustrates how the invariance is built hierarchically, and how this process is intertwined with the process of building feature selectivity in the first place: while sensitivity to increasingly complex combinations of visual features emerges across CNN layers, from the standpoint of a readout unit, it is the *insensitivity* to those same, increasingly complex visual properties that is progressively gained. In other words, while a visual

system strives to gain selectivity for increasingly complex visual features, it must also become invariant to those same features’ combinations.

Our comparison of the invariant images generated for standard and robustly-trained ResNet50 extends prior work on the topic (Feather et al., 2023) with an innovative characterization. The superior semantic coherence and cross-system recognizability (by humans and other CNNs) of the invariances obtained for robust networks strongly corroborates that adversarial training sculpts representations towards human-aligned perceptual features (Engstrom et al., 2019b; Gaziv et al., 2023; Feather et al., 2023). However, unlike computation of MEIs or metamers (Feather et al., 2023), SnS reveals additional trends that are consistent across all human and CNN observers: there are opposite trends for the interpretability of the invariant images generated for the  $L_2$ -robust and the standard networks as a function of stretching depth (Fig. 4c, e). Specifically, we show that stretching low-level representations generates human-interpretable invariances in robustly trained networks, whereas it produces poorly interpretable invariances in standard networks. Conversely, stretching high-level representations enhances interpretability in standard networks but decreases it in  $L_2$ -robust networks. These trends are different from those obtained by testing the interpretability of metamers (see Table S3). This is not surprising, given that metamers are invariant images that minimize the distance from a target natural image in a given layer, rather than maximizing it, as SnS does. Thus, while SnS pushes the search for invariant images to the boundaries of the invariance manifold, metamers explore a much narrower neighborhood of the target image (see Table S4). Hence the complementary nature of the two approaches.

Finally, the model-agnostic and gradientless nature of SnS positions it as a powerful new tool for neuroscience, particularly for systems where high-fidelity “digital twins” may be hard to develop. Our demonstration of SnS’s efficacy with simulated sparse neural recordings directly addresses a critical experimental constraint, paving the way for *in vivo* applications. This could lead to the discovery of new, hierarchical invariances in biological visual systems, extending our understanding of visual object representations in both primates (DiCarlo et al., 2012; Ponce et al., 2019) and other species (Muratore et al., 2025; 2022b; Tafazoli et al., 2017; Tong et al., 2023; Soto & Wasserman, 2016; Wood & Wood, 2015; Schluessel et al., 2012).

**Limitations and broader impacts.** While SnS has proven effective in uncovering novel, meaningful invariances, it presents opportunities for refinement and broader application. One avenue is to use SnS for further, more extensive geometric characterizations of the invariance manifolds, expanding the work presented in Supplementary Material, Section F. Another interesting direction is to apply SnS to characterize invariance in other modalities, such as audition, using a generator capable of producing audio waveforms or spectrograms from a low-dimensional latent code - e.g., SpecGAN, WaveGAN (Donahue et al., 2018). Returning to the visual domain, our work has demonstrated SnS’ generative capabilities using eight different neural models (Fig. 4 and S11), belonging to three different architectural families, with parameter counts spanning roughly an order of magnitude (from 11.7M for ResNet-18 to 86.6M for ViT). This provides a strong test of our framework’s generalizability, although future work could test SnS with even larger models (e.g., DINO (Caron et al., 2021), ViT22-B (Dehghani et al., 2023)). Finally, by pinpointing where the perceptual misalignement between robustified networks and humans is strongest (i.e., in higher-order invariances), SnS could guide the development of models that achieve more human-aligned invariance. For instance, one could design a visual diet where “good” (i.e., human-like) and “bad” (i.e., nonhuman-like) invariant images are explicitly used in the training of the network.

In terms of application to neuroscience studies, while SnS’ robustness to significant representational subsampling provides a strong foundation for its neuroscientific applicability, the next step is direct validation in biological systems. Future research should focus on adapting SnS for *in vivo* experiments, addressing issues such as neural signal variability, the limited duration of neuronal recordings, and the need to optimize stimulus presentation in closed loop experiments, similarly to what was already achieved with XDREAM (Ponce et al., 2019). One challenge will be to leverage the wealth of neuronal data afforded by high-density silicon probes (Siegle et al., 2021), that can simultaneously record from hundreds to thousands of units. One strategy would be to apply SnS in parallel to multiple recorded neurons with non-overlapping receptive fields (RFs), thus independently evolving the distinct portions of the image processed by each unit. Conversely, for neurons with overlapping RFs, one could first screen them using natural images (Ponce et al., 2019) to search for subpopulations of units with similar selectivity. SnS could then be applied to the subpopulation to find the collective invariance field of the neuronal ensemble.

## 6 REPRODUCIBILITY STATEMENT

To enhance the reproducibility of our research, we provide the following resources. The SnS algorithm is described in detail within the main paper’s Methods Section 3.1, with its corresponding pseudocode available in the Supplementary Material (see Algorithm 1). Furthermore, the complete source code used to conduct the experiments is available at <https://github.com/zoccolan-lab/SnS>, while experimental data were open sourced at Tausani et al. (2025). For clarity on our experimental setup, a general description of the experimental parameters can be found in the methods section. For a more granular understanding, further information on methodological details, hyperparameter settings and the computational resources used is reported in Supplementary Material, Section A. The Supplementary Material also includes a comprehensive section on the statistical analysis employed in our experiments (see Section B.1).

## 7 ETHICS STATEMENT

The human behavioral study was conducted under a pre-existing, approved Institutional Review Board (IRB) protocol. The task, which involved participants viewing images on a computer screen and making responses by clicking buttons, presented no more than minimal risk. Further details, including a complete reproduction of test instructions and information on participant compensation, are provided in Supplementary Material Section B.

## 8 ACKNOWLEDGEMENTS

We would like to thank Sebastiano Quintavalle for his help in implementing the code, and Eugenio Piasini for his suggestions on the Pareto front optimization. This work was funded by the European Union – NextGenerationEU – NRRM4C2-I.1.1, in the framework of the PRIN Project no. 2022WX3FM5, CUP:G53D23003220006 (DZ), by NIGMS Award T32GM144273 (MT), by NIH grant R01EY026025 (GK), and by the NSF Center for Brains, Minds, and Machines NSF CCF-1231216 (GK). The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of any of the above organizations.

## REFERENCES

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Luca Baroni, Mohammad Bashiri, Konstantin F Willeke, Ján Antolík, and Fabian H Sinz. Learning invariance manifolds of visual sensory neurons. In *NeurIPS workshop on symmetry and geometry in neural representations*, pp. 301–326. PMLR, 2023.

David GT Barrett, Ari S Morcos, and Jakob H Macke. Analyzing biological and artificial neural networks: challenges with opportunities for synergy? *Current opinion in neurobiology*, 55:55–64, 2019.

Mohammad Bashiri, Luca Baroni, Ján Antolík, and Fabian H. Sinz. Learning and aligning single-neuron invariance manifolds in visual cortex. In *International Conference on Learning Representations (ICLR)*, 2025.

Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, 2019.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

Alexander Berardino, Valero Laparra, Johannes Ballé, and Eero Simoncelli. Eigen-distortions of hierarchical representations. *Advances in neural information processing systems*, 30, 2017.

Santiago A. Cadena, Marissa A. Weis, Leon A. Gatys, Matthias Bethge, and Alexander S. Ecker. Diverse feature visualizations reveal invariances in early layers of deep neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. URL <https://arxiv.org/abs/2104.14294>.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David Cox, and James J DiCarlo. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *Advances in Neural Information Processing Systems*, 33:13073–13087, 2020.

Joshua R De Leeuw. jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, 47:1–12, 2015.

Kalyanmoy Deb. *Multi-objective Optimisation Using Evolutionary Algorithms: An Introduction*, pp. 3–34. Springer London, 2011. ISBN 9780857296528. doi: 10.1007/978-0-85729-652-8\_1. URL [http://dx.doi.org/10.1007/978-0-85729-652-8\\_1](http://dx.doi.org/10.1007/978-0-85729-652-8_1).

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin F. Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Patrick Collier, Alexey Gritsenko, Vighnesh Birodkar, Cristina Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetić, Dustin Tran, Thomas Kipf, Mario Lučić, Xiaohua Zhai, Daniel Keysers, Jeremiah Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters, 2023. URL <https://arxiv.org/abs/2302.05442>.

Misha Denil, Babak Shakibi, Laurent Dinh, Marc’Aurelio Ranzato, and Nando De Freitas. Predicting parameters in deep learning. *Advances in neural information processing systems*, 26, 2013.

James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, February 2012. ISSN 0896-6273. doi: 10.1016/j.neuron.2012.01.010. URL <http://dx.doi.org/10.1016/j.neuron.2012.01.010>.

Zhiwei Ding, Dat T Tran, Kayla Ponder, Erick Cobos, Zhuokun Ding, Paul G Fahey, Eric Wang, Taliah Muhammad, Jiakun Fu, Santiago A Cadena, et al. Bipartite invariance in mouse primary visual cortex. *bioRxiv*, 2023.

Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*, 2018.

Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems*, 29, 2016.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019a. URL <https://github.com/MadryLab/robustness>.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations, 2019b. URL <https://arxiv.org/abs/1906.00945>.

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International conference on machine learning*, pp. 1802–1811. PMLR, 2019c.

Elena Facco, Maria d’Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140, 2017.

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

Thomas Fel, Ivan F Rodriguez Rodriguez, Drew Linsley, and Thomas Serre. Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in neural information processing systems*, 35:9432–9446, 2022.

Thomas Fel, Thibaut Boissin, Victor Boutin, Agustin Picard, Paul Novello, Julien Colin, Drew Linsley, Tom Rousseau, Rémi Cadène, Lore Goetschalckx, et al. Unlocking feature visualization for deep network with magnitude constrained optimization. *Advances in Neural Information Processing Systems*, 36:37813–37826, 2023.

Félix-Antoine Fortin, François-Michel De Rainville, Marc-André Gardner, Marc Parizeau, and Christian Gagné. DEAP: Evolutionary algorithms made easy. *Journal of Machine Learning Research*, 13:2171–2175, jul 2012.

Guy Gaziv, Michael Lee, and James J DiCarlo. Strong and precise modulation of human percepts via robustified anns. *Advances in Neural Information Processing Systems*, 36:65936–65947, 2023.

Tal Golan, Prashant C Raju, and Nikolaus Kriegeskorte. Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, 117(47):29330–29337, 2020.

Ian Goodfellow, Honglak Lee, Quoc Le, Andrew Saxe, and Andrew Ng. Measuring invariances in deep networks. *Advances in neural information processing systems*, 22, 2009.

Chong Guo, Michael Lee, Guillaume Leclerc, Joel Dapello, Yug Rao, Aleksander Madry, and James DiCarlo. Adversarially trained neural representations are already as robust as biological neural representations. In *International Conference on Machine Learning*, pp. 8072–8081. PMLR, 2022.

Nikolaus Hansen, Sibylle D Müller, and Petros Koumoutsakos. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (cma-es). *Evolutionary computation*, 11(1):1–18, 2003.

Nikolaus Hansen, Youhei Akimoto, and Petr Baudis. CMA-ES/pycma on Github. Zenodo, DOI:10.5281/zenodo.2559634, February 2019. URL <https://doi.org/10.5281/zenodo.2559634>.

Olivier J Hénaff and Eero P Simoncelli. Geodesics of learned representations. *arXiv preprint arXiv:1511.06394*, 2015.

Jörn-Henrik Jacobsen, Jens Behrmann, Richard Zemel, and Matthias Bethge. Excessive invariance causes adversarial vulnerability, 2020. URL <https://arxiv.org/abs/1811.00401>.

James J Jun, Nicholas A Steinmetz, Joshua H Siegle, Daniel J Denman, Marius Bauza, Brian Barbarits, Albert K Lee, Costas A Anastassiou, Alexandru Andrei, Çağatay Aydin, et al. Fully integrated silicon probes for high-density recording of neural activity. *Nature*, 551(7679):232–236, 2017.

Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

Saeed Reza Kheradpisheh, Masoud Ghodrati, Mohammad Ganjtabesh, and Timothée Masquelier. Deep networks can resemble human feed-forward vision in invariant object recognition. *Scientific reports*, 6(1):32672, 2016.

Daiichiro Kuroki. A new jspysch plugin for psychophysics, providing accurate display duration and stimulus onset asynchrony. *Behavior Research Methods*, 53:301–310, 2021.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5188–5196. IEEE, June 2015. doi: 10.1109/cvpr.2015.7299155. URL <http://dx.doi.org/10.1109/CVPR.2015.7299155>.

Girik Malik, Dakarai Crowder, and Ennio Mingolla. Extreme image transformations affect humans and machines differently. *Biological Cybernetics*, 117(4):331–343, 2023.

Giulio Matteucci, Eugenio Piasini, and Davide Zoccolan. Unsupervised learning of mid-level visual representations. *Current Opinion in Neurobiology*, 84:102834, 2024.

Paolo Muratore, Sina Tafazoli, Eugenio Piasini, Alessandro Laio, and Davide Zoccolan. Prune and distill: similar reformatting of image information along rat visual cortex and deep neural networks. *Advances in Neural Information Processing Systems*, 35:30206–30218, 2022a.

Paolo Muratore, Sina Tafazoli, Eugenio Piasini, Alessandro Laio, and Davide Zoccolan. Prune and distill: similar reformatting of image information along rat visual cortex and deep neural networks, 2022b. URL <https://arxiv.org/abs/2205.13816>.

Paolo Muratore, Alireza Alemi, and Davide Zoccolan. Unraveling the complexity of rat object vision requires a full convolutional network and beyond. *Patterns*, 6(2):101149, February 2025. ISSN 2666-3899. doi: 10.1016/j.patter.2024.101149. URL <http://dx.doi.org/10.1016/j.patter.2024.101149>.

Aran Nayebi, Nathan CL Kong, Chengxu Zhuang, Justin L Gardner, Anthony M Norcia, and Daniel LK Yamins. Mouse visual cortex as a limited resource system that self-learns an ecologically-general representation. *PLOS Computational Biology*, 19(10):e1011506, 2023.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *arXiv preprint arXiv:1602.03616*, 2016.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Understanding neural networks via feature visualization: A survey. *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 55–76, 2019.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. <https://distill.pub/2017/feature-visualization>.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL [https://proceedings.neurips.cc/paper\\_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf).

Carlos R Ponce, Will Xiao, Peter F Schade, Till S Hartmann, Gabriel Kreiman, and Margaret S Livingstone. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, 177(4):999–1009, 2019.

Alexandre Pouget, Peter Dayan, and Richard Zemel. Information processing with population codes. *Nature Reviews Neuroscience*, 1(2):125–132, 2000.

Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better?, 2020. URL <https://arxiv.org/abs/2007.08489>.

V. Schluessel, G. Fricke, and H. Bleckmann. Visual discrimination and object categorization in the cichlid *pseudotropheus* sp. *Animal Cognition*, 15(4):525–537, March 2012. ISSN 1435-9456. doi: 10.1007/s10071-012-0480-3. URL <http://dx.doi.org/10.1007/s10071-012-0480-3>.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, pp. 407007, 2018.

Odelia Schwartz, Jonathan W Pillow, Nicole C Rust, and Eero P Simoncelli. Spike-triggered neural characterization. *Journal of vision*, 6(4):13–13, 2006.

Tatyana O Sharpee. Computational identification of receptive fields. *Annual review of neuroscience*, 36(1):103–120, 2013.

Joshua H Siegle, Xiaoxuan Jia, Séverine Durand, Sam Gale, Corbett Bennett, Nile Graddis, Greggory Heller, Tamina K Ramirez, Hannah Choi, Jennifer A Luviano, et al. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, 592(7852):86–92, 2021.

Nicholas James Sofroniew, Daniel Flickinger, Jonathan King, and Karel Svoboda. A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. *elife*, 5:e14472, 2016.

Fabian A. Soto and Edward A. Wasserman. Promoting rotational-invariance in object recognition despite experience with only a single view. *Behavioural Processes*, 123:107–113, February 2016. ISSN 0376-6357. doi: 10.1016/j.beproc.2015.11.005. URL <http://dx.doi.org/10.1016/j.beproc.2015.11.005>.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D Harris. Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437):eaav7893, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Sina Tafazoli, Houman Safaai, Gioia De Franceschi, Federica Bianca Rosselli, Walter Vanzella, Margherita Riggi, Federica Buffolo, Stefano Panzeri, and Davide Zoccolan. Emergence of transformation-tolerant representations of visual objects in rat lateral extrastriate cortex. *eLife*, 6, April 2017. ISSN 2050-084X. doi: 10.7554/elife.22794. URL <http://dx.doi.org/10.7554/elife.22794>.

Morgan B. Talbot, Gabriel Kreiman, James J. DiCarlo, and Guy Gaziv. L-wise: Boosting human visual category learning through model-based image selection and enhancement. *International Conference on Learning Representations (ICLR)*, 2025.

Lorenzo Tausani, Paolo Muratore, Morgan B. Talbot, Giacomo Amerio, Gabriel Kreiman, and Davide Zoccolan. Stretching beyond the obvious: A gradient-free framework to unveil the hidden landscape of visual invariance, September 2025. URL <https://doi.org/10.5281/zenodo.17191765>.

Rudi Tong, Ronan da Silva, Dongyan Lin, Arna Ghosh, James Wilsenach, Erica Cianfarano, Pouya Bashivan, Blake Richards, and Stuart Trenholm. The feature landscape of visual cortex. November 2023. doi: 10.1101/2023.11.03.565500. URL <http://dx.doi.org/10.1101/2023.11.03.565500>.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Shimon Ullman, Liav Assif, Ethan Fetaya, and Daniel Harari. Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences*, 113(10):2744–2749, 2016.

Edgar Y Walker, Fabian H Sinz, Erick Cobos, Taliah Muhammad, Emmanouil Froudarakis, Paul G Fahey, Alexander S Ecker, Jacob Reimer, Xaq Pitkow, and Andreas S Tolias. Inception loops discover what excites neurons most using deep predictive models. *Nature neuroscience*, 22(12):2060–2065, 2019.

Bin Xu Wang and Carlos R Ponce. High-performance evolutionary algorithms for online neuron control. In *Proceedings of the genetic and evolutionary computation conference*, pp. 1308–1316, 2022a.

Bin Xu Wang and Carlos R Ponce. Tuning landscapes of the ventral stream. *Cell Reports*, 41(6), 2022b.

Bin Xu Wang and Carlos R Ponce. Neural dynamics of object manifold alignment in the ventral stream. *bioRxiv*, pp. 2024–06, 2024.

Zhou Wang and Eero P Simoncelli. Maximum differentiation (mad) competition: A methodology for comparing computational models of perceptual quantities. *Journal of vision*, 8(12):8–8, 2008.

Samantha M. W. Wood and Justin N. Wood. A chicken model for studying the emergence of invariant object recognition. *Frontiers in Neural Circuits*, 9, February 2015. ISSN 1662-5110. doi: 10.3389/fncir.2015.00007. URL <http://dx.doi.org/10.3389/fncir.2015.00007>.

Will Xiao and Gabriel Kreiman. Gradient-free activation maximization for identifying effective stimuli. *arXiv preprint arXiv:1905.00378*, 2019.

Will Xiao and Gabriel Kreiman. Xdream: Finding preferred stimuli for visual neurons using generative networks and gradient-free optimization. *PLoS computational biology*, 16(6):e1007973, 2020.

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

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.

Ehud Zohary, Michael N Shadlen, and William T Newsome. Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature*, 370(6485):140–143, 1994.

# Supplementary Material

## A ADDITIONAL DETAILS REGARDING COMPUTATIONAL EXPERIMENTS

### A.1 PSEUDOCODE FOR THE SNS ALGORITHM

Here we present the pseudo-code (Algorithm 1) for the Sns algorithm. We used library implementations for the CMA-ES gradient-free optimizer (Hansen et al., 2019) and the Pareto front computation (Fortin et al., 2012). Importantly, we introduce the following `should_stop` criteria for our experiments.

Optimizations were terminated by either reaching a maximum iteration limit (500) or satisfying an early stopping rule derived from the target unit’s response statistics to a large dataset of natural images (i.e., the full ImageNet train set,  $n \approx 1.2$  million). In particular, given the vastness of the dataset, we assumed that the activation elicited by the most effective image in the natural dataset ( $a_{\max\_nat}^\ell$ ) represented a conservative threshold for an image to be considered *invariant*. On the other hand, the activation from the least effective image ( $a_{\min\_nat}^\ell$ ) represents a non-response level, functioning as a threshold for an image to be considered *adversarial*. Therefore, when searching for adversarial images, termination was reached when a large fraction (i.e.,  $\geq 90\%$ ) of the current population of images elicited activations at or below  $a_{\min\_nat}^\ell$ , thus leading to a non-responsive status. Conversely, when looking for invariant images after initialization from random noise, it was terminated when a large fraction (i.e.,  $\geq 90\%$ ) of the population of images yielded activations equal or above  $a_{\max\_nat}^\ell$ , thus reaching the desired high-activation regime.

In terms of computational complexity, both generator-based image synthesis from latent codes and the extraction of activations from the target network require only a single feedforward pass per iteration, whose computational cost, while depending on network size, is negligible in practice for all the architectures examined in our study, which span roughly an order of magnitude of parameter counts (from 11.7M for ResNet-18 to 86.6M for ViT). Consequently, the dominant computational bottleneck is the CMA-ES optimizer, whose running time scales quadratically with the dimensionality  $d$  of the search space. In practice, however, this does not pose a limitation: CMA-ES has been shown to be a highly effective method for gradient-free optimization of neural activity *in vivo* (Wang & Ponce, 2022a), and it remains computationally efficient under our hardware configuration as well (see Section A.6).

---

**Algorithm 1** Sns gradient-free optimization algorithm

---

```

Require:  $\phi$                                  $\triangleright$  Target network
Require:  $\psi$                                  $\triangleright$  Image generator
Require:  $T \geq 0$                              $\triangleright$  Maximum number of iterations
Require:  $\kappa, \ell \in \{0, 1, \dots, \text{depth}(\phi)\}$ 
Require:  $\mathbf{x}_{\text{ref}} \in \mathbb{R}^{C \times H \times W}$            $\triangleright$  Common choice is  $\mathbf{x}_{\text{ref}} = \mathbf{x}_\star$ 

 $t \leftarrow 0$ 
 $\mathbf{a}_{\text{ref}}^{\kappa, \ell} \leftarrow \Gamma(\mathbf{x}_{\text{ref}}, \phi^{\kappa, \ell})$ 
 $\xi_0 \leftarrow \mathcal{N}(0, \sigma_{\text{init}})$ 
while  $t \leq T$  and !early_stop do
     $\mathbf{x}_t = \psi(\xi_t)$                                  $\triangleright$  Compute new candidate stimuli
     $\mathbf{a}^{\kappa, \ell} = \Gamma(\mathbf{x}_t, \phi^{\kappa, \ell})$            $\triangleright$  Measure activations in target layers  $\kappa, \ell$ 
     $\mathcal{L}_{\text{stretch}}^\kappa \leftarrow -\|\mathbf{a}^\kappa - \mathbf{a}_{\text{ref}}^\kappa\|$ 
     $\mathcal{L}_{\text{squeeze}}^\ell \leftarrow +\|\mathbf{a}^\ell - \mathbf{a}_{\text{ref}}^\ell\|$ 
     $\mathcal{P} \leftarrow \text{compute\_pareto\_front}(\xi_t, \mathcal{L}_{\text{stretch}}^\kappa, \mathcal{L}_{\text{squeeze}}^\ell)$ 
     $\xi_{t+1} \leftarrow \text{evolve}(\mathcal{P}, \xi_t)$            $\triangleright$  Evolution strategy CMA-ES optimization
     $t \leftarrow t + 1$ 
     $\text{early\_stop} \leftarrow \text{should\_stop}(\mathbf{a}^\kappa, \mathbf{a}^\ell)$ 
end while

```

---

## A.2 EXPERIMENTAL HYPERPARAMETERS

The generative model  $\psi$  is instantiated as a pretrained deep neural network variant, specifically the `fc7` configuration from Dosovitskiy & Brox (2016). The choice of this model instead of alternative models, such as as large-scale GANs and diffusion models, was motivated by the fact that the latter impose very strong priors on the structure of the synthesized images, which risks evolving exceedingly complex photorealistic details that are unlikely to be encoded by units in low- and mid-level processing stages (Wang & Ponce, 2024).

The CMA-ES optimizer is configured with the following hyperparameters:

- **Initial Step Size** ( $\sigma_0$ ): 1.0
- **Population Size**: 50

The initial step size  $\sigma_0$  determines the covariance of the sampling distribution at the onset of optimization, effectively controlling the exploration radius in the latent space.

The population size specifies the number of candidate solutions (i.e. codes) evaluated per iteration.

Initial codes were randomly sampled from a Gaussian distribution  $\mathcal{N}(0, \sigma_{\text{init}})$ . The initial standard deviation,  $\sigma_{\text{init}}$ , was set to  $\sqrt{0.01 \times \mathbb{E}[\|\zeta_{\text{ref}}\|]}$ , where  $\mathbb{E}[\|\zeta_{\text{ref}}\|]$  is the mean absolute value of the reference code. For invariance experiments benchmarked against natural images, where direct reference codes were unavailable, initial codes were drawn from a standard normal distribution ( $\mathcal{N}(0, 1)$ ).

## A.3 PARETO FRONT COMPUTATION

Multi-objective optimization often involves conflicting goals (e.g., stretching vs. squeezing, see Fig. 1). One way of dealing with such problems is by structuring the fitness function as a linear combination of the different optimization objectives (weighted-sum methods). However, these methods often require fine parameter tuning to prevent one objective from dominating. Unlike weighted-sum methods, Pareto fronts (PFs) handle trade-offs without parameter tuning. At each optimization iteration, candidate solutions ( $n=50$ ) are ranked iteratively into PFs in the two-objective space. Each PF contains mutually non-dominant solutions (i.e. one is not better than another in both objectives). To get a full ordering, solutions are sorted within each front:

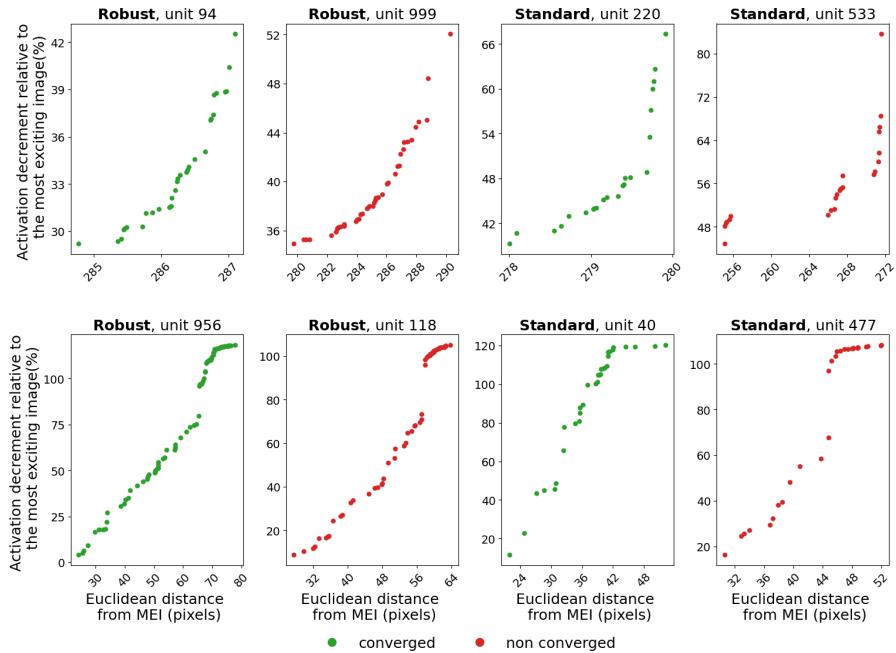
**Invariance Experiments:** In this scenario, solutions on the same Pareto front were prioritized based on their proximity to the target unit’s reference activation level ( $\min_x \|\Gamma(\mathbf{x}, \phi^\ell) - \mathbf{x}_{\text{ref}}^\ell\|_2$ ). This exploitative strategy was motivated by the empirical difficulty of converging towards the target activation level while at the same time maximizing input distance.

**Adversarial Attacks:** Solutions on the same Pareto front were selected for the next generation with uniform probability (random ordering). This promotes *exploration* along the front, preventing premature convergence to a single type of adversarial strategy.

CMA-ES then updates its distribution using the best 30% of this ranking. At the end of the optimization, all solutions are used to compute a global PF, which is thus fully data-driven (see Fig. S1 for several examples). In our study, we analyzed only the last solution from this final PF for each experiment.

## A.4 SEPARABILITY OF THE INVARIANT IMAGES IN THE PIXEL REPRESENTATION

To assess whether the representation spaces where the stretching was applied yielded sets of invariant images that were distinguishable at the pixel level, we performed an image classification analysis. We generated invariant images for  $n = 77$  readout layer units (i.e., 77 ImageNet classes) by stretching the representations in three different processing stages: (i) `low_level` (pixel space), (ii) `mid_level`, and (iii) `high_level`. For each unit and each representation space condition, we performed 10 independent optimization runs with different random seeds, yielding a total dataset of  $77 \times 3 \times 10 = 2310$  images. Principal Component Analysis (PCA) was applied to the raw pixel representations of all images. The first  $k$  principal components were then used as input features to train a Support Vector Classifier with a radial basis function kernel to discriminate the class to which the invariant images belonged (i.e., `low_level`, `mid_level`, or `high_level`). 5-fold cross-validation was used.

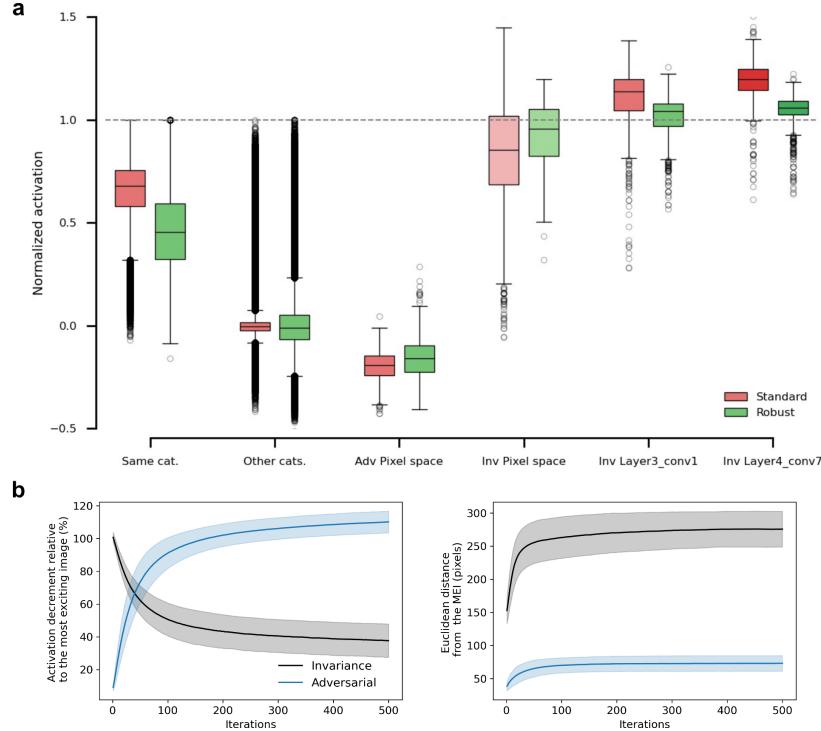


**Figure S1: Examples of solutions in the final Pareto Front.** Visualization of the final Pareto fronts for  $n = 4$  example units in Invariance experiments (top) and  $n = 4$  example units in Adversarial Attack experiments (bottom). In all reported experiments, targets were located in the readout layer, while pixel space was used as lower representation space. The ordinate axis represents the activation reduction of a readout unit of a robust or standard ResNet50 with respect to the activation produced by a reference image (MEI), when SnS is applied to synthesize either adversarial or invariant images, while the abscissa represents the  $L_2$  pixel distance between these synthesized images and the reference. In the Invariance Experiments, SnS maintains a substantial fraction of the reference image activation (regardless of whether the optimization fully converges) while allowing the synthesized images to diverge markedly from the reference in pixel space. Conversely, in the Adversarial Attack setting, SnS reliably produces images that almost completely suppress the response of the target unit, yet remaining comparatively close to the MEI in the pixel space representation. Note the different scale range of the abscissa and ordinate axes between the top and bottom plots.

## A.5 OPTIMIZATION CONVERGENCE

### A.5.1 SNS IMAGES FALL INTO FUNCTIONALLY RELEVANT ACTIVATION REGIMES

Although some optimizations terminated at maximum iterations (adversarial: 97.92% (robust), 93.12% (standard); invariant: 63.38%, 35.71%, 21.30% (robust: `low_level`, `mid_level`, `high_level`); 86.75%, 32.60%, 5.84% (standard: `low_level`, `mid_level`, `high_level`)), final populations consistently reached functionally relevant activation regimes. Most invariant images achieved activations within the top 0.1% of natural image responses, being comparable or more extreme than the readout activation of images of the same category encoded by the readout target unit (Fig. S2a). Adversarial examples drove activations to the lowest 1<sup>st</sup> percentile of the natural images, thus significantly suppressing target unit activation with respect to average activation with a natural image. As shown in Table S1, this was true also when considering only those units for which the early-stopping criterion was not reached, i.e., the optimizations that terminated at the maximum number of iterations (500). Moreover, as illustrated by the convergence curves in Fig. S2b, the average trajectories of both optimization objectives (activation of the target readout units and Euclidean distance from the MEI in pixelspace) rapidly approached stable asymptotic values consistent with the respective optimization goals (i.e., invariant and adversarial). In particular, the Euclidean distance plateaued, indicating that the optimized images no longer underwent substantial changes after a certain number of iterations.



**Figure S2: SnS images fall into functionally relevant activation regimes.** (a) Distributions of readout target unit activations (normalized by peak natural stimulus response) for SnS-generated images: invariant (stretched at pixel, mid-, high-levels) and adversarial (pixel-space). Each distribution pulls together all optimizations performed in that specific condition ( $n = 77$  target neurons  $\times 10$  random seeds = 770 experiments per condition). Comparison with natural image activations (same/different category) confirms that SnS invariant images elicit strong target responses, while adversarial examples achieve significant suppression. (b) The average convergence curves ( $\pm$  SD) for the experiments illustrated in Fig. 2 (invariance and adversarial attacks,  $n = 770$  runs per condition). Both optimization objectives (i.e., activation of the target units and Euclidean distance from the MEI in pixel space) are shown as a function of iteration number. Both objectives converged to stable asymptotic values consistent with the intended optimization target (i.e., the search for invariant or adversarial images). In particular, the Euclidean distance reached a plateau, indicating that the images no longer continued to evolve significantly following an initial period of convergence.

#### A.5.2 INVARIANT SOLUTIONS ARE MAXIMALLY DISTINCT FROM THEIR REFERENCES AND HETEROGENEOUS ACROSS MULTIPLE RUNS

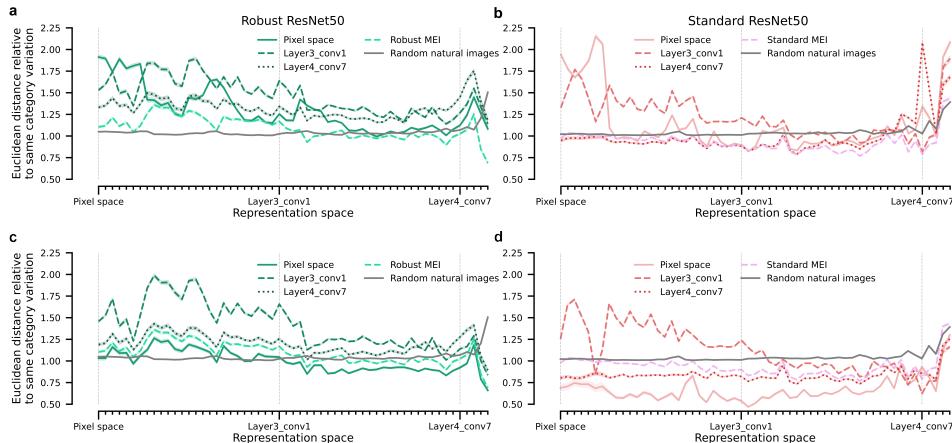
To quantify how much the invariant images generated by SnS (when targeting a given processing stage) diverged from the reference MEI ( $x^*$ ) across the depth of ResNet50, we computed the Euclidean distance between their activation vectors at each layer of the network (*Representational Distance Analysis* in short). Again, 10 different invariant images were produced for  $n = 77$  distinct readout units by stretching representations at layers `low_level`, `mid_level`, and `high_level`. We calculated the mean ( $\pm$  SEM) Euclidean distance between each of these  $77 \times 10$  invariant images and the corresponding reference images at every convolutional, pooling and linear layer of the network, separately for each stretching stage. Analogously, to quantify heterogeneity between different invariant solutions, for each aforementioned stage we measured the average pairwise distance between invariant images, then averaged across each target neuron ( $n = 77$ ). These distances were compared against three control metrics:

**Reference Variability distance:** Mean distance between all pairs drawn from 10 independent XDREAM runs for each unit ( $n = \binom{10}{2} = 45$  pairs per unit,  $n = 77$ )

**Within-Category distance:** Mean distance between pairs of images randomly sampled from the same ImageNet category (mean over 1000 categories, 10 images/category).

**Table S1: Mean activations of target units to SnS-synthesized images from optimizations that failed to converge.** Reported activations are normalized by the response to the maximally activating natural image for the corresponding unit.  $N$  denotes the number of runs in which the early-stopping criterion was not reached for each experimental condition, out of a total of 770 optimization runs per condition. Successful maintenance of activation levels in Invariance runs would be indicated by positive activation values in the Mean column near to or exceeding 1, while successful suppression of activity in Adversarial runs would result in low or negative values.

Experiment	Layer	Model Type	Mean	SEM	N
Invariance	low_level	Standard	0.7682	0.0103	668
Invariance	mid_level	Standard	0.9581	0.0110	251
Invariance	high_level	Standard	0.9623	0.0233	45
Adversarial	low_level	Standard	-0.1920	0.0027	718
Invariance	low_level	Robust $L_2$	0.8467	0.0056	489
Invariance	mid_level	Robust $L_2$	0.9104	0.0059	276
Invariance	high_level	Robust $L_2$	0.9271	0.0075	166
Adversarial	low_level	Robust $L_2$	-0.1547	0.0036	755



**Figure S3: Representation of the invariant images across the feedforward hierarchy.** (a) Average distance between the invariant images generated by SnS and their reference MEIs across the different stages of a robust ResNet50. Euclidean distances were normalized by the mean, within-category distance of Imagenet images. Different lines indicate the different stretching layers used in the optimization (also reported in the  $x$  axis). The average distance between randomly selected natural images from Imagenet (solid gray line) and the average distance between multiple MEIs generated via XDREAM are reported for comparison. (b) Same as (a) for standardly trained network. (c) Normalized average distance between the invariant images generated by SnS across the different stages of a robust ResNet50. Different lines indicate the different stretching layers used in the optimization. The average distance between randomly selected natural images from Imagenet (solid gray line) and the average distance between multiple MEIs generated via XDREAM are reported for comparison. (d) Same as (c) for standardly trained network.

**Between-Category distance:** Mean distance between pairs of images randomly sampled across all ImageNet categories (using the same total number of pairs as the within-category control).

In our analysis (see Fig. S3) we normalized all Euclidean distances by the mean, within-category distance of Imagenet images. Fig. S3a-b show that invariant images that were generated to be maximally different from the MEI in a specific layer had the largest normalized distance in that layer, as compared to invariant images obtained by applying the stretching in other layers. Moreover, this distance consistently surpassed the one among natural images (both within the same category and randomly sampled) and multiple MEIs of the same unit, demonstrating SnS’s efficacy in exploring larger spans of a unit’s invariance manifold, as compared to searching multiple times for different

MEIs. Moreover, Fig. S3c-d show that distances between invariant images were significantly above 0 for all stretching layers (pixel space, mid, and late layers) in both standard and robust ResNet50, indicating a substantial diversity between invariant images. Heterogeneity is lowest in the pixel space (standard: 0.69, robust: 1.03) and higher in the mid (standard: 1.24, robust: 1.66) and late layers (standard: 0.85, robust: 1.27). Interestingly, invariances in the mid-layers of both networks - and those in the late layer of the robust network - are more heterogeneous than both random ImageNet pairs and multiple MEIs for the same unit. Instead, the invariant images in the late layer of the standard network exceed MEI heterogeneity, but not that of random ImageNet pairs. This heterogeneity is supported by SnS random initialization, which promotes broad exploration of the invariance landscape across multiple runs for the same unit and reduces the likelihood of converging to the same local minimum.

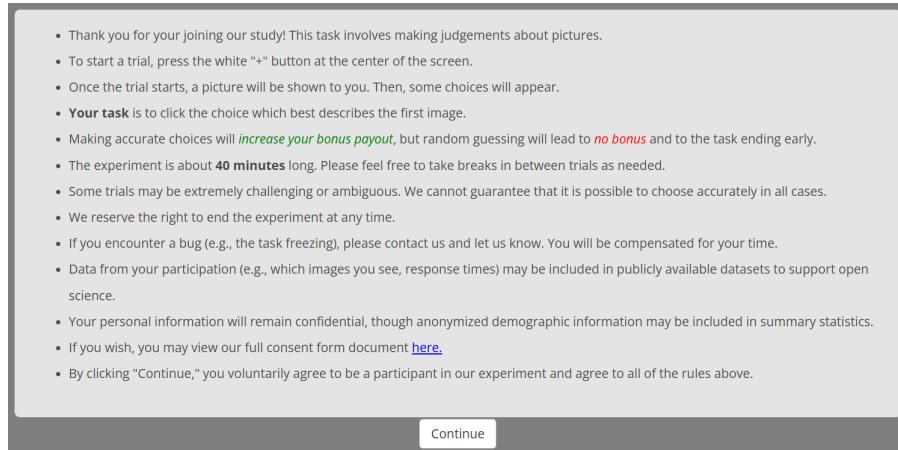
#### A.6 COMPUTATIONAL RESOURCES

All experiments reported in this work were executed on a Dell Precision 7960 Tower (Ubuntu 22.04.5 LTS) with the following configuration:

- **CPU:** Intel (R) Xeon(R) w5-3425 (24 cores, 48 threads);
- **GPU:** NVIDIA RTX A6000 (48 GB GDDR6);
- **System Memory:** 128 GiB;
- **Storage:** 8.2 TB.

On this setup, a full optimization of a single SnS run (500 iterations) lasted approximately 120 seconds. The complete source code used to conduct the experiments is available at <https://github.com/zoccolan-lab/SnS>. Moreover the experimental data is accessible via Tausani et al. (2025).

#### B PROCEDURE TO PROBE THE INTERPRETABILITY OF THE INVARIANT IMAGES BY HUMANS AND OBSERVER NETWORKS



Continue

Figure S4: **Instructions shown to human participants.** These instructions were shown to the participants at the beginning of the task.

For our interpretability experiments, we chose 12 ImageNet categories (goldfish, chickadee, frilled lizard, Dungeness crab, Staffordshire bull terrier, cicada, candle, grand piano, minibus, reflex camera, soccer ball, cup). For each category, we selected 10 reference images  $x_{\text{ref}}$  and, for each of them, we generated one invariant image by stretching its representation in `low_level`, `mid_level`, and `high_level` in both standard and robust ResNet50. This yielded 6 kinds of invariant images per reference image, where invariance was always defined relative to the readout unit corresponding to the reference image's category. For each category, we also included in the classification pool the 10 reference images and 20 MEIs (generated with XDREAM) of the readout unit corresponding to the chosen category (10 for the standard and 10

for the robust network). This yielded a total of  $12 \text{ categories} \times 10 \text{ randomly sampled images} \times 9 \text{ stimulus types}$  (i.e., 1 reference image, 2 MEIs and 6 invariant images). Observer networks had to classify all 1080 of these images; humans had to classify a subset of 540 images (i.e., 5 rather than 10 randomly sampled images per category). Classification accuracy (12-AFC percent correct) was calculated for each of the 9 stimulus types. To assess the generalizability of our findings, we repeated the same experiment with invariant images generated by two additional architectures (ResNet18 and VGG16\_bn, in both standard and robust versions; see details above). The pool of observer networks was identical to the previous experiment. As for ResNet50, 3 levels of stretching were applied: Pixel space, mid level (VGG16\_bn: 2<sup>nd</sup> convolutional layer in the 3<sup>rd</sup> block; ResNet-18: 4<sup>th</sup> convolutional stage in layer 2) and high level (VGG16\_bn: 3<sup>rd</sup> convolutional layer in the 5<sup>th</sup> block; ResNet-18: 4<sup>th</sup> convolutional stage in layer 4).

The psychophysics experiment administered to human participants was conducted as follows. After clicking on a fixation cross at the start of each trial, participants were presented with an image for 50 ms at  $\sim 10^\circ$  of visual angle. After a 20ms blank screen, a backward mask image (random RGB pixel noise) was then presented for 200ms. Images were presented in random order. Participants then clicked one of 12 category buttons that subsequently appeared in a circular arrangement, with a 10-second timeout (Fig. 4a). To avoid directional bias, the sequence of category buttons was randomly rotated for each trial.

Participants for the human experiment were recruited using the online Prolific platform under a preexisting IRB protocol. Participants were compensated at an overall rate calibrated to \$15 USD per hour. To facilitate engagement with the task, compensation included a 1-cent bonus for each correct trial, and participants were shown a green check after correct responses or a black “X” otherwise. The typical bonus amount that participants would receive was overestimated based on pilot data. To adjust for this discrepancy, after recruitment for the study was complete, participants were uniformly provided with an additional bonus to bring compensation to the \$15/hour level.

The main task was preceded by a screening phase, in which participants classified 24 natural images (2 per category), and were allowed to proceed if they correctly classified  $\geq 1$  image per category and  $\geq 16$  in total (maximum 3 attempts allowed per participant). Data from the screening phase were not included in any analyses. Participants who did not pass the screening phase were compensated for the time spent during the screening process.

The experiment posed minimal risks to participants. It is unlikely but conceivable that the brief presentation of images and masks could be hazardous to a subset of individuals with photosensitive epilepsy: out of an abundance of caution, a warning was shown in the description of the task on the Prolific website to be viewed by participants before joining the study. To ensure consistent stimulus presentation, participants were required to complete the task on a desktop or laptop computer; access via tablets and smartphones was programmatically blocked. The title and description of the task on the Prolific website are reproduced as follows:

**Title:** *Identify objects in photos, earn roughly \$4.00 bonus for accurate responses*

**Description:**

*View 540 photos and identify the animal or other object in each photo. Earn a bonus for each accurate response, typically around \$4.00 USD in total (maximum \$5.40).*

**PLEASE NOTE:**

1. *The task begins with a screening phase you must pass with a certain accuracy level before starting the experiment proper. You may re-attempt the screening phase up to 2 times if you wish, but you will only be compensated for a maximum of 5 minutes spent in the screening phase.*
2. *Please avoid resizing the task window during the experiment, as this can break the task. It's best to maximize your browser window and get comfortable before starting.*

**WARNING:** *this task contains bright, rapidly flashing images: it is not suitable for individuals with photosensitive epilepsy.*

Please see Fig. S4 for the additional in-task instructions provided to participants. The task was implemented using the JsPsych library (De Leeuw, 2015) and the JsPsychPsychophysics plugin (Kuroki, 2021).

Table S2: **Statistical results from omnibus test (type II Wald chi-square) and post-hoc pairwise comparisons to test differences among stimulus types in the human image classification experiments.** All  $p$  values are adjusted using the Benjamini-Hochberg correction, with significant values in bold.

Comparison	Human subjects	
	z-value	p-value
Omnibus test (effect of stimulus type)	-	<b>&lt;0.0001</b>
Average Robust Stretch vs Average Standard Stretch	36.63	<b>&lt;0.0001</b>
Robust layer4_conv7 vs Robust layer3_conv1	-4.94	<b>&lt;0.0001</b>
Standard layer4_conv7 vs Standard layer3_conv1	-8.41	<b>&lt;0.0001</b>
Robust layer4_conv7 vs Robust pixel space	-9.52	<b>&lt;0.0001</b>
Standard layer4_conv7 vs Standard pixel space	7.49	<b>&lt;0.0001</b>
Robust layer3_conv1 vs Robust pixel space	-4.66	<b>&lt;0.0001</b>
Standard layer3_conv1 vs Standard pixel space	-0.99	0.32
Robust MEI vs Standard MEI	32.96	<b>&lt;0.0001</b>

### B.1 STATISTICAL ANALYSIS FOR CLASSIFICATION EXPERIMENTS

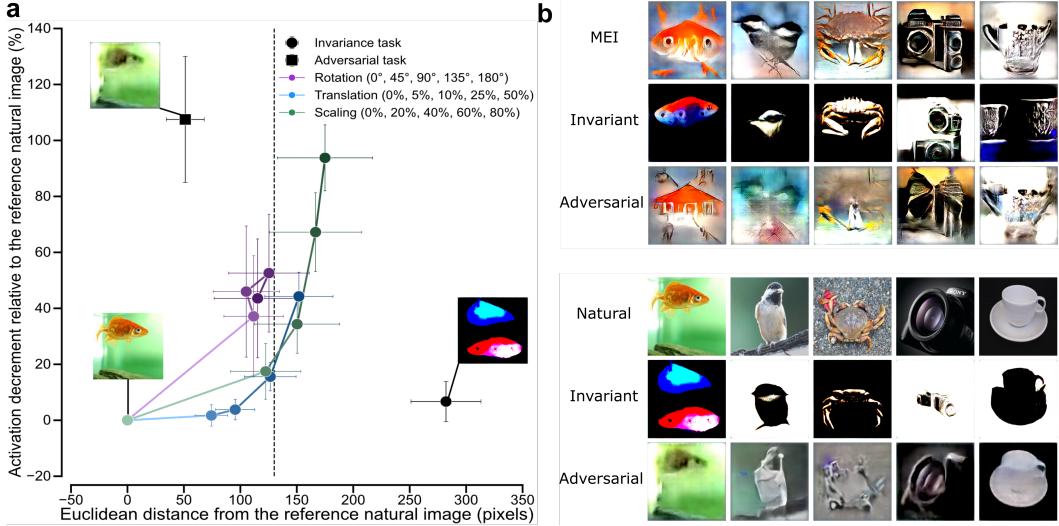
Analysis of data from classification experiments focused on identifying accuracy differences between 9 image types: natural images, MEI’s from robust and standard networks, and invariant images from 3 different layers for both network types. We fitted a generalized linear mixed model (GLMM) with a logistic link function for binary correct vs. incorrect trial responses, including random intercepts for participants and semantic image categories. We used an omnibus test (type II Wald chi-square) to assess overall differences among the 9 image types, followed by 8 planned contrasts via estimated marginal means: (i) robust vs. standard MEIs, (ii) robust vs. standard invariant images averaged across the 3 representation layers, and (iii-viii) within each network type, all 3 possible pairwise comparisons between invariant types at different layers. Our analysis employed the Benjamini-Hochberg procedure to control the false discovery rate under multiple comparisons (Benjamini & Hochberg, 1995). The results of the planned comparisons are provided in Table S2.

### C SNS GENERATES EFFECTIVE ADVERSARIAL AND INVARIANT EXAMPLES FOR NATURAL IMAGES

We replicated the experiment reported in Section 4.1 (see Fig. 2) also in the case where, for a target readout neuron, a highly effective natural image representative of the class encoded by the unit was used as reference for SnS optimization instead of a MEI. As for Section 4.1, both invariant and adversarial images were synthesized by stretching the pixel-level representation. In these experiments, we considered the 12 readout units of robust ResNet50 encoding the 12 image categories used for the interpretability experiment with human observers (see Sections 3.2, 4.3 and Supplementary Section B). For each unit, both the adversarial and invariant experiment was repeated for 10 different reference natural images, the same ones used for the interpretability experiment (for a total of  $12 \times 10$  invariant images and  $12 \times 10$  adversarial images). Example images are shown in Fig. S5b, bottom. Fig. S5a shows that, similarly to the experiment performed with MEI references (see Fig. 2), both adversarial and invariant images reached the desired functional activation regimes (i.e., for adversarial images: silencing the target unit without distancing too much from the reference image; for invariant images: departing substantially from the reference natural image, while at the same time preserving the same level of activation).

### D SINGLE-LOSS ABLATION EXPERIMENTS

To assess the functional contribution of each component of the SnS dual-loss objective (see Section 3.1 for a detailed description), we performed a set of ablation experiments in which we optimized images using only one loss term at a time. We repeated the experiment shown in Fig. 2, in which a readout unit is selected as the optimization target and pixel space is used as the representation



**Figure S5: SnS generates effective adversarial and invariant examples for natural images** **(a)** As in Fig. 2, the ordinate axis represents the activation reduction of a readout unit of  $L_2$ -robust ResNet50 with respect to the activation produced by a reference natural image, when SnS is applied to synthesize either adversarial or invariant images, while the abscissa represents the  $L_2$  pixel distance between these synthesized images and the reference. In this experiment, differently from Fig. 2, the reference images were not the MEIs of the readout units but natural images belonging to the specific categories encoded by the units. Each data point reports the average ( $\pm$  SD) of these metrics across  $12 \times 10$  image synthesis experiments, where 12 are the targeted readout units and 10 are the different reference images selected for each unit. **(b)** Examples of invariant and adversarial images obtained for 5 readout units of  $L_2$ -robust ResNet50, when the reference images were either their MEIs (top), as for the experiment shown in Fig. 2, or natural images (bottom), as for the experiment shown in **(a)**.

space. For each target unit, we synthesized images under four single-loss variants: 1) Squeeze-up: minimize the target unit activation distance between the synthesized image and the reference MEI; 2) Stretch-up: maximize the activation distance from the reference MEI; 3) Squeeze-down: minimize the distance of the synthesized image from the reference MEI in the pixel representation; and 4) Stretch down: maximize such distance.

We used the same 77 readout units and 10 initialization seeds per unit as in the main experiments, resulting in 770 runs per single-loss condition. Squeeze-up and stretch-down followed the initialization and Pareto-front sorting used for SnS invariances; squeeze-down and stretch-up followed the procedures used for SnS adversarial optimizations (see Sections A.3 and 3.1).

Across all four variants, optimizing only one objective consistently failed to drive the other metric (activation distance or pixelspace distance) to its required regime (Fig. S6). None of the single-loss optimizations produced solutions that simultaneously exhibit the trade-off characteristic of SnS invariant or adversarial synthesis. This demonstrates that the dual-loss formulation is necessary for producing meaningful invariances and adversarial examples. Although the single-loss variants did not reproduce SnS behavior, each produced a characteristic synthesis regime that sheds light on the role of its loss component:

- Squeeze-up: Optimizing only the squeeze term yielded images whose activation approached the invariant regime (i.e., close to the MEI’s activation), but whose pixel-space distance remained near the distance of the initial random seeds (see the black curve in Fig S2b, right when iterations is equal to 0). This behavior mirrors repeated XDREAM optimization with multiple initializations, producing multiple high-activation images that differ from each other but do not explore the far boundary of the invariance manifold. The explicit maximization of representation distance (e.g., in pixel space) is the factor that drives invariant images far

away from its reference MEI, exploring the maximal invariant transformations within that representation space.

- Stretch-up: Maximizing the activation distance naturally pushed images to diverge in pixel space as well. The optimization suppressed the target unit by making the image increasingly dissimilar from the reference. This behavior is essentially the opposite of SnS adversarial synthesis, where the unit is silenced with minimal pixel-space change.
- Squeeze-down: Matching the reference MEI in pixel space also preserved the high activation of the reference. This optimization is notably similar to that used for model metamers (Feather et al., 2023), which enforce representation matching at internal layers. Thus, metamer synthesis emerges as a special case of the SnS loss structure when the squeeze term is applied alone to the representation space.
- Stretch-down: Maximizing pixelspace distance alone caused images to diverge visually but produces no systematic change in target-unit activation. Indeed the activation of the target unit remained close to its initial value at the beginning of the optimization (see the black curve in Fig. S2b, left; iteration = 0), which is very far from the one produced by the MEI. This confirms that stretch-down alone does not find images relevant to the unit of interest.

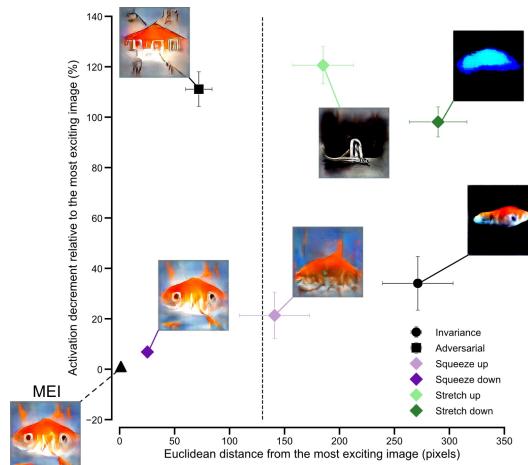
In summary, the single-loss ablations showed that none of the individual terms is sufficient to generate the invariant or adversarial images produced by SnS. The dual-loss structure is essential, with each component stabilizing and constraining the other. These results further validated SnS as a principled framework for probing model invariances and adversarial directions.

## E SNS IS EFFECTIVE IN TARGETING UNITS IN HIDDEN LAYERS

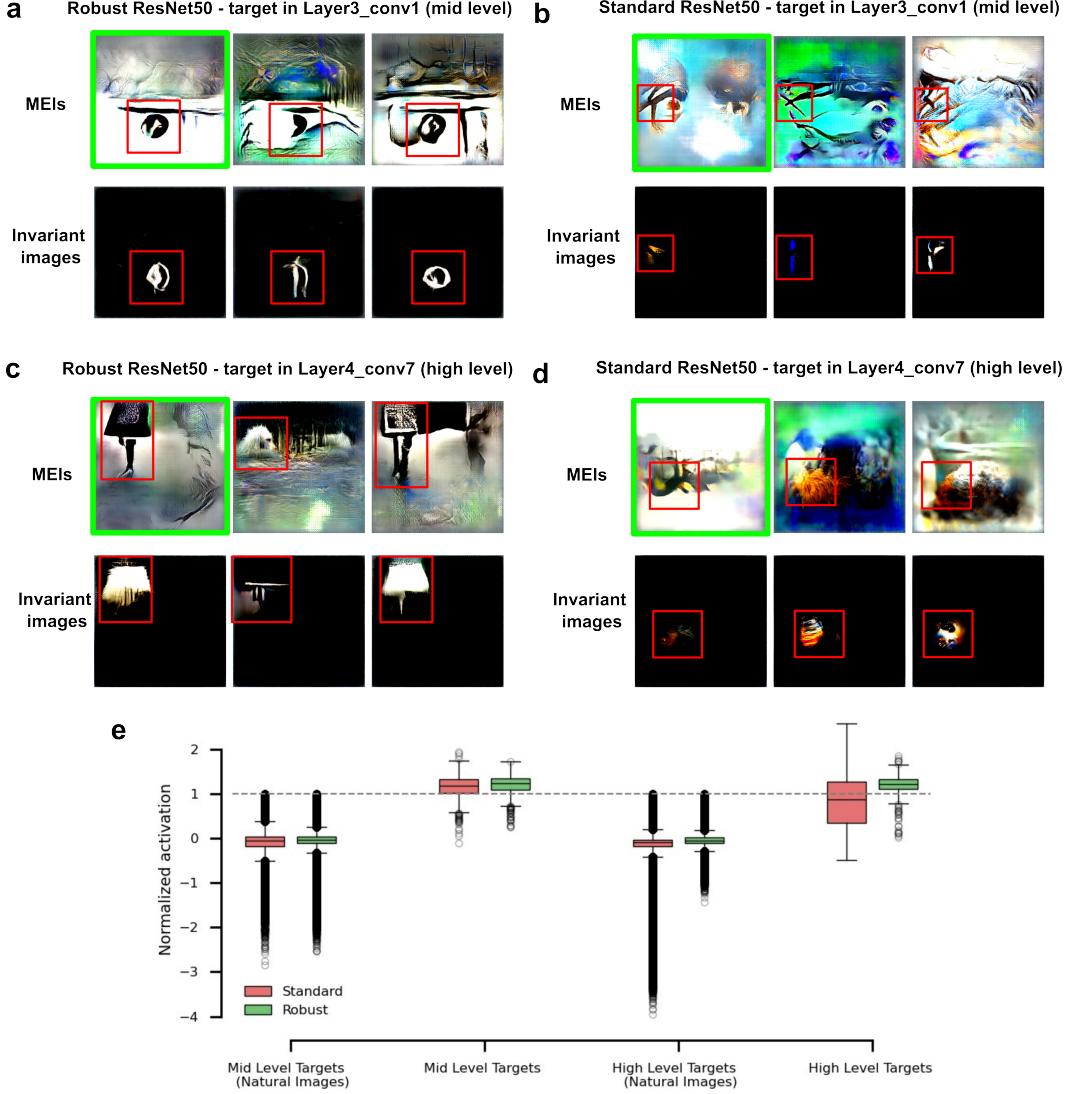
Building upon our findings in Section 4.1, we extended the application of SnS to units within hidden convolutional layers. Unlike output layer units tuned to predefined classes, hidden units are hypothesized to function more analogously to biological visual neurons, responding to intermediate-level features rather than complete objects. This makes them compelling targets for deciphering learned internal representations.

To investigate this, we conducted SnS invariance experiments targeting 50 distinct units in both `layer3_conv1` and `layer4_conv7` of the ResNet50 network, using pixel space as the input representation. For each unit, we performed 10 optimization runs initiated with different random seeds. In the standard network, SnS terminated by early stopping in 67.20% of runs at `layer3_conv1` and 27.40% at `layer4_conv7`; in the  $L_2$ -robust network, convergence was 87.80% and 89.40%, respectively. Importantly, even in non-converging instances, the optimization process often guided the input towards stimuli that elicited high unit activations, consistent with functionally relevant regimes identified through natural image statistics on ImageNet (Fig. S7e).

Qualitatively, the SnS-generated invariant images served as effective “feature visualizers.” This was particularly insightful as the precise selectivity of these hidden units is unknown a priori. The



**Figure S6: Single-loss ablation experiments demonstrate the need of SnS dual loss.** Comparison between SnS adversarial and invariance experiment distances (same as in Fig. 2) with the results of single loss optimizations (squeeze up, squeeze down, stretch up, stretch down). Each point corresponds to the mean  $\pm$  SD of 770 experiments (i.e. 77 readout units  $\times$  10 random initializations per each unit). The black dotted line represents the median image  $L_2$  distance among Imagenet images.



**Figure S7: SnS is effective with targets in hidden layers** (a) Comparison of MEIs (top row) with SnS-generated invariant images (bottom row) for a representative target unit in `Layer3_conv1` of robust ResNet50. SnS images, synthesized to be invariant in pixel space, were generated using the green-highlighted MEI as a reference. Red rectangles identify the unit's key responsive regions. This comparison highlights SnS's ability to distill the core visual features a unit detects, offering clearer insight than MEIs alone. (b-d) The same SnS-based feature visualization approach (as in (a)) is applied to different representative target units: (b) `Layer3_conv1` of a standard ResNet50; (c) `Layer4_conv7` of a robust ResNet50; and (d) `Layer4_conv7` of a standard ResNet50. (e) Final activation distributions for the target units from the SnS optimization in hidden layers. As in Fig. S2a, convergence performance is compared to natural image statistics

resulting visualizations clearly delineated the specific image features to which a target neuron was most sensitive (region of interest (ROI)), effectively isolating these core features from irrelevant contextual details (Fig. S7a-d (bottom rows)). This level of feature disentanglement and clarity surpassed that observed in images optimized by methods like XDREAM (Xiao & Kreiman, 2020), which often retain more complex, less isolated visual elements (Fig. S7a-d (top rows)). Moreover, by comparing the reference image to multiple invariant instances (i.e. initialized with multiple random

seeds), the images revealed specific transformations the identified ROI could tolerate while preserving the unit’s activation level, offering insights into the unit’s invariances (Fig. S7a-d (bottom rows)).

## F SNS CAN BE USED TO STUDY INVARIANCE MANIFOLD GEOMETRY

SnS can be used to extensively sample the invariant manifold of a target unit, paving the way to the analyses of its intrinsic dimension (ID) and geometric structure. To show the feasibility of this approach we selected a single target readout neuron (the `chickadee` unit in the  $L_2$  robust ResNet50 readout layer) and extensively sampled its maximally invariant images across three different stretching layers (the same ones used in Section 4.2): pixel space (input/low-level), `Layer3_conv1` (mid-level), and `Layer4_conv7` (high-level). For each stretching layer, we ran SnS 100 times for each of 5 different reference MEIs (obtained from independent XDREAM runs), yielding 500 invariant images per layer. To estimate the ID of each invariance manifold, we applied two distinct techniques to the vector representations of the invariant images: 1) PCA, measuring the number of components explaining 95% of the variance; and 2) two-nearest-neighbors intrinsic dimension estimation (2NN-ID; Facco et al. (2017)), a state-of-the-art, nonlinear estimator of intrinsic dimension that has been successfully applied to analyze image manifolds in both CNNs and visual cortical areas (Ansuini et al., 2019; Muratore et al., 2022a). 2NN-ID was computed for different subsamples of points (from  $n = 50$  to  $n = 500$  invariant images), repeating the process 50 times for each cardinality.

These two approaches returned dimensionalities in very different numerical ranges: PCA required hundreds of components to explain 95% of variance (from 230 to 340 PCs, Fig. S8a), whereas the ID estimated by 2NN-ID ranged between 10 and 30 across the three manifolds (Fig. S8b). This radical difference in ID estimation between a classic linear method (PCA) and 2NN-ID points to the fact that the structures of all 3 manifolds are highly nonlinear (Facco et al., 2017; Ansuini et al., 2019). This aligns with expectations for neurons high in the visual hierarchy whose responses abstract over complex real-world transformations.

Despite differences in scale, both methods agreed on a consistent ordering of dimensionality across stretching layers: lowest ID for pixel-level stretching, highest for mid-level stretching, and reduced ID again for high-level stretching. This pattern mirrors previously reported trends in the ID of image representations across deep CNNs (Ansuini et al., 2019), where ID initially increases, peaks in mid layers and then gradually diminishes in deeper layers. This result further reinforces the distinction between invariances uncovered at different stretching layers (Fig. 3), indicating the ability of SnS to discover axes of variation that are distinct not only in qualitative terms (Fig. 3a), but also in their dimensionality.

## G SNS IS ROBUST TO SUBSAMPLING OF THE REPRESENTATION SPACE

SnS, as a gradientless and model-agnostic method, offers significant advantages for neuroscience research, particularly in scenarios where constructing detailed “digital twin” models (as described in Section 2) is impractical. However, a critical challenge for applying SnS to study hierarchical invariances in the brain is the inherent undersampling of representational spaces. Despite recent advances in electrophysiology and calcium imaging enabling simultaneous recording from hundreds or thousands of neurons (Jun et al., 2017; Sofroniew et al., 2016), these recordings invariably capture only a fraction of the total neural population within a given brain area.

To evaluate SnS’s performance under conditions analogous to experimental neuroscience, we investigated how subsampling the representational space impacts the qualitative nature and quantitative characteristics (assessed via representational distance analysis, see Section A.5.2 for methodological details) of the invariances identified by SnS. Specifically, we focused on the same 77 readout target units previously analyzed in Section 4.2. For these units, we identified invariances after subsampling their corresponding representational spaces (either `layer3_conv1` or `layer4_conv7` of the CNN) to either 1000 or 100 randomly selected units. To ensure robustness, this process was repeated 10 times for each target unit and subsampling condition, using different random selections of representational space units. As a reference, 1000 units represent 0.5% of the units in `layer3_conv1` ( $n = 200704$ ), while they constitute 4% of `layer4_conv7` ( $n = 25088$ ).

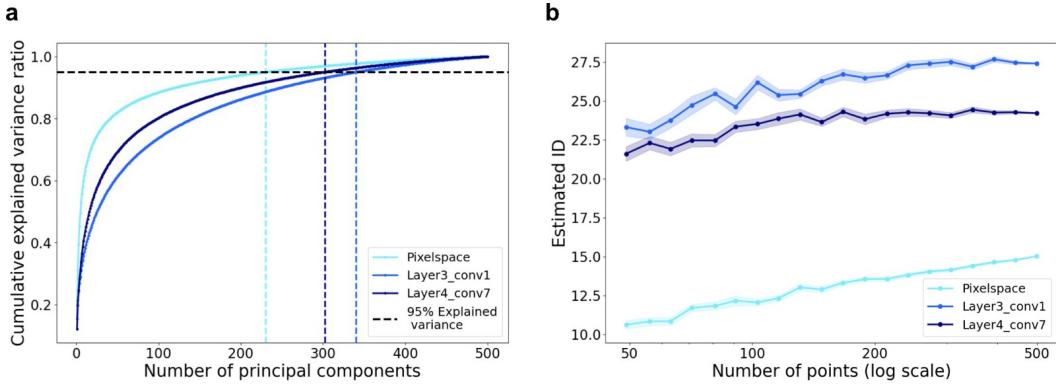


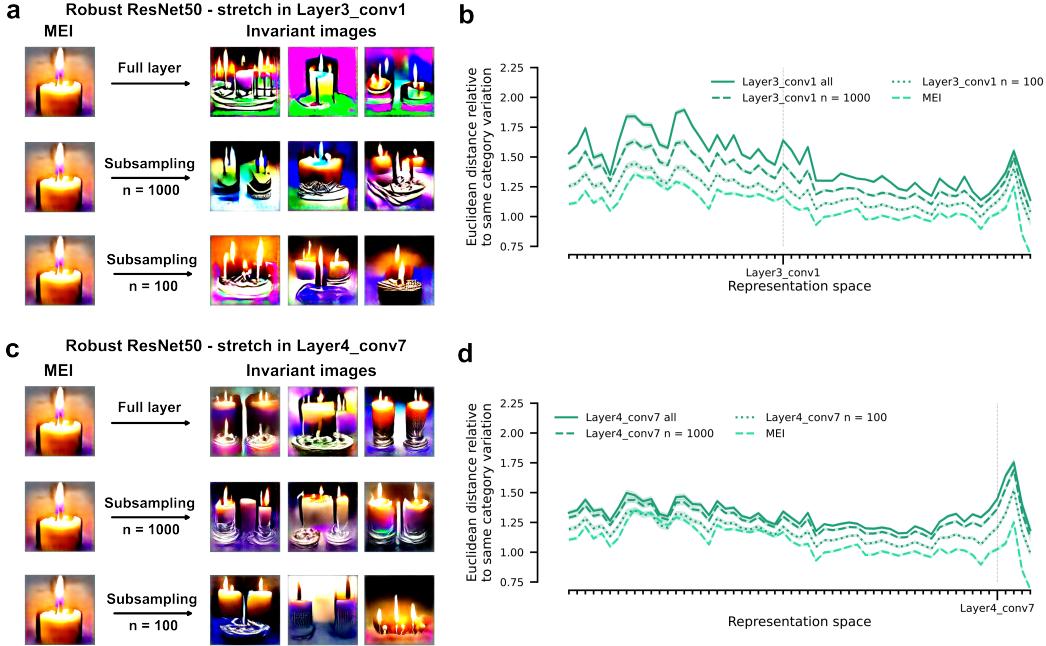
Figure S8: **Hierarchical invariance manifolds differ consistently in their dimensionality.** (a) PCA estimation of the dimensionality of the invariance manifolds obtained for an example  $L_2$  robust ResNet50 readout unit by stretching the representations of the MEIs at different depths of the network: low-level (pixel space), mid-level (Layer3\_conv1) and high-level (Layer4\_conv7). The curves show the cumulative fraction of variance of each manifold (consisting of 500 sampled images) that was explained by considering an increasingly larger number of principal components (ranked according to their eigenvalue). For each stretching depth, the number of principal components necessary to explain 95% of data variance (horizontal black dashed line) is indicated by a color-specific vertical dashed line. (b) Dimensionality of the same invariance manifolds analyzed in a, but estimated using the nonlinear intrinsic dimension method 2NN-ID. The curves show the average intrinsic dimension  $\pm$  SEM, as obtained for increasingly larger sets of invariant images sampled from the manifolds.

Our findings indicate that SnS is robust to such undersampling. Qualitatively, the invariances identified using subsampled spaces were comparable in their level of abstraction to those found using the full representational layer (Fig. S9a and c). This suggests that the SnS optimization process effectively identified relevant axes of image variation even with reduced unit information. This qualitative observation was corroborated by quantitative representation distance analysis (Fig. S9b and d), which demonstrated a strong alignment in the progression of invariance discovery between analyses using full and subsampled layers. These results highlight SnS’s potential for reliably characterizing neural invariances even when constrained by the partial observations typical of in vivo recordings.

A plausible explanation for this robustness to subsampling lies in the inherent representational redundancy present in both biological neural circuits and artificial neural networks. In biological systems, information is often encoded in a distributed manner across populations (Zohary et al., 1994) (Stringer et al., 2019), where multiple neurons may carry similar or overlapping information, contributing to robustness against noise and cell loss (Pouget et al., 2000). Similarly, deep neural networks, including CNNs, have been shown to learn redundant features or possess over-parameterization, where multiple units or weights contribute to similar computations (Denil et al., 2013). Consequently, a sufficiently large and representative subsample of units can still capture the dominant computational characteristics and drive the overall representation into similar regimes as the full layer, explaining the observed consistency in identified invariances.

## H SNS DISCOVERS LAYER SPECIFIC HIERARCHICAL INVARIANCES FOR $L_\infty$ -ROBUST MODELS

To assess whether hierarchical divergence in SnS invariances also arises under  $L_\infty$  robustification, we generated SnS invariances for an  $L_\infty$ -robust ResNet50 ( $\epsilon = 4/255$ ) from RobustBench (Croce et al., 2020) using the same  $n = 120$  natural images and the same three stretching depths (pixel-space, mid-level, high-level) as in Fig. 4. Qualitatively (Fig. S10a) and quantitatively (Fig. S10b), layer-specific invariances in the  $L_\infty$ -robust model remained highly distinct and discriminable in pixel space, mirroring those already shown in Fig. 3. We then tested interpretability across observers by presenting



**Figure S9: SnS invariance is robust to subsampling of the representation space** **(a)** Example MEI and associated invariant images obtained for a readout neuron (candle) in  $L_2$  robust ResNet50 stretching the mid\_level layer (i.e. Layer3\_conv1). Each row represents examples for different levels of subsampling: using the full layer (1<sup>st</sup> row), subsampled spaces of size  $n = 1000$  and  $n = 100$  (2<sup>nd</sup> and 3<sup>rd</sup> rows respectively). Multiple results are shown for different random initialization seeds (columns). **(b)** Normalized average distance between the invariant images generated by SnS (stretching in Layer3\_conv1) and their reference MEIs across the different stages of a robust ResNet50. Different lines indicate different cardinalities of units in the representation space used for optimization (i.e. all,  $n = 1000$  and  $n = 100$ ). The average distance between multiple MEIs generated via XDREAM is reported for comparison. **(c-d)** Same as **(a)** and **(b)** respectively, but stretching was performed on robust ResNet50 in the high\_level layer (i.e. Layer4\_conv7).

the  $L_\infty$ -robust invariances to three architectures (ResNet-18, Wide-ResNet50-2, ConvNeXt-Base), each in standard (Paszke et al., 2019) and  $L_\infty$ -robust versions ( $\epsilon = 4/255$ , also from RobustBench (Croce et al., 2020)). As in the  $L_2$  setting, invariances from the  $L_\infty$ -robust generator were more interpretable on average than those from a standard ResNet50, and interpretability again decreased when robust invariances were evaluated by standard networks (Fig. S10c). Notably, however, when  $L_\infty$  invariances were judged by other  $L_\infty$ -robust observers, classification accuracy either remained at ceiling or increased across stretching depths, suggesting that  $L_\infty$ -robust networks share a highly aligned invariance structure even at deeper layers. This observation highlights promising directions for future work comparing invariances across robustness regimes, as well as in models whose robustness arises from perceptually aligned training strategies (e.g., harmonized neural networks (Fel et al., 2022)) or biologically inspired architectures such as VOneResNet (Dapello et al., 2020).

## I HIERARCHICAL INVARIANCES IN VISION TRANSFORMERS

One of the strengths of SnS is that, being gradient-free, it can easily be applied (literally out-of-the-box) to virtually any visual processing architecture (biological or artificial), as long as one can measure the responses of its units to the images produced by SnS optimization algorithm. In cases where the goal is to uncover the invariance of the whole model to transformations of specific object classes, the additional constraint is to have access to the responses of the readout units for those

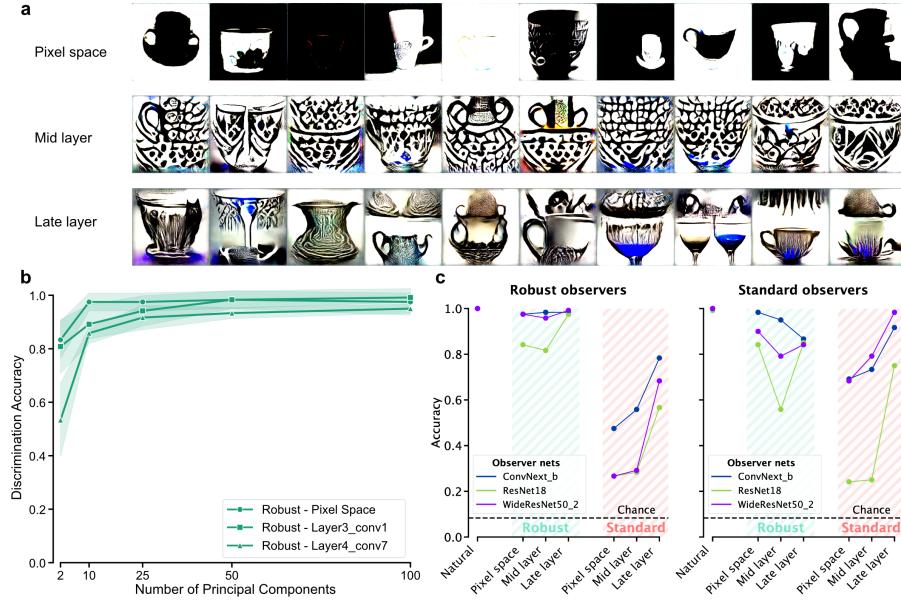


Figure S10: **Invariances in  $L_\infty$ -robust ResNet50.** (a) Example invariant images obtained for a readout neuron ("cup") in  $L_\infty$ -robust ResNet50 and computed for Pixel space, Layer3\_conv1, Layer4\_conv7 (rows). (b) Accuracy of a SVC in discriminating the three classes of invariant images produced by stretching at the aforementioned layers as a function of the number of principal components fed to the classifier. (c) Classification accuracies for  $L_\infty$ -robust and standard ResNet50 invariances across multiple  $L_\infty$ -robust and standard observer networks.

classes. As such, SnS can directly be applied to investigate the invariance of readout units in vision transformers (ViTs).

To test this application, we ran the experiments of Section 3.2 (observer networks only) on a ViT-B-16 model pretrained on ImageNet and a CLIP-trained ViT-B-16 visual encoder (Radford et al., 2021). Since the CLIP model lacks explicit readout units, we trained a linear classifier atop the frozen visual encoder for ImageNet classification to construct class-specific detectors (top 1 validation accuracy: 78.94%). For both models, we considered the output linear layer of the MLP in the sixth encoder layer as mid-level layer and the one in the twelfth encoder layer as high level representation.

Our findings reveal that, similar to CNNs, the nature of the discovered invariances in ViTs depends on the hierarchical level of the stretched representation. Stretching the pixel space primarily altered low-level properties such as luminance, yielding invariances that were poorly interpretable (Fig. S11 a,b). Contrary to the experiments with ResNet50 (see Fig. 3 a,b), where a clear, qualitative distinction was observable between invariant images stretched in mid and late layers, in both ViTs we did not observe such difference (see Fig. S11 a,b). These images were also comparable in terms of invariance interpretability (Fig. S11 c). This observation is consistent with the view that ViTs learn less strictly hierarchical representations than CNNs, and that these models integrate information more globally (Dosovitskiy et al., 2020). Overall, there seems to be a little, but consistent across conditions advantage of the CLIP-trained version of ViT-B-16 in terms of interpretability. This trend seems to be similar between standard and robust network observers.

## J SNS INVARIANCES AND METAMERS ARE STRUCTURALLY AND FUNCTIONALLY DISTINCT

To better contextualize SnS invariances within state-of-the-art techniques to probe invariances in neural networks, we carried out a systematic comparison between SnS-derived invariant images and gradient-based metamers. Metamers were generated using the original implementation of Feather et al. (2023), including the same hyperparameter settings mentioned in the original paper. Metamers

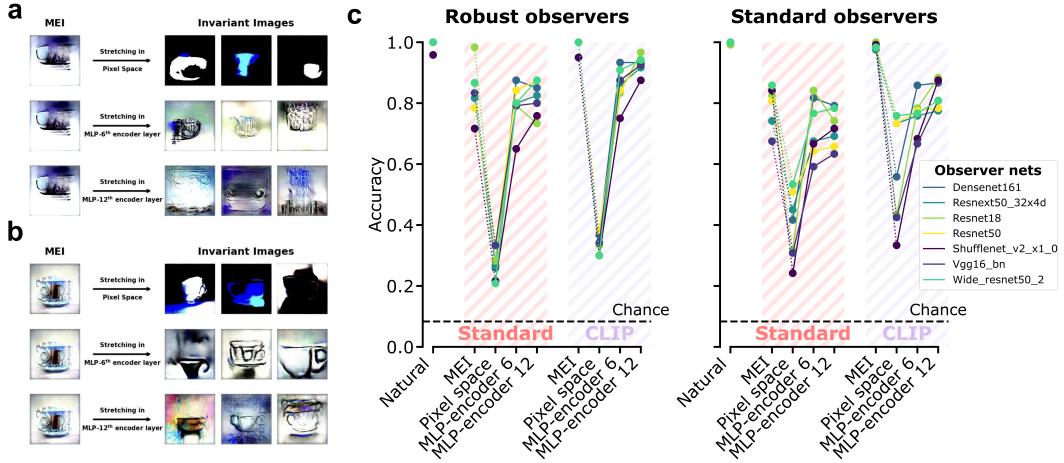


Figure S11: **Invariances in vision transformers.** (a) Example MEI and associated invariant images obtained for a readout neuron (cup) in standard ViT and computed for three different choices of the stretching representational stages (rows). Multiple results are shown for different random initialization seeds (columns) (b) Same as (a) but for a ViT encoder trained via CLIP. (c) Classification accuracies for ViT invariances across multiple robust (left) and standard (right) networks.

were synthesized for the same  $n = 120$  natural images used in our behavioral experiment (Fig. 4), yielding for each natural image one SnS invariant and one metamer at each of three representational depths (pixel space, mid-level, high-level) in both standard and  $L_2$ -robust ResNet50 (for details see Section 3.1). First we compared metamers and SnS invariances in terms of interpretability. To do this, we replicated the experiment described in Fig. 4c with metamer stimuli. Using the same observer networks as in our main experiments (six architectures, each in both standard and  $L_2$ -robust versions;  $n = 12$  total (Section 4.3)), we recovered results consistent with Feather et al. (2023) (Table S3). Specifically, metamers produced by the  $L_2$ -robust ResNet50 remained highly interpretable across observer models at all synthesis depths (pixel space, layer 3, layer 4). In contrast, standard-network metamers showed the expected sharp decline in interpretability for late-layer metamers (layer 4) across all observer networks. Critically, this pattern is the opposite of what we observed for SnS hierarchical invariances (Fig. 4c; numerical values in Table S3). Indeed, for standard networks, late-layer SnS invariances show a pronounced increase in interpretability, so much so that observer accuracies far exceed those obtained for late-layer metamers for the standard network.

A second point of stark contrast between the two methods is that of average distances from the reference natural images from which both metamers and SnS invariant images were generated (Table S4). Indeed, metamers are explicitly optimized to minimize representational distance to the target image at a given layer, whereas SnS invariances explicitly maximize that distance. As a result, metamers lie much closer to their reference images than SnS invariances at all depths and for both standard and robust generator networks. For comparison, we also report the average pairwise distances between natural ImageNet images from the same class at the same three layers (i.e., within-category distance for the three representation stages, see Supplementary Section A.5.2). While metamers lie far below these natural-image baselines, SnS invariances systematically exceed them by a wide margin (Table S4). In the context of the different image synthesis between the two methods, we remind that, as detailed in Supplementary Section D, metamer-like stimuli arise as a special case of the SnS objective (i.e. the squeeze-down single loss variant in Fig. S6). In this sense, SnS provides a gradient-free generalization of metamer-like synthesis, enabling analogous invariance characterizations in settings - such as in vivo experiments - where gradients are not accessible.

**Table S3: Interpretability comparison between metamers and SnS hierarchical invariances.** Classification accuracies (mean  $\pm$  SEM) of the standard and  $L_2$ -robust observer networks used in the interpretability experiment (Fig. 4c), evaluated on images generated either as metamers or using SnS with both standard and  $L_2$ -robust ResNet50 as the generator network (**Gen Net** column). Metrics are reported for the three different representation layers used also in Fig. 4 (**Net Layer** column). Note that accuracy results for the metamers conditions are in line with Feather et al. (2023).

Condition	Gen Net	Net Layer	Standard observers	Robust observers
Metamer	Standard	Pixel space	$0.999 \pm 0.00122$	$0.989 \pm 0.00653$
Metamer	Standard	Mid layer	$0.939 \pm 0.01878$	$0.883 \pm 0.05348$
Metamer	Standard	Late layer	$0.310 \pm 0.02817$	$0.121 \pm 0.01388$
Metamer	$L_2$	Pixel space	$0.998 \pm 0.00122$	$0.989 \pm 0.00653$
Metamer	$L_2$	Mid layer	$0.997 \pm 0.00163$	$0.985 \pm 0.00816$
Metamer	$L_2$	Late layer	$0.994 \pm 0.00245$	$0.990 \pm 0.00694$
SnS	Standard	Pixel space	$0.397 \pm 0.087$	$0.224 \pm 0.014$
SnS	Standard	Mid layer	$0.476 \pm 0.099$	$0.256 \pm 0.022$
SnS	Standard	Late layer	$0.779 \pm 0.070$	$0.636 \pm 0.024$
SnS	$L_2$	Pixel space	$0.843 \pm 0.052$	$0.936 \pm 0.040$
SnS	$L_2$	Mid layer	$0.722 \pm 0.035$	$0.857 \pm 0.041$
SnS	$L_2$	Late layer	$0.721 \pm 0.030$	$0.888 \pm 0.009$

**Table S4: Comparison of average distances with the reference natural image between metamers and SnS invariances.** Mean  $\pm$  SEM are reported for each representation layer used in the experiments (Pixel space, Mid layer, and Late layer of ResNet50 standard and  $L_2$ -robust variants). Distances between natural images belonging to the same ImageNet category are reported as reference. For every network layer, SnS invariances yield substantially larger distances than both the natural-image baseline and the metamers.

Condition	Net Layer	ResNet50	ResNet50 ( $L_2$ )
Metamer	Pixel space	$0.7998 \pm 0.01523$	$0.7998 \pm 0.01523$
Metamer	Mid layer	$978.2 \pm 6.483$	$6.181 \pm 0.5169$
Metamer	Late layer	$356.8 \pm 10.48$	$8.254 \pm 0.9402$
SnS	Pixel space	$281.4 \pm 5.991$	$281.8 \pm 4.94$
SnS	Mid layer	$1935 \pm 8.864$	$371.8 \pm 2.603$
SnS	Late layer	$858.2 \pm 22.75$	$406.3 \pm 6.442$
Natural	Pixel space	$143.4 \pm 1.447$	$142.4 \pm 1.669$
Natural	Mid layer	$1478 \pm 8.138$	$201.3 \pm 2.268$
Natural	Late layer	$384.8 \pm 7.128$	$239.1 \pm 1.845$