Emergent Neural Network Mechanisms for Generalization to Objects in Novel Orientations

Anonymous authors
Paper under double-blind review

Abstract

The capability of Deep Neural Networks (DNNs) to recognize objects in orientations outside the training data distribution is not well understood. We investigate the limitations of DNNs' generalization capacities by systematically inspecting DNNs' patterns of success and failure across out-of-distribution (OoD) orientations. We present evidence that DNNs (across architecture types, including convolutional neural networks and transformers) are capable of generalizing to objects in novel orientations, and we describe their generalization behaviors. Specifically, generalization strengthens when training the DNN with an increasing number of familiar objects, but only in orientations that involve 2D rotations of familiar orientations. We also hypothesize how this generalization behavior emerges from internal neural mechanisms — that neurons tuned to common features between familiar and unfamiliar objects enable out of distribution generalization — and present supporting data for this theory. The reproducibility of our findings across model architectures, as well as analogous prior studies on the brain, suggests that these orientation generalization behaviors, as well as the neural mechanisms that drive them, may be a feature of neural networks in general.

1 Introduction

Recognizing objects in novel orientations lies at the heart of biological and artificial intelligence, as it is a fundamental capacity necessary to understand the visual world Sinha & Poggio (1996); Ullman (1996). However, the computational mechanisms underlying this capacity in both brains and machines are not yet well understood Gazzaniga et al. (2006); Pinto et al. (2008); Li et al. (2021).

In the realm of artificial systems, Deep Neural Networks (DNNs) have recently made large strides in object recognition He et al. (2017); Carion et al. (2020). However recent studies have shown that DNNs perform poorly when objects are presented in novel orientations, even when learning from large datasets with millions of examples Barbu et al. (2019); Alcorn et al. (2019); Madan et al. (2022). More broadly, novel orientations are a special case of *out-of-distribution* (OoD) data. DNNs' generalization is often limited to the images from the training distribution, known as *in-distribution* data, while it remains difficult to give a principled account of their performance in OoD settings.

One promising approach to understand the capabilities of DNNs is to leverage the knowledge gained from studying biological intelligence Hassabis et al. (2017); Ullman (2019). In natural settings, biological intelligent agents observe instances of object categories from diverse orientations. When encountering a new object instance, these agents often demonstrate the capacity to accurately identify the object in different orientations by drawing upon past experiences with similar instances Booth & Rolls (1998); Freiwald & Tsao (2010); Ratan Murty & Arun (2015). Extensive investigations into human and mammalian perception and object recognition in unfamiliar orientations have revealed that recognition accuracy varies across novel orientations, with some orientations exhibiting superior generalization compared to others Logothetis & Pauls (1995). Additionally, studies into the neural mechanisms underlying these cognitive abilities have marshaled compelling evidence suggesting that neurons respond to their own specific set of object features when present in the visual field Desimone et al. (1984); Kobatake & Tanaka (1994); Gauthier et al. (2002); Fang & He (2005). This neural tuning has been reported to be invariant to a certain degree from the object's orientation Logothetis & Sheinberg (1996). Theoretical frameworks have proposed that such neural invariance to object orientation forms the basis for the ability to recognize objects in novel orientations within biological systems Poggio & Anselmi (2016).

In this paper we employ these same analytical tools utilized in the study of biological brains in order to understand DNNs' generalization abilities in OoD orientations. Specifically, we study DNNs under conditions akin to the operating regime of biological brains, in which some instances of an object category (*e.g.,* a 'Boeing 777 airliner' is an instance of the 'airplane' category) are seen from all orientations during training (*fully-seen* instances), while other instances are only seen in a subset of all orientations (*partially-seen* instances). During test time, we evaluate the generalization performance of the networks by measuring instance classification performance on OoD orientations (*i.e.,* those orientations not included in the training set) of *partially-seen* instances. This simple paradigm, inspired by Jang et al. (2023), facilitates analyzing the impact of several key factors that may influence OoD generalization, such as the number of *fully-seen* instances and the *in-distribution* orientations of the *partially-seen* instances. This paradigm allows us to more precisely characterize performance challenges of DNNs for OoD orientations. Figure 1 summarizes the paradigm that we follow in this work.

It remains unclear whether DNNs are at all capable on the OoD task we outlined above. We therefore quantify DNNs' classification accuracy in OoD orientations. Our novel analytical approach to these measurements yields two important findings: 1) DNNs readily generalize to certain surprising OoD orientations, 2) these generalizable orientations are parameterized by the *in-distribution* set, and are quantifiably evaluated. We also measure invariance of the neural representations at the individual-unit level, across orientations and object instances, supporting a theory on brain-like mechanisms that drive the emergence of orientation invariance Logothetis & Sheinberg (1996); Poggio & Anselmi (2016).

These findings may have broad impact on artificial general intelligence as DNNs are commonly used as both visual encoders and end-to-end models for visual tasks. We have replicated our results with several architectural variations on CNNs He et al. (2016); Huang et al. (2017); Kubilius et al. (2018) and Visual Transformers (ViT) Vaswani et al. (2017); Dosovitskiy et al. (2020). The CNN and ViT architectures comprise almost

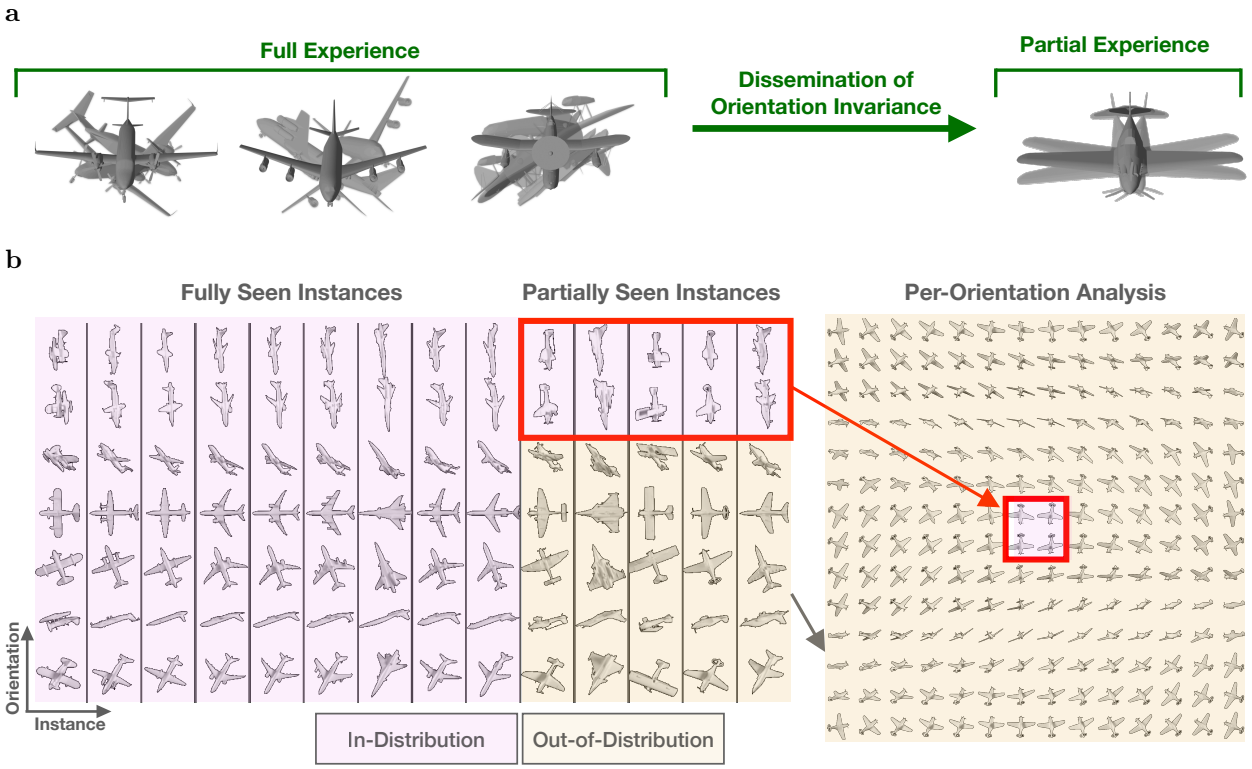

Figure 1: **Learning paradigm and network's per-orientation accuracy. (a)** The network is trained with images of certain airplanes at all orientations, constituting a 'full experience,' and for other airplanes, a small subset of orientations, constituting a 'partial experience.' The *out-of-distribution*, or OoD, generalization capacities of the network are evaluated by measuring the classification accuracy for *partially-seen* airplanes at unseen orientations. Our results suggest that OoD generalization is facilitated by the dissemination of orientation invariance developed for all orientations for the *fully-seen* airplanes to the OoD orientations of the *partially-seen* airplanes. **(b)** Left: The learning paradigm employed in this work. Each column is a sample object instance (here from the airplane dataset) and each row is a sample orientation. The training set includes all orientations for *fully-seen* instances, and a partial set of orientations (outlined in red) for *partially-seen* instances (in this example, with the airplanes' nose pointing down). The orientations included in the training set are referred to as *in-distribution* orientations (pink shading). Orientations of the *partially-seen* instances that are not included in the training set are referred to as OoD (yellow shading). Right: A visualization of per-orientation-analysis. The set of all orientations are arranged to capture proximity relationships between orientations. (Further details are provided in Fig.2a.)

all current implementations of artificial intelligence vision solutions. These solutions include scaling classic CNNs to match transformer parameter counts and data volumes Liu et al. (2022) and integrating CNN and Transformer architectures Liu et al. (2021); Radford et al. (2021). Modern state of the art vision-language models follow the same trend, with Qwen2-VL Wang et al. (2024) and Llama 3-V Dubey et al. (2024) both using ViT in the visual encoder. The continued success of these architectures suggests that the DNN generalization behaviors and mechanisms we discuss will have staying power.

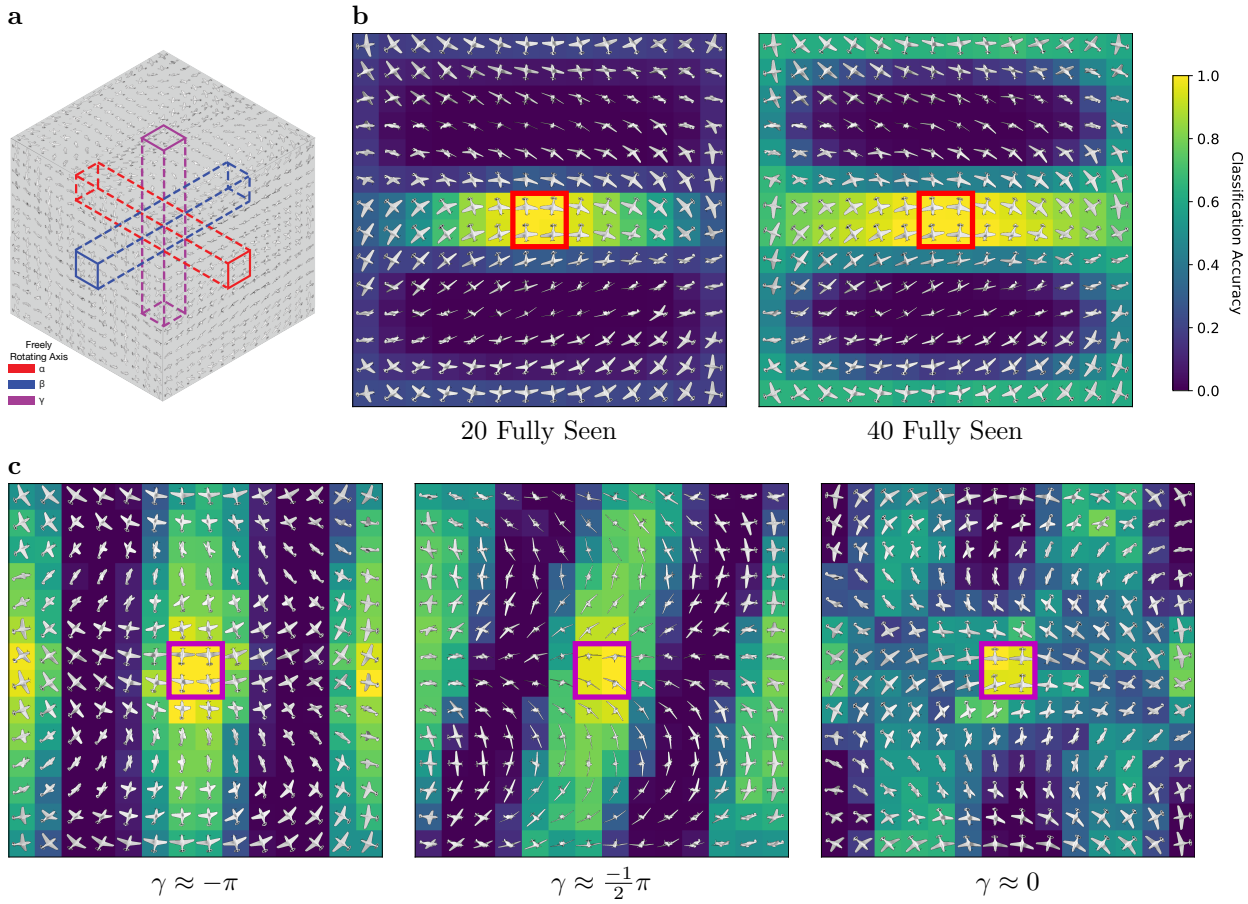

Figure 2: **Observed generalization patterns in per-orientation accuracy heatmaps.** When trained with a combination of *fully-seen* instances and *partially-seen* instances, DNNs demonstrate the ability to generalize outside of their training distribution. Generalization behaviors are demonstrated measuring per-orientation accuracy. **(a)** All orientations can be described by three Euler axes $(\alpha, \beta, \gamma,)$ and rotations are periodic around these axes. These properties allow for the visualization of all possible orientations with an orientation cube, shown here. The orientations contained within the colored rectangular prism are those orientations of the *partially-seen* instances included in training (*i.e.,* are *in-distribution*). The *in-distribution* orientations differ depending on the experiment. All other orientations are OoD. **(b)** Increased network generalization for OoD orientations, with increased instance diversity (*i.e.,* number of *fully-seen.*) Each cell in the heatmap is the average classification accuracy of the network for a given value of β and γ, across all values of α. Chance level is 0.02 (2%). **(c)** Different *in-distribution* parameters affect the generalization behaviors. The generalization patterns for a different span of *in-distribution* orientations $(-0.25 \leq \alpha \leq 0.1, -0.1 \leq \beta \leq 0.25, -\pi \leq \gamma < \pi)$ as outlined by the purple box. In this case, each cell is of a given value for α, β, γ.

2 Results

Overview. In exploration of the generalization capabilities of DNNs to novel orientations, our computational experiments show that the OoD orientation space is divided between *generalizable* and *non-generalizable* orientations, in terms of the networks behavior. We find this division to be governed by a set of rules which determine a partitioning of the orientation space, given a 'seed' of orientations for test instances seen by the DNNs at during training. Among several partitioning rules, we identify a rather intuitive one: small 3D perturbations around seen orientations will be included in the highly-generalizable partition. We find other rules to be more surprising, including that the highly-*generalizable* partition also consists of shape

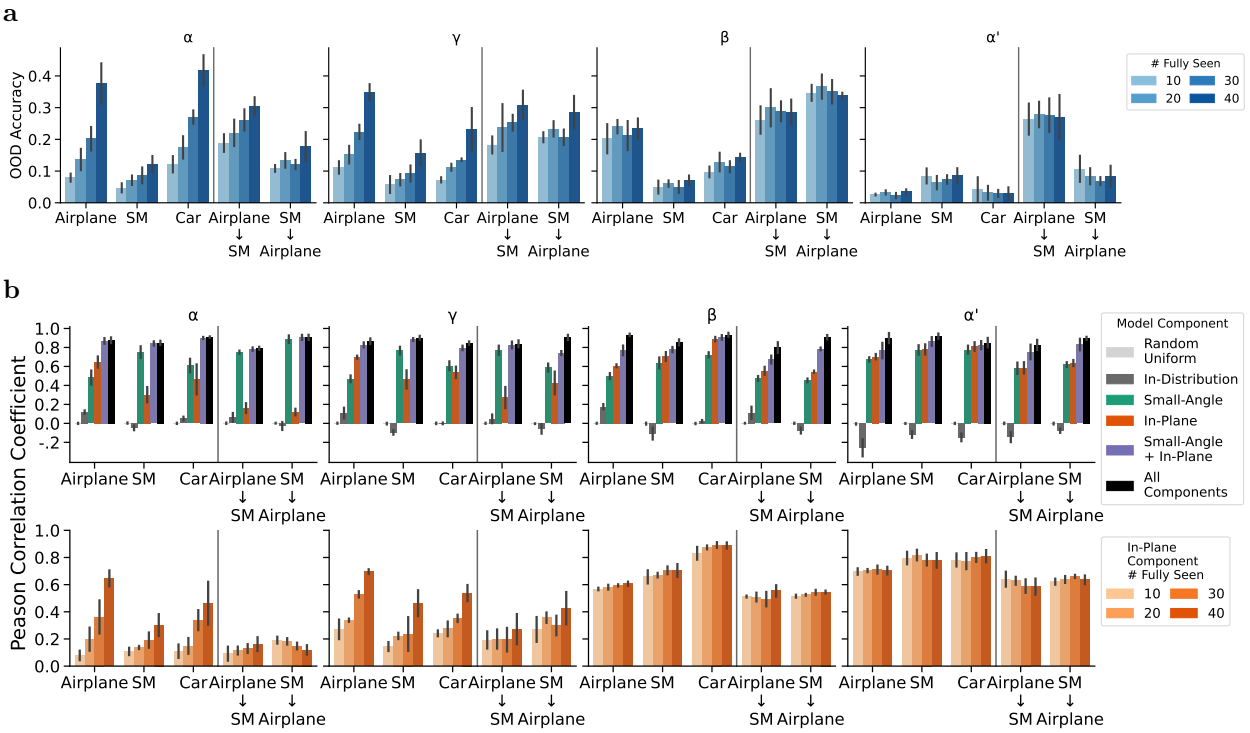

Figure 3: **Modeling generalization patterns for OoD orientations.** The bar plots show several trends related to DNN OoD classification patterns. The trends are measured under the various controls, including *in-distribution* orientations conditions $(\alpha, \gamma, \beta, \alpha')$ and object category, which is either a single object as in Airplane, SM, Car, or transfer across two categories, when the *fully-seen* instances are of a different category than the *partially-seen* instances as in Airplane → SM and vice versa. These transfer cases are visually separated from the other cases. **(a)** Network generalization for OoD orientations increases with increasing number of fully seen (blue shading.) This trend holds across object category and *in-distribution* orientations conditions. **(b)** Top: We introduce a predictive model for OoD orientation generalization (black — "All Components") which is highly predictive of experimental results, with greater than 0.8 Pearson Correlation Coefficient for all experimental controls. (Results are shown for experiments with 40 *fully-seen* instances.) Null hypothesis predictive models, including "Random Uniform" and "In-Distribution," have very low correlation coefficients. We also ablate our predictive model, including only some sub-components, like only-"Small Angle", only-"In-Plane" or only-"Small Angle + In-Plane." These ablated models have lower correlation coefficients than "All Components," and vary in relation to one another depending on the experimental condition. Bottom: We isolate the predictive power of the only-"In-Plane" component for all experiments with a range of number of *fully-seen*. The increasing predictive power of the "In-Plane" component correlates with increasing OoD accuracy as the number of *fully-seen* instances increases. This suggests that generalization to "In-Plane" orientations drives OoD accuracy.

and silhouette preserving rotations, such as *in-plane* (*i.e.,* 2D) rotations and flips along axes of symmetry of the seen seed orientations. In order to quantitatively assess this hypothesis, we evaluate the degree of correlation between predicted network behavior induced by these partitioning rules and measured behavior. We find them to be highly correlated under a variety of training regimes. We also explore the DNNs' internal representations and identify neuronal mechanisms that allow for the dissemination of orientation-invariance from familiar objects to novel objects and orientations.

Per-orientation accuracy heatmaps. A detailed inspection of the network's generalization capability for OoD orientations is enabled by introducing per-orientation accuracy heatmaps. In brief, the continuous space of object orientations, represented by Euler angles, is discretized into *cubelets* (Fig. 2a) – local areas

of the rotation space. For each *cubelet* the network's performance is evaluated in terms of the classification accuracy $\Psi(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is an orientation of interest. Accuracy heatmaps are 2D projections across a specified dimension of the full accuracy orientation cube (Fig. 2b,c; Methods). These heatmaps reveal a reproducible pattern of generalization in the form of increased classification accuracy for OoD (*i.e.,* novel) orientations. For example, Figure 2b shows that for 'seed' orientations at the center of the heatmap (red box), the network (in this experiment - ResNet18 He et al. (2016); see Methods) yields the highest accuracy (brightest *cubelets*) for adjacent orientations around the 'seed', depicting small 3D perturbations of the 'seed' orientations. Further inspection of the heatmap reveals other orientations, in this example, brighter *cubelets* forming the figure '8' (stretching 'seed' sideways and along the heatmap's boundaries, enclosing two darker 'holes'), for which the network performs better than for the rest of the OoD orientations. These orientation mainly depict *in-plane* rotations of the 'seed' orientations.

When considering the average classification accuracy across all OoD orientations, our experiments reproduce previous results. In particular, we can reliably quantify the effect of data diversity on the OoD generalization, as the amount of training examples is kept constant with our learning paradigm. Figure 3a clearly shows an increase in OoD accuracy as data diversity (*i.e.,* the number of *fully-seen* instances) increases, under various conditions, including different 'seed' orientations, different image datasets and across datasets. The accuracy heatmaps provide a complementary means of assessment to the overall average accuracy measure, depicting the generalization patterns and indicating which orientations account for the network increased performance (*e.g.,* Fig. 2b). The patterns of increased accuracy depict a partitioning of the orientation space, which reappears for various 'seed' orientations (Fig. 2c), various sizes of the training set and different object categories (e.g., *Airplane*, *Car*, *Shepard & Metzler (SM)* objects Shepard & Metzler (1971)), as shown in several experiments (Methods).

Modeling generalization patterns. We hypothesize a set of rules which govern the partitioning of the orientation space into *generalizable* and *non-generalizable* orientations. To quantitatively evaluate this hypothesis we formulate a model of the partitioning rules, which can be used to predict the OoD generalization patterns of the network, given a 'seed' of *in-distribution* orientations. Briefly, the model, denoted by $f_{\mathbf{w}}(\boldsymbol{\theta})$ has three components: $A(\boldsymbol{\theta})$, which captures small angle rotations around $\boldsymbol{\theta}$; $E(\boldsymbol{\theta})$, which captures in-plane (2D) rotations; $S(\boldsymbol{\theta})$, which captures object silhouette projections at the orientation $\boldsymbol{\theta}$ (see details in Methods). We evaluate the model's performance by measuring the Pearson correlation coefficient ρ between the accuracy of the networks as measured in our experiments and as predicted by the model, *i.e.,* $\rho(\Psi(\boldsymbol{\theta}), f_{\mathbf{w}}(\boldsymbol{\theta}))$. Figure 3b shows the predictive power of the model and its components in experiment with different 'seed' orientations and several object categories. The model's component $A(\boldsymbol{\theta})$ ('small angle' rotations), is the best predictor for the network's OoD behaviour, for highly articulated objects such as the *SM* objects. On the other hand, the model's component $E(\boldsymbol{\theta})$ ('in-plane' rotations), is a better predictor for non-articulated objects with inherent symmetries. Further analysis of this component illustrates how generalization to 'in-plane' rotations emerges with the increase in data diversity.

We conducted a large series of experiments under various settings, including different 'seed' orientation distributions, various amounts of training examples, different object scales, object categories with different levels of symmetry, image datasets (Fig. 5) and DNN architectures (see Methods). In all experiments our model highly predicts the network's behavior, indicating that indeed the networks generalization patterns for OoD orientations follow the model's partitioning rules. This is true even across categories, when the 'seed' is taken from one category (*e.g., SM*) and the 'fully-seen' instances are taken from another (*e.g., Airplane*).

Individual unit neuronal analysis. In search of how generalization and dissemination emerge in DNN's we turn to analyze the neurons' activation in the trained networks. We focus on neurons in the penultimate layer of the network, which are attuned to the highest level features in the input stimuli, but reflect a consolidated representation of the entire network for inferring the downstream task (classification in our simulations).

Figure 4a illustrates activation of individual neurons for stimuli of *fully-seen* and *partially-seen* instances. Each group of images depict input stimuli of a particular instance for which a particular neuron has the highest activation, along with the neuron's per-orientation activation heatmap for the instance. The pat-

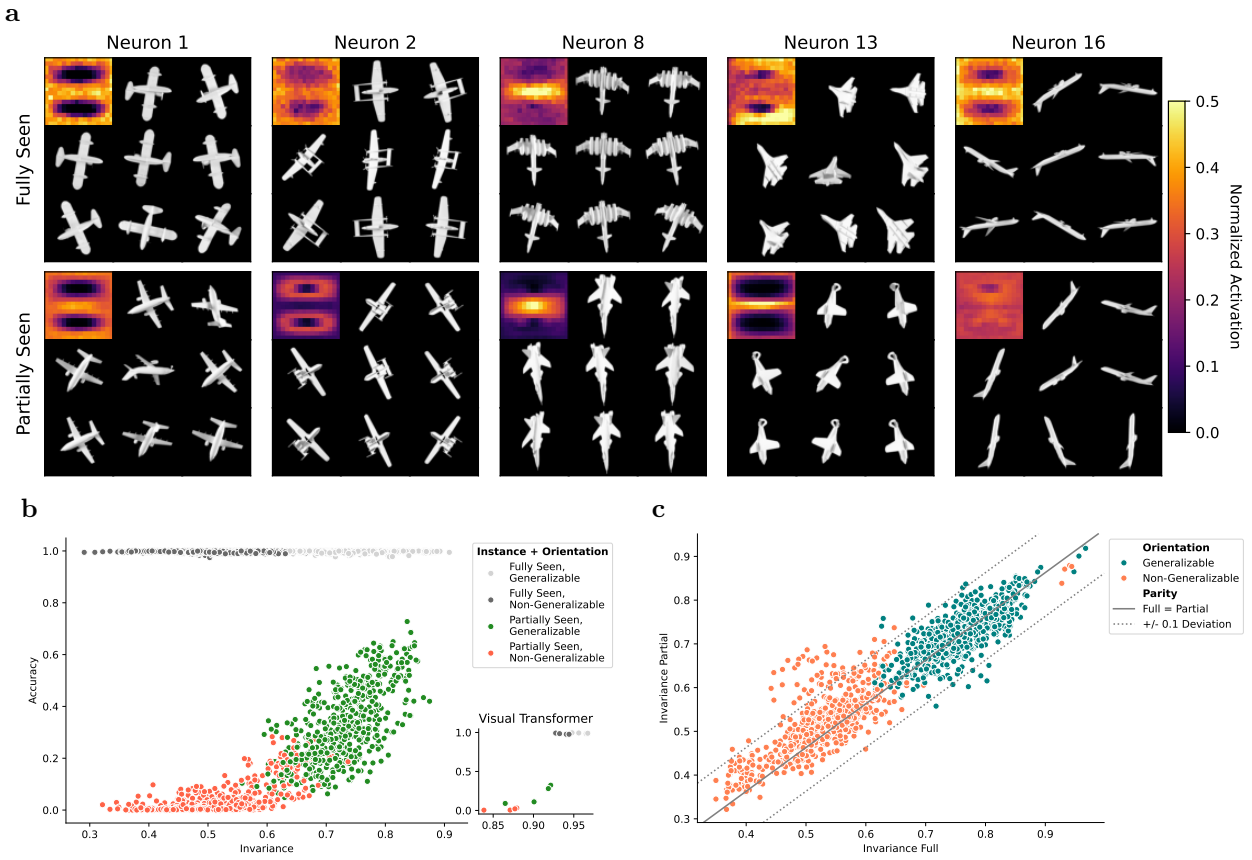

Figure 4: **Neuronal analysis, Invariance and Dissemination. (a)** An intuitive visualization of neural activity. Each square is the response of a single neuron to the airplane instance that most highly activates, it portrayed in two ways: 1) the top-8 images that most highly activate the neuron (in no particular order), 2) the heatmap of the per-orientation normalized neural activity for the airplane instance. Neurons tend to exhibit patterns of activation related to the patterns of generalization behavior (Figs. 2b for example,) and are invariant to a range of orientations that respect the partitioning of OoD orientations. Comparing the neural responses in each column demonstrates that the patterns of activation are similar between the *fully-seen* instance that most highly activates the neuron and the *partially-seen* instance that most highly activates it due to shared visual, part and semantic features between these instances. Several randomly sampled penultimate layer neurons, arranged into columns, demonstrate that these findings apply to many neurons. **(b)** Each dot represents the results of an experiment, across different number of *fully-seen* instances, object type and DNN architecture. Averaging the activations in the partitioned regions ('seed', *generalizable*, *non-generalizable*) and computing the invariances (defined here: Eq.6) between 'seed' and OoD regions captures overall generalization in the network. Plotting the generalization metrics against accuracy for those regions demonstrates a clear correlation between increasing invariance and increasing OoD classification accuracy (*i.e., partially-seen* instances in OoD orientations.) Visual Transformer results are placed separately to highlight that they follow the same trend, though their invariance is scaled higher. **(c)** Plotting *fully-seen* invariance against *partially-seen* invariance for the same experiment also yields a tight correlation, suggesting that dissemination of invariance from *fully-seen* to *partially-seen* instances enables increasing generalization in OoD orientations of *partially-seen* instances.

terns seen in the neurons' activation heatmaps resemble the partitioning patterns of the accuracy heatmaps shown in Figure 2b. Some neurons exhibit similar activation patterns for both *fully-seen* and *partially-seen* instances, while others do not.

A quantifiable measure of these neuronal responses can help with understanding how generalization occurs in the network – particularly generalization to OoD orientations of *partially-seen* instances, where generalization must stem only from 'seed' orientations seen during training. We define an activation invariance score in the range $[0, 1]$ (Eq. 6) between sets of orientations, in particular between the 'seed' orientations and OoD *generalizable* orientations or *non-generalizable* orientations. The invariance score yields higher values when a neuron fires for both sets of orientations, and lower values when it fires only for one set (see details in Methods). We expect that generalization, reflected by the accuracy level, will correlate with the invariance score.

Figure 4b depicts a scatter plot of the invariance score against the classification accuracy. Each dot represents an experiment (a tuple of number *fully-seen*, object and architecture type) and the coloring indicates the respective instance set (*fully-seen* or *partially-seen*) and orientation set (*generalizable* or *non-generalizable*). As expected, there is a clear correlation between increasing levels of classification accuracy and increasing invariance score for the *partially-seen* instances. Furthermore, the plot shows a clear partition between *generalizable* and *non-generalizable* orientations with respect to the invariance score, where significantly higher invariance scores are measured for the *generalizable* orientations.

For *fully-seen* instances (Fig. 4b gray dots), all orientations are *in-distribution*, including the 'seed', *generalizable* and *non-generalizable*. Therefore, the network easily achieves accuracy at ceiling levels regardless of the neuronal invariance score. Nevertheless, the *fully-seen* instances exhibit the same invariance partitioning between *generalizable* and *non-generalizable* orientations as the *partially-seen* instances.

Figure 4c depicts a direct comparison between the invariance score of the *fully-seen* and *partially-seen* sets for both the *generalizable* and *non-generalizable* orientations. The partition between *generalizable* and *non-generalizable* orientations is exhibited again — the *non-generalizable* invariances are in the bottom left corner, while the *generalizable* invariances are in the top right corner. Each point in this plot represents the joint invariance of *fully-seen* and *partially-seen* instances at a given orientation. The plot shows a tight correlation between the invariance scores of the *fully-seen* and *partially-seen* instances, as most of the points lie within a band roughly 0.1 units away from the line of parity, $x = y$. This correlation suggests that an increase in the invariance score of the network at a set of orientations for the *fully-seen* instances will be disseminated to the *partially-seen* instances.

3 Methods

3.1 Per-Orientation Accuracy Visualization

Previous works have typically reported average performance over all orientations. In contrast, we evaluate the network's performance for each orientation across the entire range of orientations. To express an orientation of an object instance we use $\boldsymbol{\theta} := (\alpha, \beta, \gamma)$, the Euler angles with respect to the orthogonal axes of a reference coordinate system $\mathbb{R}^3$ Goldstein et al. (2002), with the convention that α and γ are bounded within 2π radians, and β is bounded within π radians. We define $\Psi(\boldsymbol{\theta}) \in [0, 1]$ to be the network's average classification accuracy at an orientation $\boldsymbol{\theta} = (\alpha, \beta, \gamma)$ over either the *fully-seen* or *partially-seen* instances.

To facilitate intuition of Ψ we introduce a visual representation of this function. Since orientations are continuous values and are related spatially we map the range of bounded values of orientations (α, β, γ) onto a Cartesian coordinate system, resulting in a cube—the basis of our visualization. We discretize the continuous space of orientations into *cubelets*, which are sub-cubes with a width of $\frac{1}{\#Cubelets}$ of the full range of each respective angle. This approach preserves local behavior in aggregate analysis. In addition, we outline the range of orientations which are *in-distribution* for the *partially-seen* instances — the rest are OoD orientations. To illustrate the object orientation at a given *cubelet*, we sample one representative image and overlay it onto the heatmap at the location of the *cubelet*.

See Fig. 2a, which shows this visual representation scheme, and Figs. 2b and 2c for examples.

3.2 Model of DNN Per-Orientation Generalization

In the Results section, we briefly introduce the hypothesis that DNNs are capable of generalizing to orientations which are small angle rotations of the *in-distribution* orientations images and to orientations that are *in-plane* relative to the *in-distribution* images. In this section we formalize this model.

Recall that we defined $f_{\mathbf{w}}(\boldsymbol{\theta})$ as the predictive model for generalization per each orientation. To measure the goodness of our prediction, we employ the Pearson correlation coefficient to measure how closely our model correlates with DNN recognition accuracy, $\Psi(\boldsymbol{\theta})$. We choose this metric because it normalizes data with respect to amplitude and variance, and therefore measures patterns of behavior across $\boldsymbol{\theta}$ and relative to other $\boldsymbol{\theta}$, rather than the exact performance for every $\boldsymbol{\theta}$.

Our model $f_{\mathbf{w}}(\boldsymbol{\theta})$ is composed by three components ($A(\boldsymbol{\theta})$, $E(\boldsymbol{\theta})$ and $S(\boldsymbol{\theta})$), which we introduce next. These three components easily lend themselves to formalization with Euler's rotation theorem Goldstein et al. (2002). The theorem states that any rotation can be uniquely described by a single axis, represented by a unit vector $\hat{\mathbf{e}} \in \mathbb{R}^3$, and an angle of rotation, denoted as $\phi \in [0, \pi]$ around the axis $\hat{\mathbf{e}}$. We employ this representation to describe the rotation between an arbitrary orientation of interest, $\boldsymbol{\theta}$, and an orientation in the set of *in-distribution*, denoted $\boldsymbol{\theta}_s \in \Omega_s$. We use $\hat{\mathbf{e}}_{\boldsymbol{\theta},\boldsymbol{\theta}_s}$ and $\phi_{\boldsymbol{\theta},\boldsymbol{\theta}_s}$ to denote the unit vector (axis) and the angle of this rotation, respectively.

Component 1: Small Angle Rotation, $A(\boldsymbol{\theta})$. The first component of the model captures orientations that are small angle rotations from the orientations in the training distribution. Visually similar orientations are those that are arrived at by small rotations from *in-distribution* orientations, or small $\phi_{\boldsymbol{\theta},\boldsymbol{\theta}_s}$. We therefore define the first component $A(\boldsymbol{\theta})$ as

$$A(\boldsymbol{\theta}) := \max_{\boldsymbol{\theta}_s \in \Omega_s} \left| 1 - \frac{\phi_{\boldsymbol{\theta},\boldsymbol{\theta}_s}}{\pi} \right| \in [0, 1]. \tag{1}$$

The $\max_{\boldsymbol{\theta}_s \in \Omega_s}$ operator chooses the *in-distribution* orientation that is closest to $\boldsymbol{\theta}$ of interest.

Component 2: In-plane Rotation, $E(\boldsymbol{\theta})$. The second component of the model captures orientations which appear as *in-plane* rotations of *in-distribution* images. Let $\mathbf{c} \in \mathbb{R}^3$ be the unit vector representing the camera axis. *In-plane* rotations are those for which the axis of rotation is parallel to the camera axis. Thus, an orientation appear as an *in-plane* rotations of an *in-distribution* images when $\mathbf{c} \in \mathbb{R}^3$ and $\hat{\mathbf{e}}_{\boldsymbol{\theta},\boldsymbol{\theta}_s} \in \mathbb{R}^3$

(*i.e.,* the vector of object instance rotation) are parallel. Taking their standard inner product yields the proximity to being parallel, which is therefore the degree to which the rotation is *in-plane*.

Thus, we define the second component $E(\boldsymbol{\theta})$ as follows:

$$E(\boldsymbol{\theta}) := \max_{\boldsymbol{\theta}_s \in \Omega_s} \left| \mathbf{c}^\top \hat{\mathbf{e}}_{\boldsymbol{\theta},\boldsymbol{\theta}_s} \right| \in [0,1], \tag{2}$$

where $\mathbf{c}^\top$ denotes the transpose of $\mathbf{c}$.

Component 3: Silhouette, $S_A(\boldsymbol{\theta})$, $S_E(\boldsymbol{\theta})$. The third component of the model captures orientations which project object silhouettes onto the camera that are similar to the silhouettes of the object when *in-distribution* — for example, the airplane when viewed from above, and the silhouette being the airplane viewed from below. These orientations are defined as a π radians rotation around the γ axis, which results in a silhouette orientation. We transform all the *in-distribution* orientations, Ω_s, in this way, and we call these silhouette *in-distribution* orientations $\Omega_{\hat{s}}$. We then compute $S_A(\boldsymbol{\theta})$ and $S_E(\boldsymbol{\theta})$, substituting $\Omega_{\hat{s}}$ for Ω_s in $A(\boldsymbol{\theta})$ and $E(\boldsymbol{\theta})$ respectively.

Nonlinearities. The components described above capture a general trend, but do not match the range of values given by a 0-100% accuracy metric. We therefore fit the components with a logistic function. The 'S'-like shape of the logistic function allows for the highest and lowest values of $E(\boldsymbol{\theta})$, $A(\boldsymbol{\theta})$, $S_A(\boldsymbol{\theta})$ and $S_E(\boldsymbol{\theta})$ to be close to the highest and lowest values of $\Psi(\boldsymbol{\theta})$. In addition, it allows for a smooth transition between these highest and lowest values. Most importantly, the simplicity of the logistic function allows for fitting while preserving the interpretability of the model components, ensuring that the models remains related to small angle, *in-plane* and silhouette rotations. We employ the following logistic function:

$$\sigma(x; (a,b,c)) = \frac{1}{1 + e^{b(-x^c + a)}}, \tag{3}$$

where $x \in \{E(\boldsymbol{\theta}), A(\boldsymbol{\theta}), S_A(\boldsymbol{\theta}), S_E(\boldsymbol{\theta})\}$. a and b translate and scale the values of the predictive components and c spreads out saturated values of the component.

Fitting the Model with Gradient Descent. The model combines four components $A(\boldsymbol{\theta})$, $E(\boldsymbol{\theta})$, $S_A(\boldsymbol{\theta})$ and $S_E(\boldsymbol{\theta})$ by taking the sum of their respective values after applying the logistic function σ:

$$\begin{aligned} & f_{\mathbf{w}}(\boldsymbol{\theta}) \\ := & \sigma(A(\boldsymbol{\theta}); \mathbf{w}_A) + \sigma(E(\boldsymbol{\theta}); \mathbf{w}_E) + \\ & \sigma(S_A(\boldsymbol{\theta}); \mathbf{w}_{SA}) + \sigma(S_E(\boldsymbol{\theta}); \mathbf{w}_{SE}), \end{aligned} \tag{4}$$

where $\mathbf{w}$ represents the parameters of the logistic functions *i.e.,* $\mathbf{w} = (\mathbf{w}_A, \mathbf{w}_E, \mathbf{w}_{SA}, \mathbf{w}_{SE})$. The logistic fitting function is differentiable, and $f_{\mathbf{w}}(\boldsymbol{\theta})$, the linear combination of these logistic functions, is also differentiable. Further, the Pearson correlation coefficient is also differentiable. Therefore we employ gradient descent to fit $\mathbf{w}$ with the Pearson correlation coefficient as the cost function.

3.3 Neural Analysis

In the Results section, we discussed our findings that OoD generalization in the network is allowed for by dissemination of orientation invariance from *fully-seen* instances to *partially-seen* instances. In this section, we outline the process by which we quantify several different network invariance metrics. We first formalize the notation for neural activations for single orientations and for sets of orientations. We then define the invariance score (Eq. 6). Finally, we average together many invariance calculations to arrive at the network invariance metric.

We begin by formalizing our approach to neural activations. In Sec.3.1 we introduced $\Psi(\boldsymbol{\theta})$, the network's average accuracy at a specific orientation. We can similarly define the neural activation at a specific orientation, though we do so with more granularity. Namely, we introduce $\Phi_i^n(\boldsymbol{\theta})$, which is the average activation

of a neuron n from the set of all penultimate-layer neurons N (*i.e.,* $n \in N$) across all images of an object instance i from the set of all object instances I (*i.e.,* $i \in I$) for a given orientation $\boldsymbol{\theta}$. We normalize the activity of each neuron by dividing the activity level of each image by the maximum activity generated by any image. We exclude any neurons with a maximum activation of 0 from further analysis.

Having defined Φ we note that it is useful to perform analysis not on single orientations only, but sets of orientations. We demonstrated that under our experimental conditions, orientations can be partitioned into coherent subsets — *in-distribution* and OoD orientations. Further, the OoD orientations can be partitioned into *generalizable* orientations, *i.e.,* those OoD orientations that the network can generalize to, and *non-generalizable* orientations. We refer to the *in-distribution*, *generalizable* and *non-generalizable* orientation sets as *InD*, G and $\neg G$ respectively. The determination of membership of the *generalizable* and *non-generalizable* orientation sets is as follows: We compute 10% of the maximum value of $f_{\mathbf{w}}(\boldsymbol{\theta})$, the predictive model, in the experiment with 40 *fully-seen* instances. All orientations for which f is greater than the 10% threshold are considered *generalizable*, otherwise they are considered *non-generalizable*. We can now compute the average activation of a set or orientations. For example, the average activation for a given neuron n and object instance i of the *generalizable* orientations is defined in the following way:

$$\bar{\Phi}_i^n(G) = \frac{1}{|G|} \sum_{\theta \in G} \Phi_i^n(\theta). \tag{5}$$

The same may be computed for *in-distribution* and *non-generalizable* orientations.

To determine how dissemination occurs in the network, we calculate the degree of similarity in a neuron's response to a given instance across different orientations. Specifically, given a neuron n and instance i, we calculate the similarity between the neuron's response at an orientation pair $\Phi_i^n(\boldsymbol{\theta}_1)$, and $\Phi_i^n(\boldsymbol{\theta}_2)$, or pair of sets of orientations $\bar{\Phi}_i^n(InD)$, $\bar{\Phi}_i^n(G)$ for example. We use δ, *invariance score*, as the similarity metric, which is defined (based on previous work Madan et al. (2022)) in the following way:

$$\delta(\bar{\Phi}_i^n(InD), \bar{\Phi}_i^n(G)) = 1 - \left| \frac{\bar{\Phi}_i^n(G) - \bar{\Phi}_i^n(InD)}{\bar{\Phi}_i^n(G) + \bar{\Phi}_i^n(InD)} \right|. \tag{6}$$

We note that under some conditions, δ reports a high, yet trivial, invariance. Namely, if the response of a neuron is low or zero for both elements of the pair, the denominator approaches zero and the invariance becomes large. However in this case the neuron is not responding to anything — any activity is most likely noise. We therefore calculate a threshold of activity for neural response invariances to be considered to contribute to the generalization capability of the network. Otherwise, these invariances are not integrated into the overall network invariance metric. The threshold, τ, is the 95th percentile of activity for all neurons across all images. We employ τ with an indicator function as follows:

$$\mathbf{1}(\bar{\Phi}_i^n(InD), \bar{\Phi}_i^n(G)) := \begin{cases} 1 & \text{if } \bar{\Phi}_i^n(InD) \geq \tau \wedge \bar{\Phi}_i^n(G) \geq \tau \\ 0 & \text{otherwise} \end{cases}.$$

Finally, we can compute the overall network *generalizable* and *non-generalizable* invariance scores. To do so, we compute a triple average: an average activation over the set of orientations (Eq. 5) and averaged over the invariance of all neurons and object instances. We say that the *generalizable* invariance score is the invariance between the *in-distribution* orientations and the *generalizable* orientations determined as follows:

$$\frac{1}{L} \sum_{n \in N} \sum_{i \in I} \mathbf{1}(\bar{\Phi}_i^n(InD), \bar{\Phi}_i^n(G)) \cdot \delta(\bar{\Phi}_i^n(InD), \bar{\Phi}_i^n(G)), \tag{7}$$

where L is the quantity of activity pairs above the threshold τ, *i.e.,*

$$L = \sum_{n \in N} \sum_{i \in I} \mathbf{1}(\bar{\Phi}_i^n(InD), \bar{\Phi}_i^n(G)). \tag{8}$$

The definition of the network's *non-generalizable* invariance score is the same, though $\neg G$ replaces G.

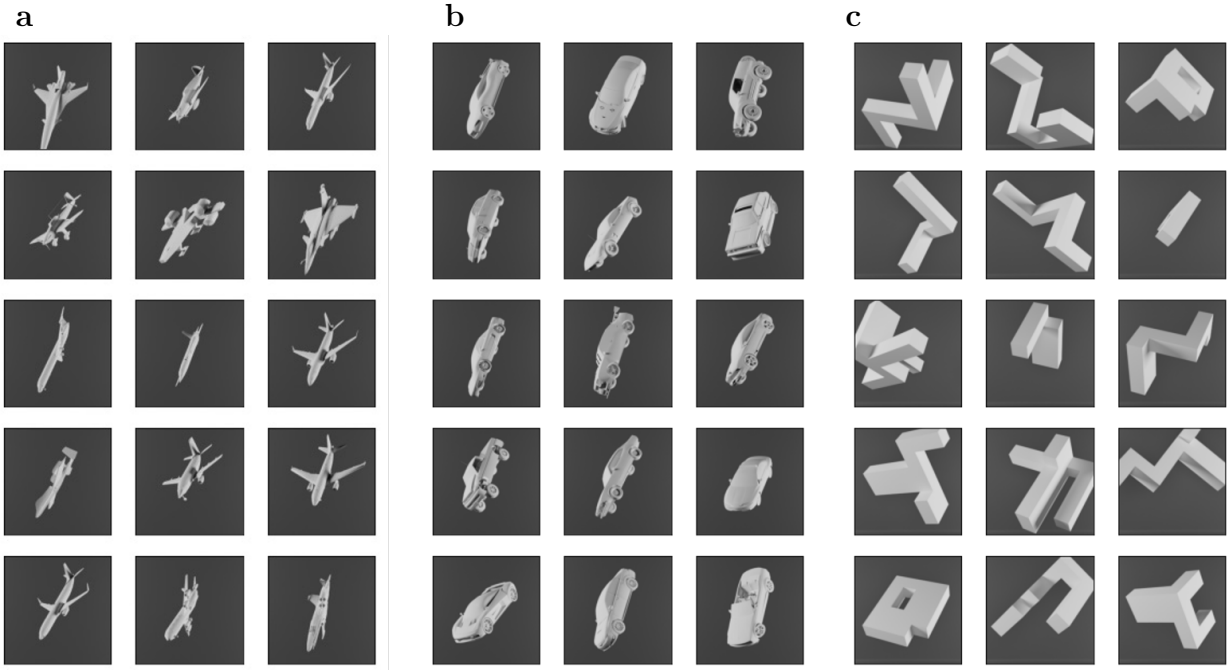

Figure 5: **Object Datasets.** In our experiments we used three object categories: (**a**) Airplanes, (**b**) Cars, and (**c**) Shepard&Metzler objects. The first two were curated from ShapeNet Chang et al. (2015) and we procedurally generated the last one. There are 50 instances per object category (*e.g.,* 'Concorde' or 'Spitfire' for the Airplanes). Images were rendered from the 3D models under fixed lighting conditions, and the models were centered and fully contained within the image frame. For *fully-seen* instances (see Fig. 1), orientations were uniformly sampled at random using Euler angles in the range of $-\pi \leq \alpha < \pi, -\frac{\pi}{2} \leq \beta < \frac{\pi}{2}, -\pi \leq \gamma < \pi$. For *partially-seen* instances, orientations were uniformly sampled from a subset of these ranges.

3.4 Experimental Controls

Proportion of *fully-seen* instances. We vary diversity in terms of the number of *fully-seen* instances N between 10 (20% of the total number of instances) and 40 (80%). The remaining instances are *partially-seen*. For a fair evaluation of the effect of data diversity, the amount of training examples is kept constant as we vary the data diversity.

Object Categories. We used three categories of objects: Airplanes, Cars and *Shepard&Metzler* objects. For the airplanes and cars we curated 50 high quality object instances of each category from the ShapeNetChang et al. (2015) database. Both airplanes and cars have clear axes of symmetry, which allow for intuition of how networks generalize to OoD orientations. We therefore also experimented with highly asymmetric objects similar to those tested for 3D mental rotations in Shepard & Metzler (1971) (which we denote as *Shepard&Metzler* objects; Fig. 5).

DNN Architectures. We used ResNet18 He et al. (2016), DenseNet Huang et al. (2017), CORnet Kubilius et al. (2018) and Visual Transformers (ViT) Vaswani et al. (2017); Dosovitskiy et al. (2020) in our experiments. The first two were chosen as they are representative feed-forward DNNs. The architecture of CORnet is brain-inspired and includes recurrence at higher layers in addition to convolutions in lower layers.

Repetition. We re-run each experiment five times, each time randomly sampling the specific instances which comprise the *fully-seen* and *partially seen* sets.

Hyperparameters for training. We trained the three deep convolutional neural networks using the Adam Optimizer Kingma & Ba (2017) with following learning rates and batch seizes, respectively:

Architecture	Learning Rate	Batch Size
1. ResNet18	10^{-3}	230
2. DenseNet	10^{-3}	64
3. CORnet	10^{-4}	128
4. ViT	10^{-4}	256

Table 1: Training Hyperparamters

Batch sizes were chosen to be as large as possible while still fitting the model, the batch of images and forward-pass computations in memory. Learning rates were chosen from $10^x, x \in \{-1, -2, -3, -4, -5\}$ to be as large as possible while ensuring that OoD generalization remained stable. Each network was trained for 10 epochs. After this point *in-distribution* performance was stabilized at 100% and OoD performance reached an asymptote.

Dataset Size. Each dataset is 200k images, 4k image for each of the 50 object instances. A training epoch iterates through every image in the dataset once.

Hardware details. Experiments were run with one CPU, 25GB of memory and on several generations of Nvidia GPUs with a minimum of 11GB of memory.

ViT Experiments Experiments involving visual transformers proved to be far more expensive to run, due to their increased size, required training time, and more expensive hardware. For this reason we ran only a subset of controls for ViT: only airplanes, freely rotating on α, and only two repetitions. These experiments yielded almost identical results to the other experiments (see Fig. 4b,c) and therefore running the full range of controls was unnecessary.

4 Discussion

A large number of previous works have explored the generalization capacities in DNNs. For example Lenc & Vedaldi (2015); Gruver et al. (2023) investigated the emergence of invariance, and specifically rotational invariance, in an array of architectures. Other works, including by Cohen and Welling (Group Equivariant CNN's Cohen & Welling (2016) and Steerable CNN's Cohen & Welling (2017)) induce invariance by construction with modifications to the CNN architecture. These works generalize the translation inductive bias inherent in CNN's to broader mathematical groups, including rotations. However, they only focus on affine rotations and don't address non-affine, or even more challenging OoD cases. This limits the relevance of these findings to more realistic scenarios.

In this work we analyze the generalization behaviors of DNN's on rendered images and observe dissemination of orientation-invariance for orientations that appear like 2D rotations (*in-plane*) of *in-distribution* orientations. In some cases, when the network relies on the object instance's silhouette for recognition, the *in-distribution* orientations also include orientations that have the same silhouette as the seen orientations. For *non-generalizable* orientations, the network has not developed orientation-invariance with respect to the seed orientations (demonstrated by the lower invariance score in our results). It is worth noting that despite the absence of orientation-invariance, the network is still able to recognize *fully-seen* instances in such *non-generalizable* orientations. This is due to the fact that these orientations fall within the training distribution and the network has learned to associate them with their corresponding object instances. However, in the case of *non-generalizable* orientations, the dissemination of orientation-invariance is not feasible. This is even the case when neurons are tuned to features shared with *partially-seen* instances, as they do not exhibit orientation-invariance for these *non-generalizable* orientations and the training process does not provide any information to establish associations with the corresponding object instances.

Further, our results support the hypothesis that the network disseminates orientation-invariance of *fully-seen* instances to *partially-seen* instances using brain-like mechanism similar to those reported by Logothetis & Sheinberg (1996); Poggio & Anselmi (2016). Neurons are feature detectors, and during training neurons are tuned to detect the features of *fully-seen* objects at multiple orientations — *i.e.,* the neurons become selective to the feature, but invariant to the orientation. Some features that neurons are tuned to are shared between *fully-seen* and *partially-seen* instances (Fig.4a). Therefore the invariance that develops for features of *fully-seen* instances are gained "for free" for *partially-seen* instances in the same orientations. Our results provide a quantitative assessment of this hypothesis and elucidate the intricate neural processes involved in object recognition, underscoring the critical role of individual neuron, feature-based representations for OoD object recognition.

This study reveals discernible patterns in the successes and failures of DNNs across diverse orientations which can be effectively characterized and explained through the analysis of neural activity. This underscores the potential for more comprehensive analyses of DNNs that transcend the conventional approach of solely focusing on average accuracy.

A key question arising from our results is to explain why DNNs disseminate orientation-invariance only to *in-plane* orientations. All object instances are distinguishable at all orientations, as evidenced by the high *in-distribution* accuracy achieved by the DNNs. Therefore the lack of orientation-invariance for such *non-generalizable* orientations is an outcome of the DNN's learning process. We speculate that this may be because orientations that are not *in-plane* are affected by self-occlusion, which poses a particular challenge for DNNs. Various efforts have been made to enhance DNNs' generalization capabilities to OoD orientations including leveraging preconceived components for DNNs, such as 3D models of objects Angtian et al. (2021) or sophisticated sensing approaches like omnidirectional imaging Cohen et al. (2018). However, these approaches rely on *ad-hoc* approaches tailored to specific objects and do not address the fundamental limitations of the DNN learning process in recognizing objects in OoD orientations. Instead, novel network architectures that extend the emergent orientation-invariance inherent within networks might allow for further gains of OoD generalization. Biological agents may overcome the difficulties associated with recognizing OoD orientations by leveraging the temporal dimension to associate orientations and learn invariant representations Ruff (1982); Johnson & Aslin (1996); Ratan Murty & Arun (2015). The mechanisms that utilize temporal association may hold fundamental significance, given that they have access to a plentiful source of training

data that does not rely on external guidance and task specific labels. This data is readily available prior to any visual task and has the potential to contribute to the emergence of orientation-invariant representations beyond *in-plane* orientations.

Previous studies have extensively compared the behavioural and electrophysiological aspects of brains and DNNs Yamins et al. (2014); Yamins & DiCarlo (2016). However, a direct comparison between these systems alone has limitations in providing insights into the underlying mechanisms of object recognition in DNNs. This is due to the possibility that while certain fundamental mechanisms may be shared across these systems, the manifestation of these fundamental mechanisms can differ at the behavioral and electrophysiological levels. Our study has provided compelling evidence of brain-like neural mechanisms in DNNs that facilitate object recognition in novel orientations, even though these mechanisms are manifested differently than in biological systems. For instance, while humans and primates can recognize objects in orientations that are not simply 2D rotations, this capability is not fully replicated in DNNs. Thus, we can conclude that the neural mechanisms that have been observed to govern recognition in biological systems largely apply to DNNs, albeit with distinct manifestations across these systems. It will be interesting to follow this line of investigation across biological and artificial systems to envision a general theory to explain emergent mechanisms in both brains and machines.

References

Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition*, pp. 4845–4854, 2019.

Wang Angtian, Adam Kortylewski, and Alan Yuille. Nemo: Neural mesh models of contrastive features for robust 3d pose estimation. In *Proc of the Int Conf on Learning Representations*, 2021.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, pp. 9448–9458, 2019.

MC Booth and Edmund T Rolls. View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral cortex (New York, NY: 1991)*, 8(6):510–523, 1998.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016.

Taco S. Cohen and Max Welling. Steerable CNNs. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=rJQKYt5ll`.

Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical CNNs. In *Proc of the Int Conf on Learning Representations*, 2018.

Robert Desimone, Thomas D Albright, Charles G Gross, and Charles Bruce. Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, 4(8):2051–2062, 1984.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Fang Fang and Sheng He. Viewer-centered object representation in the human visual system revealed by viewpoint aftereffects. *Neuron*, 45(5):793–800, 2005.

Winrich A Freiwald and Doris Y Tsao. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, 330(6005):845–851, 2010.

Isabel Gauthier, William G Hayward, Michael J Tarr, Adam W Anderson, Pawel Skudlarski, and John C Gore. Bold activity during mental rotation and viewpoint-dependent object recognition. *Neuron*, 34(1): 161–171, 2002.

Michael S Gazzaniga, Richard B Ivry, and GR Mangun. Cognitive neuroscience. the biology of the mind,(2014), 2006.

Herbert Goldstein, Charles Poole, and John Safko. *Classical mechanics*. Addison-Wesley, 3rd edition, 2002.

Nate Gruver, Marc Anton Finzi, Micah Goldblum, and Andrew Gordon Wilson. The lie derivative for measuring learned equivariance. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JL7Va5Vy15J.

Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. We used the following implementation in our experiments: https://pytorch.org/vision/stable/models.html#torchvision.models.resnet18.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2961–2969, 2017.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017. We used the following implementation in our experiments: https://pytorch.org/vision/stable/models.html#torchvision.models.densenet121.

Hojin Jang, Syed Suleman Abbas Zaidi, Xavier Boix, Neeraj Prasad, Sharon Gilad-Gutnick, Shlomit Ben-Ami, and Pawan Sinha. Robustness to transformations across categories: Is robustness driven by invariant neural representations? *Neural Computation*, 35(12):1910–1937, 2023.

Scott P Johnson and Richard N Aslin. Perception of object unity in young infants: The roles of motion, depth, and orientation. *Cognitive Development*, 11(2):161–180, 1996.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint, arXiv:1412.6980, 2017.

Eucaly Kobatake and Keiji Tanaka. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of neurophysiology*, 71(3):856–867, 1994.

Jonas Kubilius, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel L. K. Yamins, and James J. DiCarlo. CORnet: Modeling the neural mechanisms of core object recognition. bioRxiv preprint, 408385, 2018. We used the following implementation in our experiments: https://github.com/dicarlolab/CORnet.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 991–999, 2015.

Bin Li, Yuki Todo, and Zheng Tang. The mechanism of orientation detection based on local orientation-selective neuron. In *2021 6th International Conference on Computational Intelligence and Applications (ICCIA)*, pp. 195–199. IEEE, 2021.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Nikos K Logothetis and Jon Pauls. Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cerebral Cortex*, 5(3):270–288, 1995.

Nikos K Logothetis and David L Sheinberg. Visual object recognition. *Annual review of neuroscience*, 19 (1):577–621, 1996.

Spandan Madan, Timothy Henry, Jamell Dozier, Helen Ho, Nishchal Bhandari, Tomotake Sasaki, Frédo Durand, Hanspeter Pfister, and Xavier Boix. When and how convolutional neural networks generalize to out-of-distribution category–viewpoint combinations. *Nature Machine Intelligence*, 4(2):146 – 153, 2022.

Nicolas Pinto, David D Cox, and James J DiCarlo. Why is real-world visual object recognition hard? *PLoS computational biology*, 4(1):e27, 2008.

Tomaso Poggio and Fabio Anselmi. *Visual cortex and deep networks: learning invariant representations*. MIT Press, 2016.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

N Apurva Ratan Murty and Sripati P Arun. Dynamics of 3d view invariance in monkey inferotemporal cortex. *Journal of Neurophysiology*, 113(7):2180–2194, 2015.

Holly A Ruff. Effect of object movement on infants' decision of object structure. *Developmental Psychology*, 18(3):462, 1982.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Roger N. Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972): 701–703, 1971.

Pawan Sinha and Tomaso Poggio. Role of learning in three-dimensional form perception. *Nature*, 384(6608): 460–463, 1996.

Shimon Ullman. *High-level vision: Object recognition and visual cognition*. MIT Press, 1996.

Shimon Ullman. Using neuroscience to develop artificial intelligence. *Science*, 363(6428):692–693, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 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.

Supporting Information

- S1: Accuracy heatmaps: alternative 'seed' orientations.
- S2: Accuracy heatmaps: effect of data diversity - alternative object categories.
- S3: Accuracy heatmaps: alterative training conditions - pretraining and augmentation.
- S4: Accuracy heatmaps: alternative backbone architectures.
- S5: Modeling generalization patterns for OoD orientations, continued.
- S6: OoD accuracy, split between *generalizable* and *non-generalizable* orientations
- S7: Invariance and Dissemination: controls.
- S8: tSNE analysis on the penultimate layer of a representative experiment.

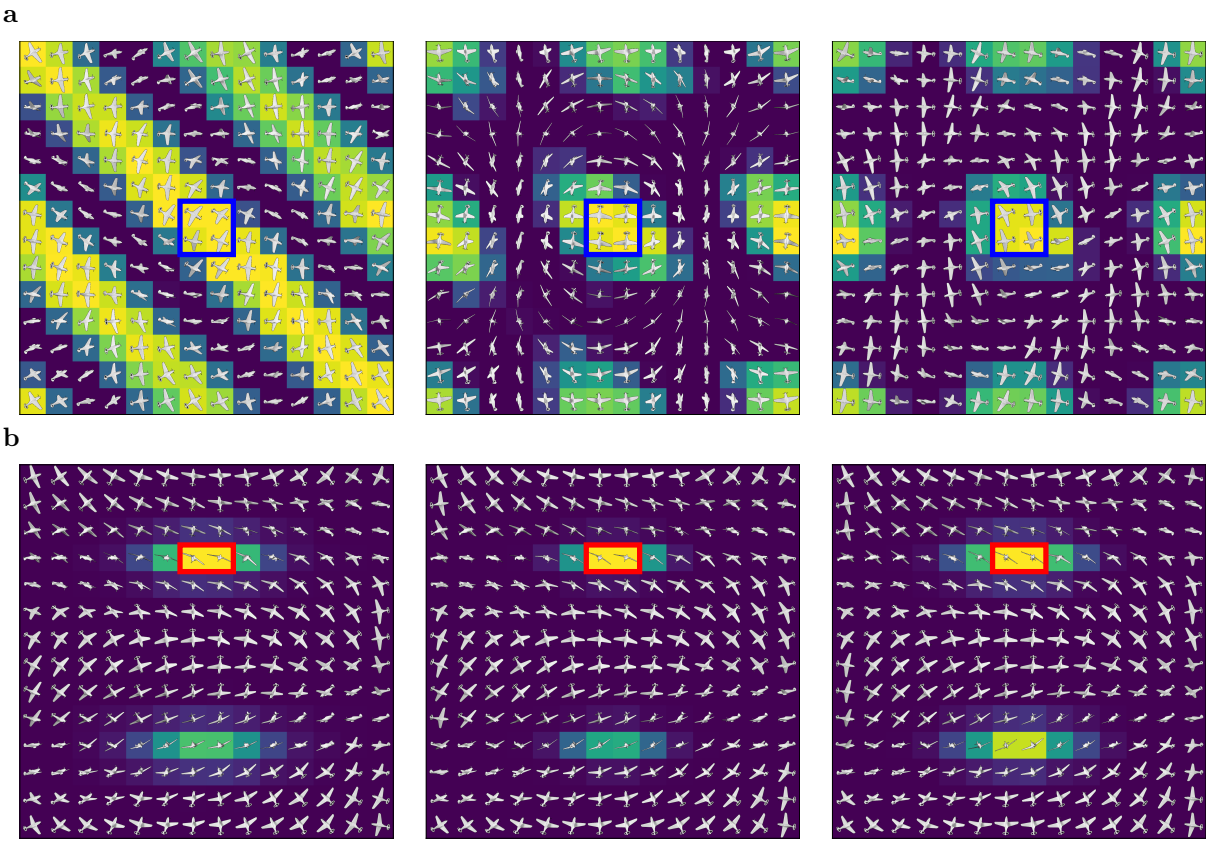

Figure S1: **Accuracy heatmaps: alternative 'seed' orientations.** (**a**) 'Seed' (*in-distribution*) orientations include $-0.25 \leq \alpha \leq 0.25, -0.25 \leq \gamma \leq 0.25, -1/2\pi \leq \beta < 1/2\pi$. (**b**) 'Seed' orientations include $-0.1 \leq \beta \leq 0.1, -0.25 \leq \gamma \leq 0.25, -1.8\pi \leq \alpha < -1.3/\pi$.

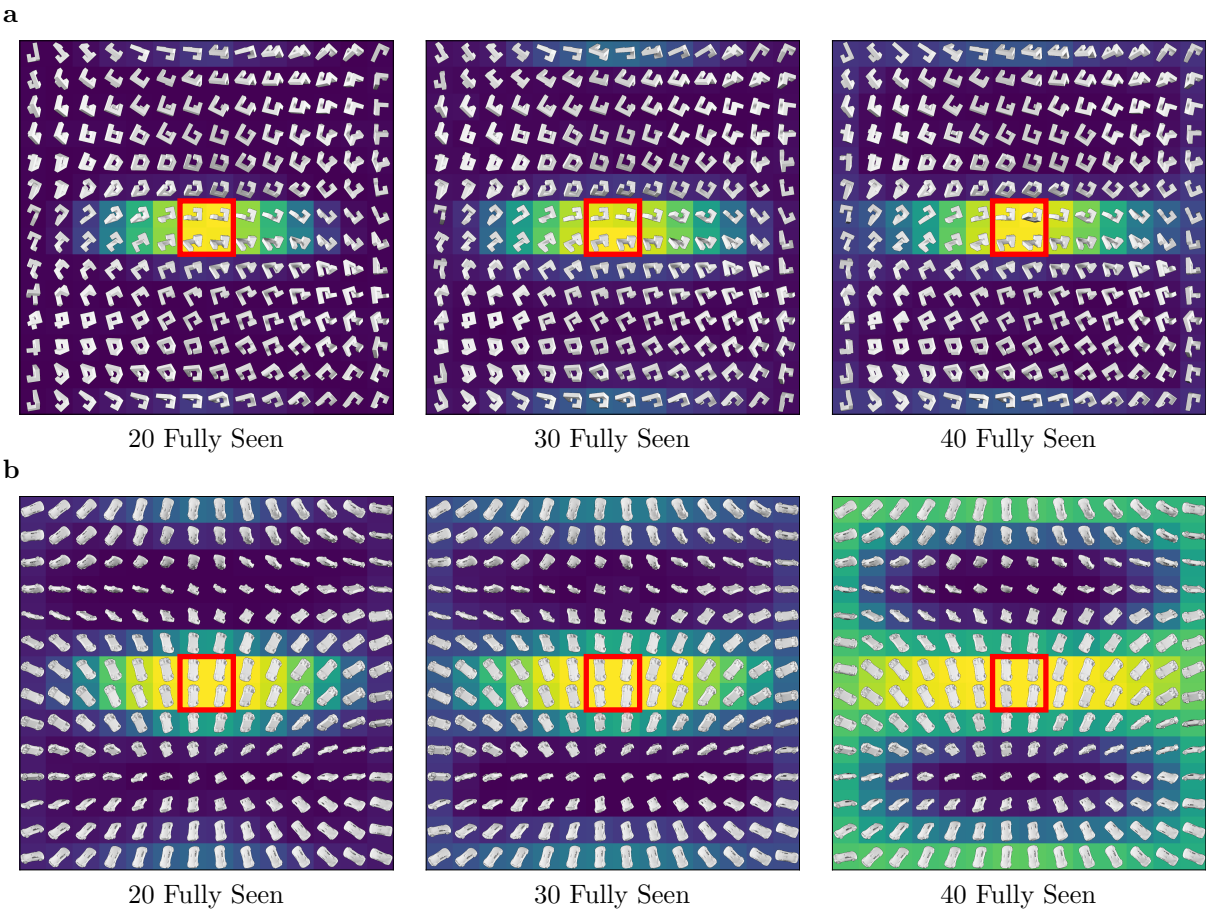

Figure S2: **Accuracy heatmaps: effect of data diversity - alternative object categories.** Increasing number of *fully-seen* instances, with different object classes. (**a**) Shepard-Metzler Objects. (**b**) Cars.

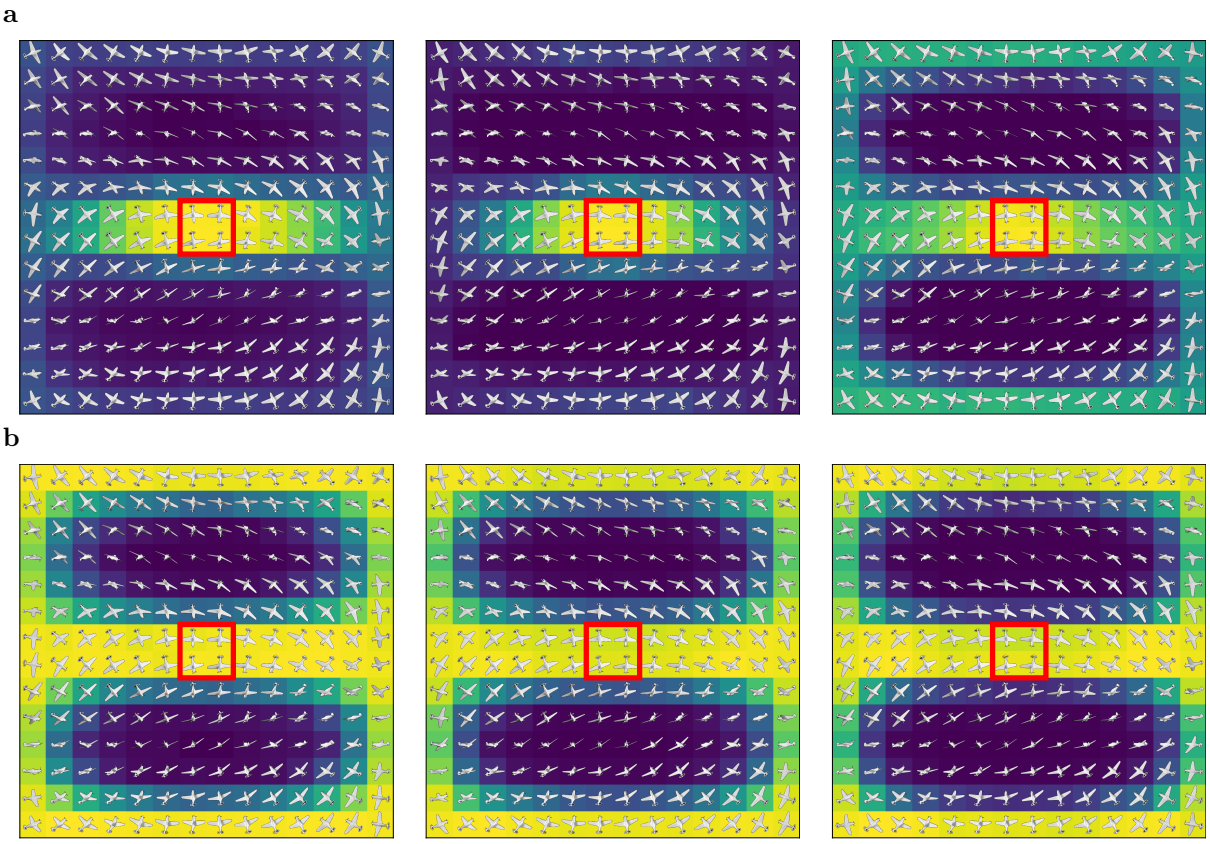

Figure S3: **Accuracy heatmaps: alterative training conditions - pretraining and augmentation.** (**a**) ResNet-18 pretrained on ImageNet Russakovsky et al. (2015), finetuned on our learning paradigm with airplanes. Network behavior isn't meaingfully altered. (**b**) All data (both from *fully-seen* and *partially-seen* instances) were augmented with random 2D image rotations. This effectively expands the *in-distribution* set to include all *generalizable* orientations. This results in *generalizable* orientations with high accuracy.

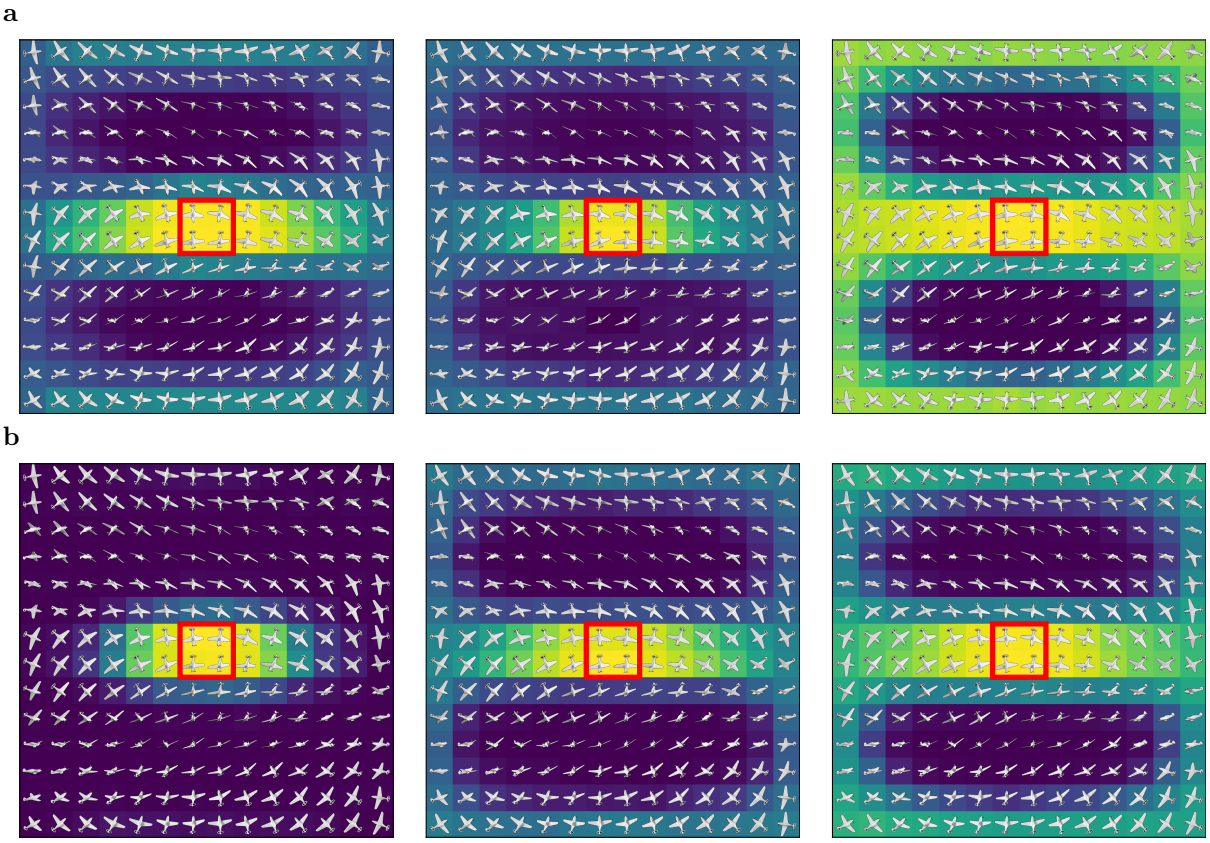

Figure S4: **Accuracy heatmaps: alternative backbone architectures.** Network's backbone used (in place of ResNet-18): (**a**) DenseNet. (**b**) CORnet.

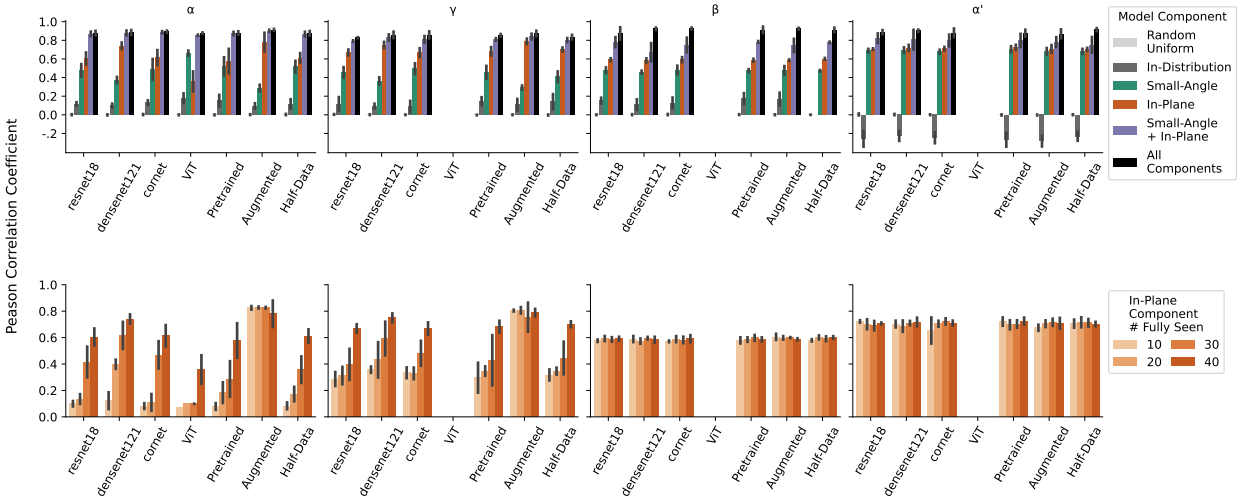

Figure S5: **Modeling generalization patterns for OoD orientations, continued.** The same analysis as Figs. 3b is applied to the controls introduced in Figs. S3, S4.

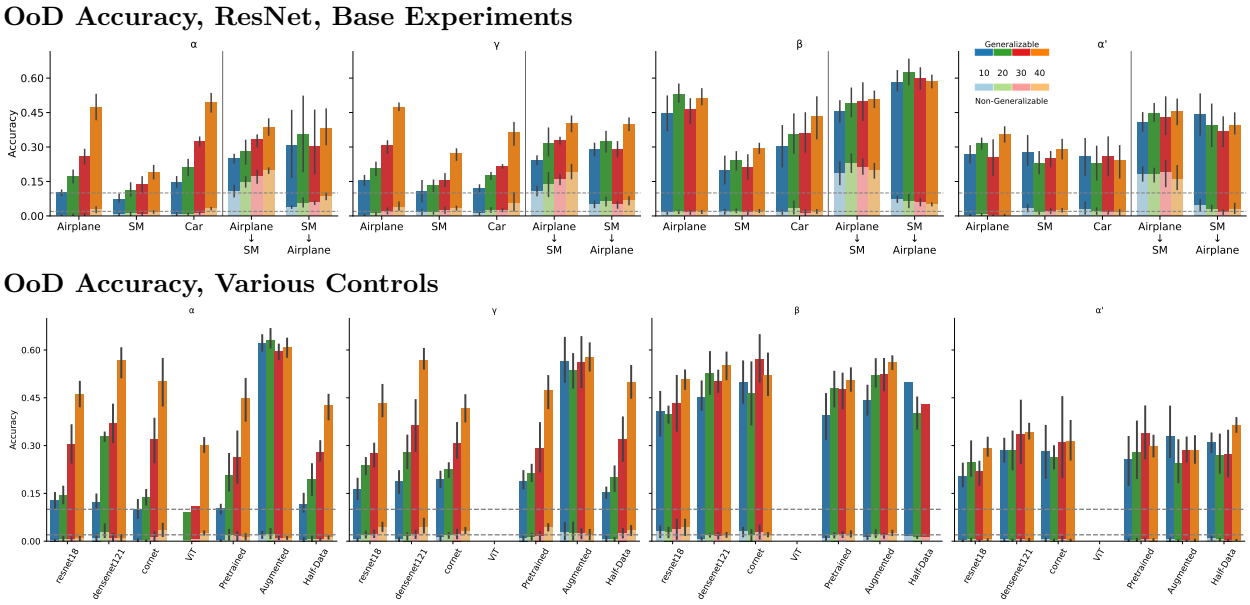

Figure S6: **OoD accuracy, split between *generalizable* and *non-generalizable* orientations.** In Fig. 3a we report the average accuracy across all OoD orientations. As we note, however, accuracy behavior is differentiated between *generalizable* and *non-generalizable* orientations. Here we report the average accuracy for these two orientation groups. Gray horizontal lines indicate chance performance of 2% and 10% (the latter relevant in the case where *fully-seen* and *partially-seen* instances are of two different classes.) *Generalizable* accuracy is always greater than *non-generalizable* accuracy. The former is always well above chance, while the latter is below or at chance level. **(a)** The *generalizable* and *non-generalizable* average accuracy for the same set of experiments presented in Fig. 3a. **(b)** The average accuracies for several other conditions. These other conditions are explained in Figs. S3, S4.

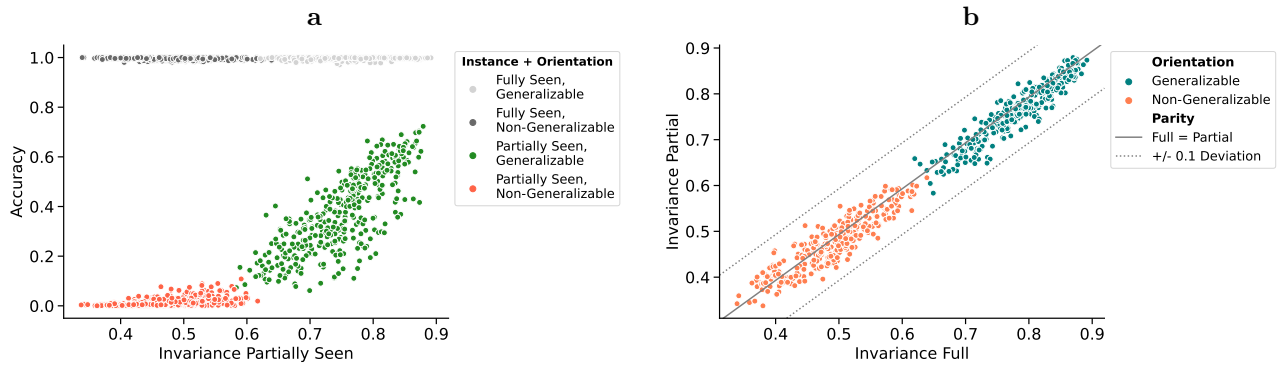

Figure S7: **Invariance and Dissemination: controls.** The same analysis as Figs. 4b, c is applied to the controls introduced in Figs. S3, S4.

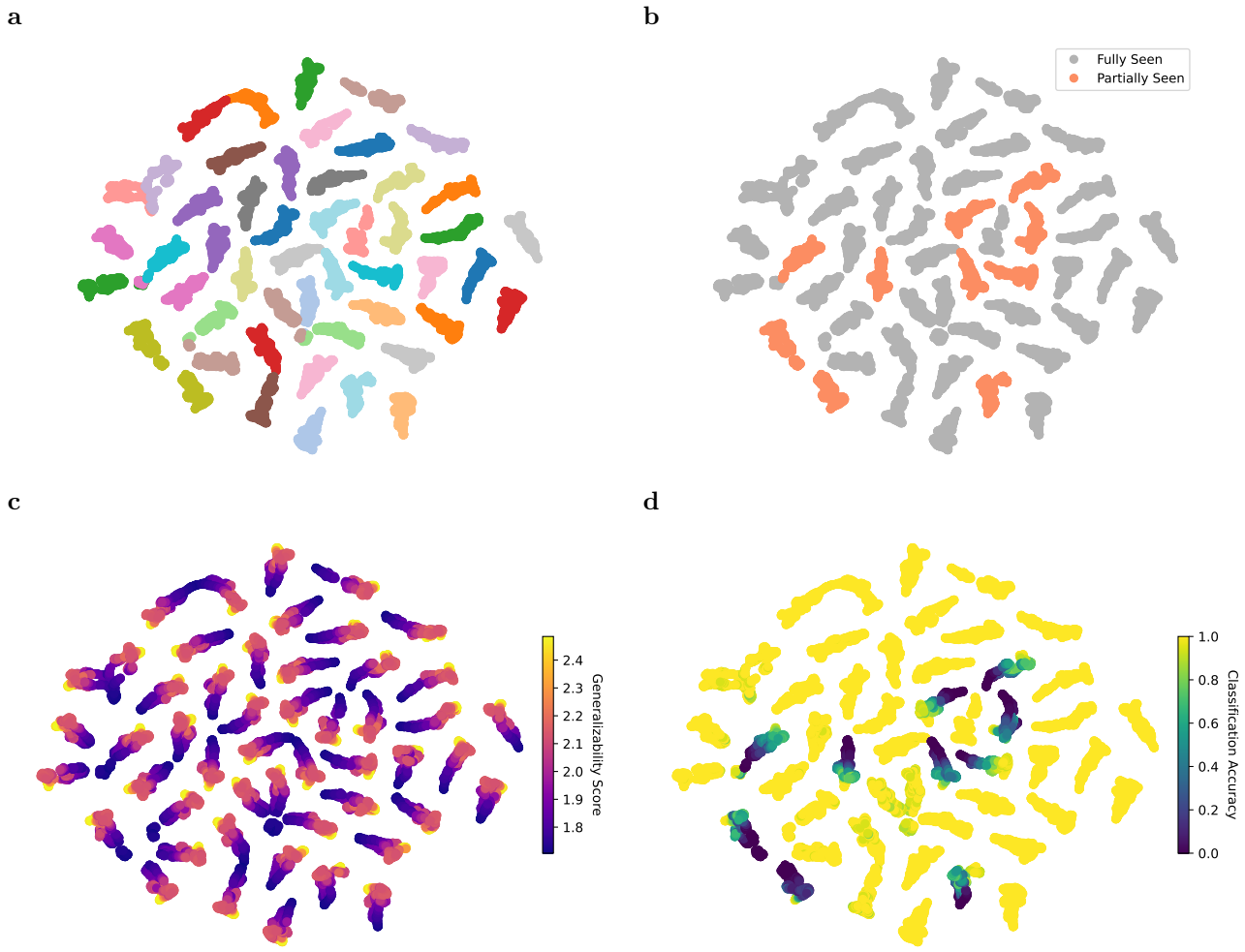

Figure S8: **tSNE analysis on the penultimate layer of a representative experiment.** The 512 dimension activation vectors in the penultimate layer for each instance and orientation are recorded. We employ tSNE to reduce these 512 dimensions down to two dimensions. **(a)** Object instances are colored (semi-) uniquely. (20 colors are distributed to 50 instances due to the limits of choosing many perceptually different colors.) For the most part, instances cluster together without much overlap between clusters of different instances. This indicates that representations of instances are separable, and that the task is solved by DNN. **(b)** Each point is colored based on whether the instance it represents is *fully-seen* or *partially-seen*. *Partially-seen* clusters are independent of other clusters, both *fully-seen* and *partially-seen*. It is therefore difficult to determine the range of behaviors for *partially-seen* instances — namely, why certain OoD orientations are *generalizable*, while others are not. **(c)** Points are colored with the degree of generalizabilty, as predicted by the predictive model of DNN generalization behavior. Note that points within each cluster are ordered — they are arranged such that *generalizable* orientations are far from *non-generalizable* orientations with a smooth transition between them. **(d)** Points are colored with the classification accuracy of the network (for the given instance and orientation.) While *fully-seen* instances have near 100% accuracy across all orientations, *partially-seen* show differentiation in accuracy between *generalizable* and *non-generalizable* orientations.