To Which Out-Of-Distribution Object Orientations Are DNNs Capable of Generalizing?

Avi Cooper[⋆,1,5,6], **Xavier Boix**[⋆,1,6], **Daniel Harari**[1,2,6], **Spandan Madan**[3,6], **Hanspeter Pfister**[3], **Tomotake Sasaki**[4,6], **Pawan Sinha**[1]

⋆ *equal first authorship*

[1] Massachusetts Institute of Technology [2] Weizmann Institute of Science [3] Harvard University
[4] Fujitsu Limited [5] Yale University [6] Center for Brains, Minds and Machines

{avic,xboix,psinha}@mit.edu, hararid@weizmann.ac.il,{spandan_madan,pfister}@seas.harvard.edu, tomotake.sasaki@fujitsu.com

Abstract

The capability of Deep Neural Networks (DNNs) to recognize objects in orientations outside the distribution of the training data, *ie.,* out-of-distribution (OoD) orientations, is not well understood. For humans, behavioral studies showed that recognition accuracy varies across OoD orientations, where generalization is much better for some orientations than for others. In contrast, for DNNs, it remains unknown how generalization abilities are distributed among OoD orientations. In this paper, we investigate the limitations of DNNs' generalization capacities by systematically inspecting patterns of success and failure of DNNs across OoD orientations. We use an intuitive and controlled, yet challenging learning paradigm, in which some instances of an object category are seen at only a few geometrically restricted orientations, while other instances are seen at all orientations. The effect of data diversity is also investigated by increasing the number of instances seen at all orientations in the training set. We present a comprehensive analysis of DNNs' generalization abilities and limitations for representative architectures (ResNet, Inception, DenseNet and CORnet). Our results reveal an intriguing pattern—DNNs are only capable of generalizing to instances of objects that appear like 2D, *ie.,* in-plane, rotations of in-distribution orientations.

Keywords: Evaluation and Analysis; Deep Neural Networks; Object Recognition; Out-of-Distribution Generalization

1 Introduction

The orientation of an object with respect to the viewer is a key factor that impacts image structure. An intelligent viewer should be able to recognize objects across a variety of orientations. During training, the orientations included in the datasets define the entire distribution available for learning systems. However, these datasets may lack some orientations from the full, true distribution of all orientations—they may be biased towards certain object orientations due to conventions and constraints during the data collection process (Torralba and Efros 2011). Thus, learning systems trained with these datasets may never have encountered examples of those "out-of-distribution" (OoD) orientations.

It has been shown that while Deep Neural Networks (DNNs) achieve high *in-distribution* test accuracy, their accuracy substantially degrades when tested with objects at OoD orientations, even when learning from large datasets with millions of examples (Barbu et al. 2019; Alcorn et al. 2019). Efforts to address OoD orientations leverage preconceived components for DNNs, such as using 3D models of objects (Angtian, Kortylewski, and Yuille 2021) or sophisticated sensing approaches such as omnidirectional imaging (Cohen et al. 2018). However, understanding the role of image-based learning in recognizing objects in OoD orientations has received less attention. It remains as an outstanding question at the heart of artificial and biological intelligence (Sinha and Poggio 1996; Ullman 1996; Poggio and Anselmi 2016). For state-of-the-art DNNs, little is known beyond the observed accuracy reduction for OoD orientations (Madan et al. 2020).

A potentially useful strategy for investigating the behavior of DNNs is to employ evaluation metrics that provide more details than just the conventional average accuracy number (Hoiem, Chodpathumwan, and Dai 2012). For humans and other primates, studies showed that recognition accuracy varies across OoD orientations, where generalization is much better for some orientations than for others (Logothetis and Pauls 1995). Are DNNs more likely to fail in recognizing an object at some OoD orientations than others? If so, which are these OoD orientations, and are they consistent across different characteristics of the dataset and DNN architectures?

In this paper, we answer these questions by analyzing patterns of success and failure of DNNs across a range of individual OoD orientations. To this end, we build upon the paradigm introduced by Zaidi et al. (2020), in which some instances of an object category (*e.g.,* a 'Boeing 777 airliner' is an instance of 'airplane' category) are seen from any orientation during training (*ie., fully-seen* instances), while other instances are only seen in few orientations (*ie., restricted-seen* instances). This is a simple paradigm that 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 *restricted-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.

To foreshadow the results, we find a pattern of behavior that is consistent across a diverse set of object categories (airplanes, cars, lamps, Shepard-Metzler objects) and multiple network architectures. We find that DNNs are only capa-

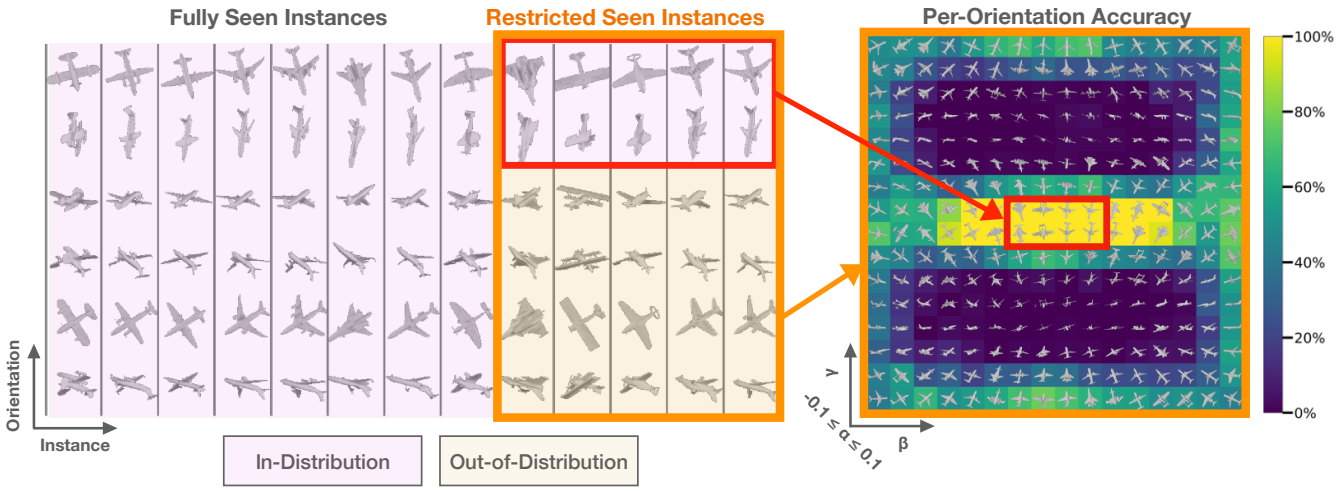


Figure 1: **Learning paradigm and network's per-orientation accuracy.** (**Left**) The learning paradigm employed in this work. Each column is a sample object instance (here from the airplane dataset) and each row is a sample orientation. The training set includes all orientations for *fully-seen* instances, and a restricted set of orientations (marked in red) for *restricted-seen* instances (in this example, with the airplane's nose pointing down). The orientations included in the training set are referred to as *in-distribution* (pink shading) orientations. Orientations of the *restricted-seen* instances that are not included in the training set are referred to as *out-of-distribution* or OoD (yellow shading). (**Right**) A visualization of the network's performance for the *restricted-seen* instances by means of per-orientation accuracy. Orientations outlined in red at the center are *in-distribution*, while the rest are OoD. The heatmap clearly shows the network's generalization to OoD orientations, including those which differ significantly from the *in-distribution* orientations, but appear as their 'in-plane' rotations.

ble of generalizing to new orientations that can be approximated as *in-plane* 2D rotations of *in-distribution* orientations.

2 Generalization-Ready Learning Paradigm

In a natural setting, biological intelligent agents observe instances of an object category from multiple, diverse viewpoints. When agents observe a novel object instance, they are often able to correctly classify the object, leveraging past experience with other instances of the same or similar category seen at same or similar orientations. Inspired by the setting of biological agents, we introduce a learning paradigm that facilitates analyzing the capabilities of DNNs to generalize across object instances at OoD orientations.

2.1 *Fully-seen* and *restricted-seen* instances

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, Poole, and Safko 2002), to express an orientation of an object instance. We use the convention that α and γ are bounded within 2π radians, and β is bounded within π radians. In our paradigm, the following two sets of instances are included in the training set:

– ***Fully-seen* instances** are object instances whose orientations are unrestricted (we employ uniform sampling along the three axes of rotations to generate the training set).

– ***Restricted-seen* instances** are biased towards a sub-range of object orientations during training. Specifically, one axis is allowed to freely rotate, *ie.,* rotate along its full range, and the other two axes are restricted to within a sub-range of a defined orientation. Evaluation of OoD generalization is measured using the networks' instance classification performance on OoD orientations of *restricted-seen* instances.

Figure 1 portrays our learning paradigm, including an illustration (on the left) of the *fully-seen* and *restricted-seen* sets of instances for the 'airplane' category.

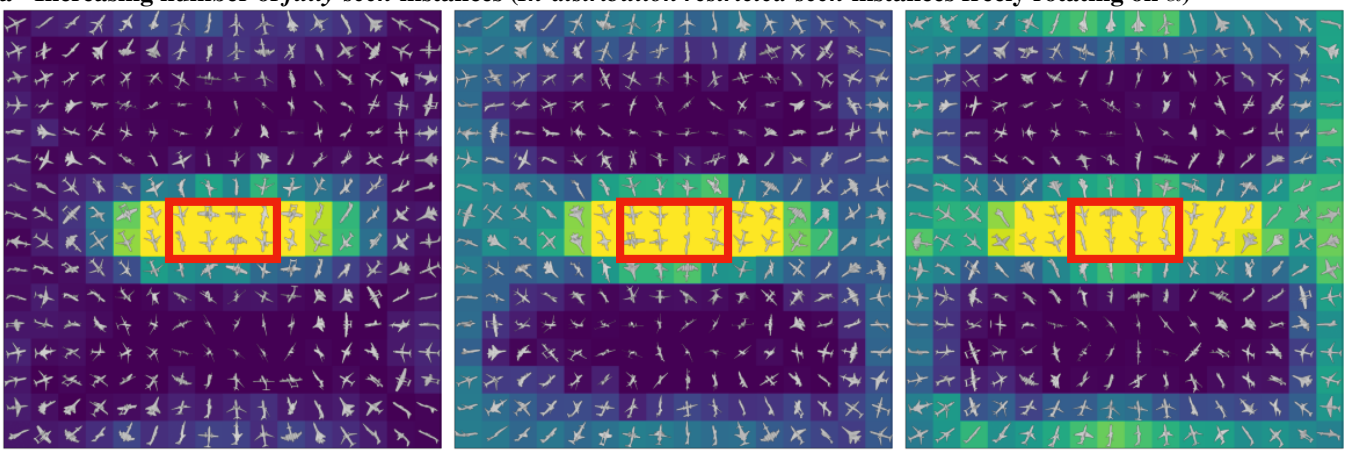
2.2 Data diversity

In order to better understand how the proportion of *fully-seen* compared to *restricted-seen* instances affects the network's generalization performance, we vary the proportion in the following two ways:

– **Proportion of *fully-seen* instances.** We vary diversity in terms of the number of *fully-seen* instances N between 20% and 80% of the total number of instances. The remaining instances are *restricted-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. Also, we always test with the same 20% of instances that are never part of the *fully-seen* set.

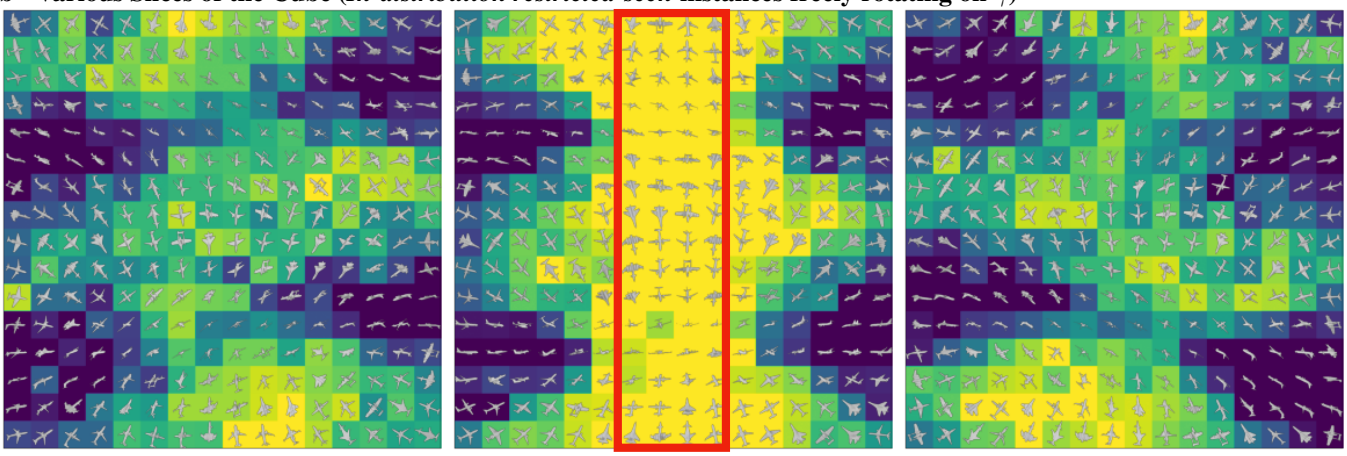
– **Orientations seen during training.** We vary diversity in terms of orientations by means of setting the ranges of orientations sampled for the *restricted-seen* instances. We do this in two ways: 1) choosing different axes to be freely rotating, which results in different types of orientations in the *in-distribution* set, 2) choosing different range-centers for the other two axes. For example, when rotating on α, sampling γ from $[-0.25, 0.25]$ or $[\pi/2-0.25, \pi/2+0.25]$ or the union of the two ranges (see Fig. 4a).

a - Increasing number of *fully-seen* instances (*In-distribution restricted-seen* instances freely rotating on α)



Fully Seen Instances: 20 / 50 | # Fully Seen Instances: 30 / 50 | # Fully Seen Instances: 40 / 50

b - Various Slices of the Cube (*In-distribution restricted-seen* instances freely rotating on γ)



$-2.75 \leq \alpha \leq -2.36$ | $0 \leq \alpha \leq 0.29$ | $2.36 \leq \alpha \leq 2.749$

Figure 2: **Observed generalization in accuracy patterns.** The heatmaps show per-orientation accuracy for test samples included in a slice of our visualization cube. The dimensions of a full cube correspond to the Euler angles of the rotation α, β, γ. Outlined in red is the range of orientations that are *in-distribution*. Orientations outside the red box are *out-of-distribution* (OoD). **(a)** Increased network generalization for OoD orientations, with increasing instance diversity (left to right). Each heatmap is a slice at $\alpha \approx 0$ from three different cubes, each with the indicated number N of *fully-seen* instances. **(b)** The generalization patterns for a different span of *in-distribution* orientations ($-0.25 \leq \alpha \leq 0.25, -0.25 \leq \beta \leq 0.25, -\pi \leq \gamma < \pi$) as outlined by the red box in the middle slice. Each heatmap is from the same cube at the indicated α value (see Fig. 4a).

3 Per-Orientation Accuracy Visualization

We introduce a systematic way to develop an understanding of DNN behavior in response to OoD examples. The following visualization approach provides rich insights and brings forward hypotheses about DNN generalization behavior.

3.1 Formulation of the visualization

Previous works typically report average performance over all orientations (Hoiem, Chodpathumwan, and Dai 2012). In contrast, we evaluate the network's performance for each orientation across the entire range of orientations. 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 the *restricted-seen* instances.

To facilitate intuition of Ψ we introduce a visual representation for this function. Orientations are continuous values and are related spatially. We leverage this property in the following way: We map the range of bounded values of orientations (α, β, γ) onto a Cartesian coordinate, 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 1/20 of the full range of each respective angle (0.1π radians for α and γ; 0.05π radians for β). We choose this size because the width is sufficiently narrow to preserve local behavior in aggregated analysis, while also being wide enough to include sufficiently large number of random samples from our dataset in almost every *cubelet*.

We evaluate our visualization cube only for the *restricted-seen* instances, where we average the classification accuracy across each *cubelet* to obtain *per-orientation* performance. Figure 4a shows this visual representation scheme.

3.2 Observed generalization in accuracy patterns

Our *per-orientation* accuracy visualization cube reveals interesting patterns, which reflect the network's behavior at different orientations. We show slices of the full cube for easy viewing. Each square slice is an accuracy heatmap extracted from the cube at a given angle for one of the cube dimensions. See Figs. 1 and 2. We outline in red the range of *in-distribution* orientations, 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*. Across a series of experiments detailed in the sequel, we observe accuracy patterns that indicate two main aspects of the network's generalization capabilities:

Generalization increases with instance diversity. As the number of *fully-seen* instances is increased in the training set, the network overall performance on classifying *restricted-seen* instances in OoD orientations markedly improves. Furthermore, the visualization heatmap reveals a pattern, which develops for this increase in performance and it occurs for some OoD orientations, but not for others (see Fig. 2a and Appendix A). A pattern of generalization clearly develops, becoming more defined and pronounced as the number of *fully-seen* instances is increased.

Generalization to *in-plane* orientations. Visual inspection of the accuracy heatmaps allows for the recognition of a reproducible pattern, as well as for generating hypotheses about the behavior of generalization. We noticed generalization to orientations that appeared as 2D rotations of *in-distribution* orientations, in the plane normal to the camera viewing direction. We call these orientations *in-plane*. The overlaid object images on-top of the heatmaps in Fig. 2 depicts the network's generalization to orientations which are image transformations of the *in-distribution* set (outlined in red). Figs. 2a and 2b differ in their respective *in-distribution* orientations, which yield different patters of generalization at the respective sets of *in-plane* orientations (see more examples in Appendix A).

This result is a surprising twist to the generally accepted view. We note that the general view is that networks are capable of generalizing and therefore performing well only for images fairly indistinguishable from *in-distribution* images. This would imply in our visualization, high accuracy only for orientations inside and possibly adjacent to *in-distribution* orientations (indicated in red in Fig. 2). Nevertheless, our experimentation shows that networks are capable to generalize to many other orientations beyond those expected by the general view. We conclude that an alternative hypothesis for networks generalization capabilities is necessary—one that includes *in-plane* orientations.

4 Model for Per-Orientation Generalization

In this section, we introduce a quantitative predictive model for the generalization behaviours of DNNs, which we denote as $f_{\mathbf{w}}(\boldsymbol{\theta})$. We evaluate this model by measuring the Pearson correlation between the performance of the networks in our experiment and our model, *ie.,* $\rho(\Psi(\boldsymbol{\theta}), f_{\mathbf{w}}(\boldsymbol{\theta}))$. We choose the Pearson correlation as a 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 $\boldsymbol{\theta}$.

We base our model $f_{\mathbf{w}}(\boldsymbol{\theta})$ on the hypothesis we derived in Sec. 3.2, *ie.,* DNNs are capable of generalizing to orientations which are visually similar to the *in-distribution* images and to orientations that are *in-plane* relative to the *in-distribution* images. These two components easily lend themselves to formalization with Euler's rotation theorem (Goldstein, Poole, and Safko 2002). The theorem states that any rotation can be uniquely described by a single axis, which can be represented by a unit vector $\hat{\mathbf{e}} \in \mathbb{R}^3$, and an angle, which is denoted as $\phi \in [0, \pi]$ and represents the amount of rotation around axis $\hat{\mathbf{e}}$. We compute $\hat{\mathbf{e}}$ and ϕ for the rotation from an arbitrary orientation $\boldsymbol{\theta}_s$ in the set of *in-distribution* orientations of the *restricted-seen* instance, denoted by Ω_s, to the orientation of interest $\boldsymbol{\theta}$. 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 visually similar to those 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 our 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 $\hat{\mathbf{e}}_{\boldsymbol{\theta},\boldsymbol{\theta}_s} \in \mathbb{R}^3$ (*ie.,* 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 $\mathbf{c}$.

Definition of the Predictive Model. Following the definitions above, we finally define $f_{\mathbf{w}}(\boldsymbol{\theta})$, the predictive model for generalization per each orientation. The model combines $A(\boldsymbol{\theta})$ and $E(\boldsymbol{\theta})$ by taking the maximum of their respective

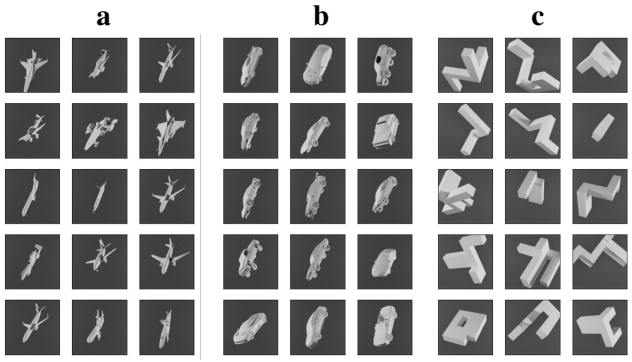


Figure 3: **Object Data Sets.** In our experiments we used three object categories: (**a**) Airplanes, (**b**) Cars, (**c**) Shepard&Metzler (the first two curated from ShapeNet (Chang et al. 2015); the last generated by ourselves). There are 50 instances per object category (*e.g.,* 'Concorde' or 'Spitfire' for the Airplanes). Images were rendered from the 3D models under fixed lighting conditions, but with orientations varied. Objects 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 $(-\pi \leq \alpha < \pi, -\frac{\pi}{2} \leq \beta < \frac{\pi}{2}, -\pi \leq \gamma < \pi)$.

values (and we therefore also denote it as AE). We choose to employ the logistic function in order to better match experimental results:

$$f_{\mathbf{w}}(\boldsymbol{\theta}) := \max\{\sigma(A(\boldsymbol{\theta}); \mathbf{w}_A),\ \sigma(E(\boldsymbol{\theta}); \mathbf{w}_E)\} \in [0, 1], \quad (3)$$

where $\sigma(x; (a, b)) = 1/(1 + \exp(a(-x + b)))$. Note that σ has two parameters, (a, b), and these are in $[0, \infty)$. Thus, $\mathbf{w}$ represents the parameters of the two logistic functions in equation 3, *ie.,* $\mathbf{w} = (\mathbf{w}_A, \mathbf{w}_E) \in [0, \infty)^2 \times [0, \infty)^2$.

The 'S'-like shape of the logistic function allows for the highest and lowest values of $\sigma(E(\boldsymbol{\theta}); \mathbf{w}_E)$ and $\sigma(A(\boldsymbol{\theta}); \mathbf{w}_A)$ 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, ensuring that $f_{\mathbf{w}}(\boldsymbol{\theta})$ remains a model related to small angle and *in-plane* rotations.

Finally, we fit $\mathbf{w}$ with a grid search, finding the optimal parameters which maximize the score on our evaluation metric, $\rho(\Psi(\boldsymbol{\theta}), f_{\mathbf{w}}(\boldsymbol{\theta}))$. The range of values for each of $\mathbf{w}_A$ and $\mathbf{w}_E$ is $\{0.1, 0.2, 0.3, \ldots, 1.5\} \times \{1, 2, 3, \ldots, 30\}$. The outputs of E are saturated and therefore our application of σ does not work well. We apply an exponential function, *ie.,* $E = E(\boldsymbol{\theta})^k$, which spreads out its values (we choose $k = 20$).

5 Experimental Setup

Datasets. We used ShapeNet (Chang et al. 2015) airplanes and cars. We curated 50 high quality object instances for each of these categories. Both airplanes and cars have clear axes of symmetry, which allow for intuition of how networks generalize to OoD orientations. We also experimented with highly asymmetric objects similar to those tested for 3D mental rotations in (Shepard and Metzler 1971) (denoted here as *Shepard&Metzler* objects; Fig. 3).

DNN Architectures. We use four deep convolutional neural networks in our experiments, namely, ResNet18 (He et al. 2016), DenseNet (Huang et al. 2017), Inception (Szegedy et al. 2016) and CORnet (Kubilius et al. 2018). The first three 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.

More details on the experimental setup, including the dataset sizes, hyperparameters for training, and hardware information, are given in Appendix B.

6 Results

We now report the experimental results. For all experiments, we evaluated five different training distributions of orientations for the *restricted-seen* instances. These five distributions are depicted in Fig. 4 and results are displayed in each column of the plots. In the first column ("Rotating on β"), almost all the *in-plane* orientations are *in-distribution* (note that the airplane's nose rotates 180° in a concave arc). In the second column ("Rotating on γ"), the *in-distribution* set contains fewer, but still some, *in-plane* orientations, namely the nose pointing either down or up. In the third column ("Rotating on α"), only one *in-plane* orientation is *in-distribution*. Hence, the first three columns evaluate training distributions as the number of *in-plane* rotations included in the training is reduced. The fourth and fifth columns capture a similar trend with a distribution of orientations not centered around 0. For the fourth column ("Rotating on α ($\gamma \approx \pi/2$)"), almost all the *in-plane* orientations are *in-distribution*, which is analogous to the first column ("Rotating on β"). The fifth column is the union of the distribution of the third and the fourth columns.

In the following, we report results on ResNet (the rest of the networks are in Appendix C). In all figures, each row of plots indicates the object category. Results are depicted by the average across three trials (randomizing the object instances that are *fully-seen*) and the 95% confidence intervals.

Average Accuracy. Fig. 4b shows the average classification accuracy averaged across all orientations, to analyze the common networks behavior on our datasets. Results clearly show the drop in accuracy between the top performance for *in-distribution* orientations (marked with dashed lines) and the poor performance for OoD orientations (marked with solid lines). Also, we observe an increase of the OoD accuracy as the number of *fully-seen* instances, *ie.,* data diversity, is increased. Recall that the amount of training examples is kept constant to analyze the effect of data diversity (see Appendix C for results with varying amount of training examples). Our results are in accordance to previous works that already noted the importance of data diversity to facilitates OoD generalization, in contrast to the number of training examples (Zaidi et al. 2020; Madan et al. 2020). Fig. 4b also presents a comparison between using pretrained

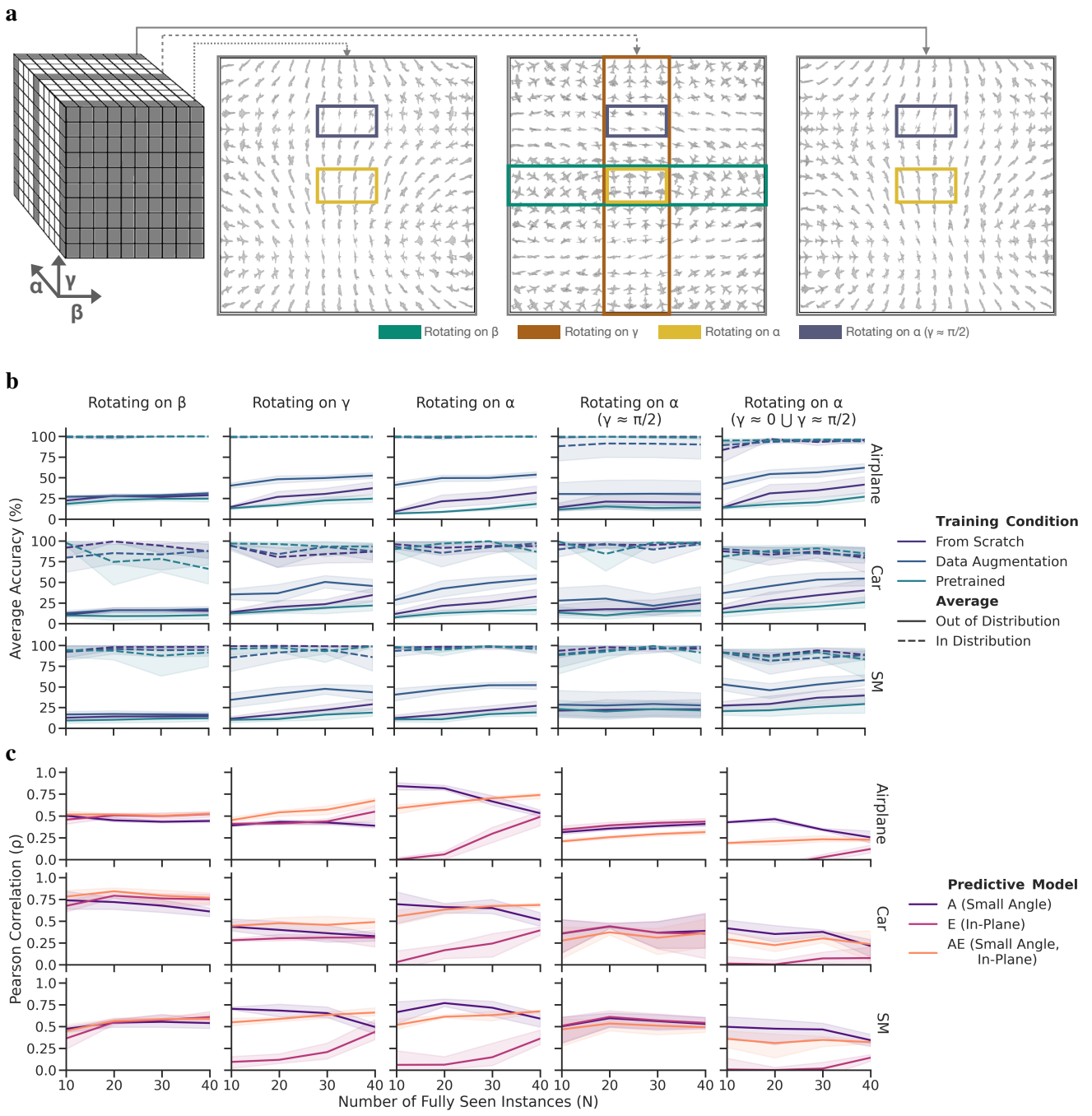


Figure 4: **Average accuracy and evaluation of the predictive accuracy model.** (**a**) A visual depiction of the orientations cube and sample slices illustrating the various *in-distribution* sets used in our experiments. (**b**) The average performance for *in-distribution* and OoD orientations. Columns refer to the different *in-distribution* sets. Experiments include: 1) training from scratch with random weights initialization, 2) training with data augmentation including 2D transformations, and 3) only fine-tuning networks pretrained on ImageNet. Each experiment was run three times with different *restricted-seen* instances and error bars indicate 95% confidence interval. Performance increases with more *fully-seen* instances, but always far lower for OoD orientations. (**c**) The Pearson correlation coefficient (ρ) between the measured network behavior and the different model components: A (small angle rotations), E (*in-plane* rotations), and AE (combination of A and E, *ie.,* the full model $f_{\mathbf{w}}(\boldsymbol{\theta})$, see Sec. 4).

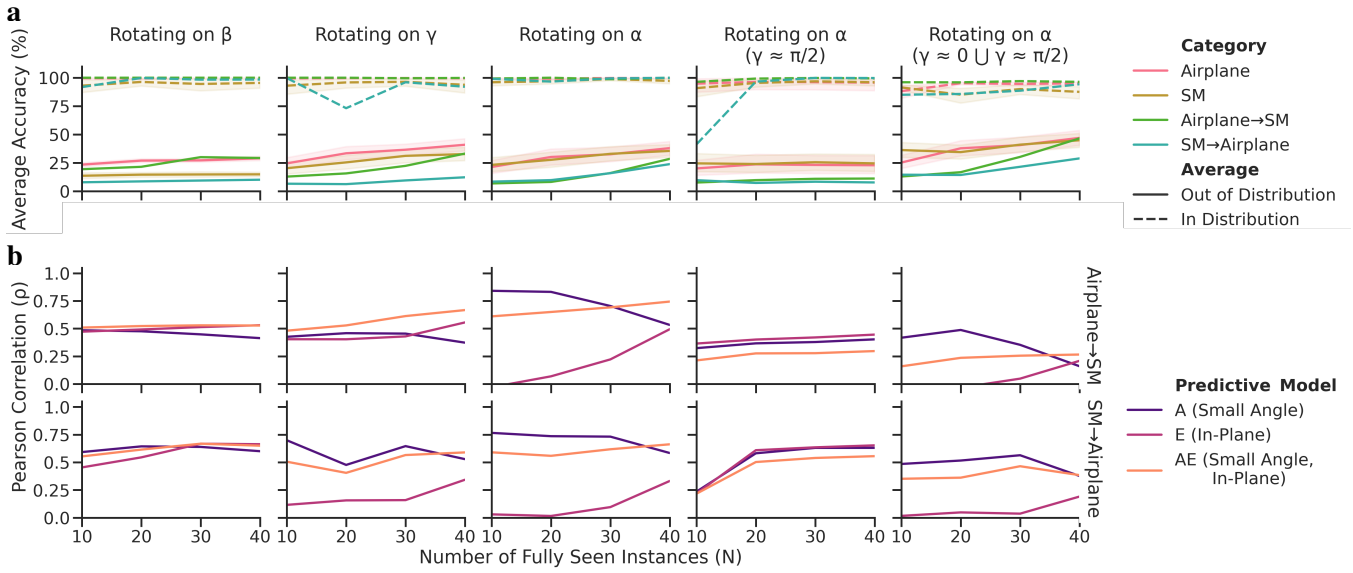


Figure 5: **Cross-Category Generalization.** Conventions follow Figs. 4b and 4c. Here *fully-seen* and *restricted-seen* instances are from different categories (airplanes and Shepard & Metzler objects) with very different geometries and symmetries. Trends closely align with those in Fig. 4, which indicates that DNNs are capable of generalization to OoD orientations of instances whose appearance after geometric transformations behaves quite differently than the *in-distribution* instances.

weights based on ImageNet, which yields the lowest accuracy, and training from scratch on our datasets, which yields higher accuracy. Including data augmentation during training (2D rotations and scaling) yields the highest accuracy, as expected, but there still exists a big gap between OoD and *in-distribution* accuracy.

Evaluating the Predictive Model. Recall we evaluate the predictive model with the Pearson correlation between the performance of the networks in our experiment and our model, *ie.,* $\rho(\Psi(\boldsymbol{\theta}), f_{\mathbf{w}}(\boldsymbol{\theta}))$. Results are presented in Fig. 4c. We evaluate the model and each of its two components (fitting the parameters for each of them from scratch). The first two columns and the fourth column in Fig. 4c, show that all model components highly correlate with the network behavior. In these cases, since many *in-plane* orientations are *in-distribution*, little generalization is necessary to perform well for all *in-plane* orientations, as predicted in the model by component E. For the third and fifth columns, the trends are more illustrative of how generalization to *in-plane* orientations emerges as we increase the number *fully-seen* instances, *ie.,* data diversity. Initially, for few *fully-seen* instances, A is the best predictor for OoD behavior and E is a poor predictor, *ie.,* the network generalizes to small variations of *in-distribution* samples. As the number of *fully-seen* instances is increased, A becomes less correlated with the network behaviour, and E more correlated. This follows the qualitative results discussed in Sec. 3.2, where as we increased the number of *fully-seen* instances, the networks gained generalization to *in-plane* rotations. The predictive model AE (*ie.,* the max of A and E, which is the full model $f_{\mathbf{w}}(\boldsymbol{\theta})$) is ultimately best correlated with the networks behavior, indicating that generalization occurs in regions of the orientations cube consisting of small-angle as well as *in-plane* rotations of the *in-distribution* orientations. Notably, these trends hold true across object categories, even though their symmetries differ, especially with the *Shepard & Metzler* objects. In Appendix C we demonstrate the generality of these results by experimenting with different network architectures, varying the amount of training examples, and using datasets with object at different scales and with objects that have stronger symmetries (such as lamps that tend to be solids of revolution).

Cross-Category Generalization. To further demonstrate the generality of our conclusions, we extend the previous results with a set of experiments that alter the paradigm slightly. Namely, *fully-seen* instances are from one object category and *restricted-seen* instances are from a different object category. Fig. 5 presents both the average accuracy and the predictive modeling for these experiments (in this case only one trial was performed). We specifically chose to compare cross-category generalization between airplanes and *Shepard& Metzler* objects, since their respective geometries and symmetries are quite different. The results are largely the same as those described above for single-category generalization, indicating that DNNs are capable not only of generalization to OoD orientations for similar instances, but even to instances whose appearance after geometric transformations behaves quite differently.

7 Conclusions and Future Works

We have demonstrated that increasing the number of *fully-seen* instances (while keeping the number of training examples constant), results in an increase of the DNN capability of recognizing instances of objects in OoD orientations that appear like 2D rotations (*in-plane*) of *in-distribution* orientations. Note that this result does not rule out the possibil-

ity that in experimental settings different from ours or with DNNs with training mechanisms yet to be discovered, there may also be generalization to orientations that are not *in-plane*. Nonetheless, our finding demonstrates that there are patterns in the failures of DNNs across orientations that can be clearly characterized.

A key question that is derived from our results and will be tackled in future works is to explain why DNNs generalize only to *in-plane* orientations. A hypothesis is that *restricted-seen* instances in orientations that are not *in-plane* may be too difficult or impossible to recognize (even for humans). Even though all object instances are distinguishable at all orientations, as indicated by the high *in-distribution* accuracy achieved by the DNNs, it is unclear whether there is sufficient information to recognize them when they are *restricted-seen* instances. Another hypothesis is that orientations that are not *in-plane* are affected by self-occlusion and DNNs may particularly suffer from it.

Other key open questions derived from our results are analyzing the DNN behaviour to more complex transformations than changes of orientation, including changes of illumination, texture and shape deformations. Also, we are intrigued about the neural mechanisms that facilitate recognizing object instances in OoD orientations. Some hints towards an answer were provided by Poggio and Anselmi (2016), as they demonstrated that neurons that are tuned to an object category and are invariant to a sequence or orbit of orientations, facilitate recognition of novel object instances. Current state-of-the-art network architectures were introduced by taking into account *in-distribution* generalization, and not OoD generalization. Thus, novel network architectures may be necessary to allow for further gains of OoD generalization, and in particular, architectures that facilitate the emergence of invariant representations may be of crucial importance.

Author Contributions

AC, XB, DH, SM designed research; AC performed experiments with contributions of XB; AC, XB, DH, SM and TS analyzed data; AC, XB, DH, SM and TS wrote the paper with contributions of PS; HP and PS supervised the research with contributions of XB, DH and TS.

Acknowledgments

We are grateful to Tomaso Poggio and Shimon Ullman for their insightful advice and warm encouragement. This work was supported by Fujitsu Limited (Contract No. 40008819 and 40009105) and by the Center for Brains, Minds and Machines (funded by NSF STC award CCF-1231216). PS and XB are supported by the R01EY020517 grant from the National Eye Institute (NIH), AC is supported by Fujitsu Research of America, Inc. as an intern and the Yale Class of 1960 Fellowship, and DH is supported by the Robin Chemers Neustein Artificial Intelligence Fellows Program.

Conflicts of Interests Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. Fujitsu Limited funded this study (Contract No. 40008819 and 40009105) and also participated in the study through AC and TS (AC was the first half of the study at MIT and the second half at Fujitsu Research of America). All authors declare no other competing interests.

Data and Code Availability Statement

The raw data and code supporting the conclusions of this article are publicly accessible upon request.

References

Alcorn, M. A.; Li, Q.; Gong, Z.; Wang, C.; Mai, L.; Ku, W.-S.; and Nguyen, A. 2019. 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*, 4845–4854.

Angtian, W.; Kortylewski, A.; and Yuille, A. 2021. NeMo: Neural Mesh Models of Contrastive Features for Robust 3D Pose Estimation. In *Proc of the Int Conf on Learning Representations*.

Barbu, A.; Mayo, D.; Alverio, J.; Luo, W.; Wang, C.; Gutfreund, D.; Tenenbaum, J.; and Katz, B. 2019. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, 9448–9458.

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; Xiao, J.; Yi, L.; and Yu, F. 2015. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago.

Cohen, T. S.; Geiger, M.; Köhler, J.; and Welling, M. 2018. Spherical CNNs. In *Proc of the Int Conf on Learning Representations*.

Goldstein, H.; Poole, C.; and Safko, J. 2002. *Classical mechanics*. Addison-Wesley, 3rd edition.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition*, 770–778.

Hoiem, D.; Chodpathumwan, Y.; and Dai, Q. 2012. Diagnosing error in object detectors. In *Proc of the European Conf on Computer Vision*, 340–353.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition*, 4700–4708.

Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Kubilius, J.; Schrimpf, M.; Nayebi, A.; Bear, D.; Yamins, D. L. K.; and DiCarlo, J. J. 2018. CORnet: Modeling the Neural Mechanisms of Core Object Recognition. bioRxiv:408385.

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

Madan, S.; Henry, T.; Ho, H.; Bhandari, N.; Sasaki, T.; Durand, F.; Pfister, H.; and Boix, X. 2020. On the Capability of Neural Networks to Generalize to Unseen Category-Pose Combinations. Technical Report CBMM Memo No. 111, Center for Brains, Minds and Machines.

Poggio, T.; and Anselmi, F. 2016. *Visual cortex and deep networks: learning invariant representations*. MIT Press.

Shepard, R. N.; and Metzler, J. 1971. Mental Rotation of Three-Dimensional Objects. *Science*, 171(3972): 701–703.

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

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the Inception architecture for computer vision. In *Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition*, 2818–2826.

Torralba, A.; and Efros, A. A. 2011. Unbiased look at dataset bias. In *Proc of the IEEE Conf on Computer Vision and Pattern Recognition*, 1521–1528.

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

Zaidi, S. S. A.; Boix, X.; Prasad, N.; Gilad-Gutnick, S.; Ben-Ami, S.; and Sinha, P. 2020. Is Robustness To Transformations Driven by Invariant Neural Representations? arXiv:2007.00112.

Appendix

A Additional Per-Orientation Accuracy Visualization (Section 3)

We include several more qualitative results in addition to those presented in Sec. 3. These provide further insights for the hypothesis that networks generalize to small angles and *in-plane* orientations. The first set shows similar results with different object categories (Fig. 6 and 7). The second shows similar generalization capabilities with varying *in-distribution* sets (Fig. 8, 9 and 10). The final set demonstrates the generalization capabilities when data augmentation is applied during training (Fig. 11), which effectively includes *in-plane* rotations of *restricted-seen* orientations in the training set, extending the *in-distribution* set. The following list provides a summary of all the additional per-orientation accuracy visualizations provided:

- Additional Datasets
  - Fig. 6: Car Dataset, Rotating on α
  - Fig. 7: *Shepard & Metzler* Dataset, Rotating on α
- Different *In-Distribution* Sets
  - Fig. 8: Airplane Dataset, Rotating on $\alpha, (\gamma \approx \pi/2)$
  - Fig. 9: Airplane Dataset, Rotating on $\alpha, (\gamma \approx \pi/2 \cup \gamma \approx 0)$
  - Fig. 10: Airplane Dataset, Rotating on β
- Data Augmentation During Training
  - Fig. 11: Airplane Dataset, Rotating on α, Data Augmentation During Training

B Experiment Details (Section 5)

Network Architectures We test with four deep convolutional neural networks, using a learning rate of 0.001 for the Adam Optimizer (Kingma and Ba 2017):

- ResNet18 (He et al. 2016) (batch size - 230), https://pytorch.org/vision/stable/models.html#torchvision.models.resnet18
- DenseNet (Huang et al. 2017) (batch size - 64), https://pytorch.org/vision/stable/models.html#torchvision.models.densenet121
- Inception (Szegedy et al. 2016) (batch size - 98), https://pytorch.org/vision/stable/models.html#torchvision.models.inception_v3
- CORnet (Kubilius et al. 2018) (batch size - 128, learning rate - 0.0001), https://github.com/dicarlolab/CORnet

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.

C Additional Results (Section 6)

We provide additional evidence across different conditions that further strengthen the conclusions in the paper:

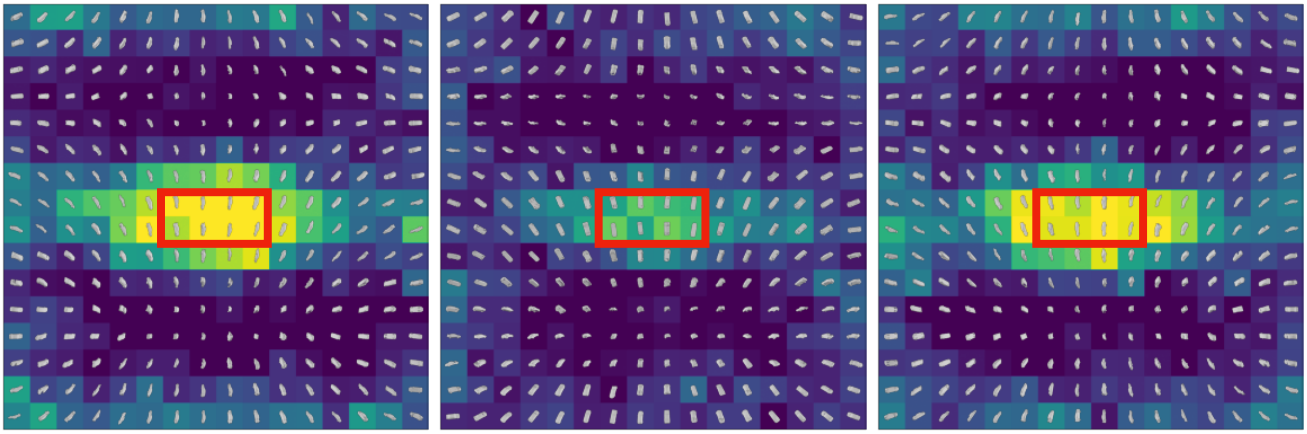
Network architectures. We show consistent results with different network architectures, namely DenseNet, Inception and CORnet. Fig. 12 compares the accuracy of the network architectures. While all them perform similarly *in-distribution*, for OoD Inception and DenseNet tend to perform best and CORnet the worst. The OoD accuracy of all networks tends to increase as the data diversity is increased. We also evaluate the model for per-orientation generalization. Fig. 13, 14 and 15, show the Pearson correlation of the model and the per-orientation accuracy for DenseNet, Inception and CORnet, respectively. Results are consistent with the results of ResNet presented in the paper.

Training regimes. In the paper we reported the accuracy of ResNet trained with data augmentation and pretrained in ImageNet. Fig. 16 and 17 evaluate the predictive model of the per-orientation accuracy. As expected, for the network trained with data augmentation, the *in-plane* component of the model (E) is more prominent, even for small amount of data diversity. This is because data augmentation facilitates generalization to *in-plane* orientations as they are included in the training set. For the network pretrained in ImageNet, the model behaviour is the same as when the network is trained from scratch, *ie.,* generalization to *in-plane* orientations emerges when data diversity is increased.

Amount of training examples. We verify that conclusions in the paper are not dependent on the number of training examples. Fig. 18a shows that decreasing the number of training examples to half leads to a decrease of both *in-distribution* and OoD accuracy, and Fig. 18b shows that the per-orientation behaviour of the network is the same with half of the data.

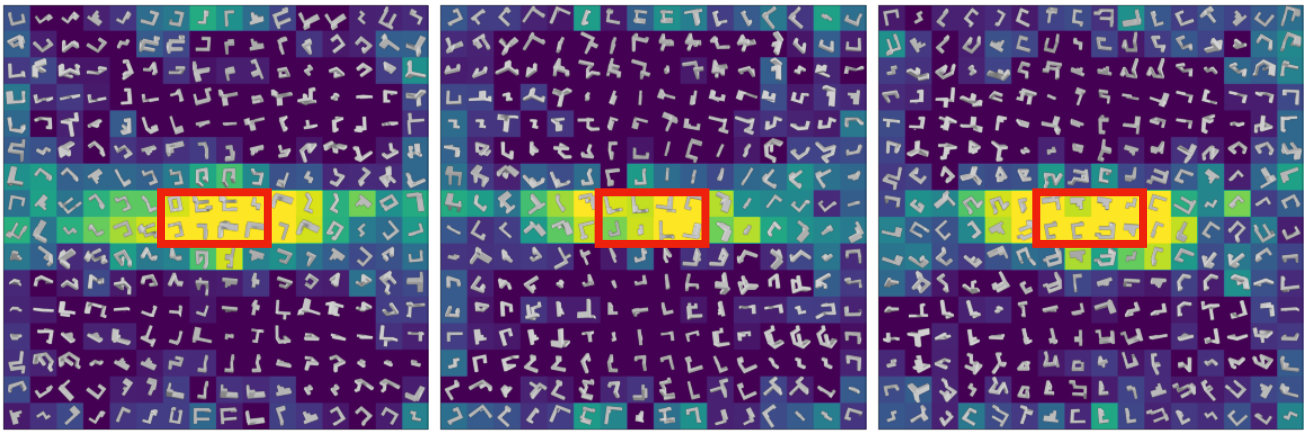
Object Scale. In all presented experiments the objects appear at the same scale. We verified that the conclusions in the paper are not dependent on this factor. We generated a dataset in which the *fully-seen* instances appear at different scales, from 65% of the image to 100% of it. The *restricted-seen* instances appear only at 85% of the image size. Thus, in this experimental setting OoD generalization requires tackling scale and orientation. Fig. 19a shows the OoD for different testing scales. We observe that the accuracy decreases as the scale of the object instance is more dissimilar from the scale in the training set (note that the maximum is at around 85%, which is the *in-distribution* of the *restricted-seen* instances). The drop of accuracy is relatively small though, which suggests that the network learned some degree of robustness to changes of scale. We also observe that the OoD accuracy increases for all scales as the data diversity is increased. Regarding the Pearson correlation of the model and the per-orientation accuracy, in Fig. 19b we observe similar trends as with the experiments without scale variations.

Symmetric objects. Finally, we also verify that the conclusions in the paper are not dependent on the symmetry properties of the object. In principle, symmetry should facilitate generalization to OoD orientations due to more similarity across *in-distribution* and OoD orientations. It could also harm OoD generalization if the *in-distribution* orientations are more similar between them due to symmetry. We experimented with a dataset of lamps as most of these objects are solids of revolution, which are highly symmetrical. We use the lamps 3D models in ShapeNet in the same way as we used the airplanes and cars. Fig. 20a shows examples of instance of lamps (note that most of them are solids of revolution, and when seen from the top, they have an infinite amount of axis of symmetry). Fig. 20b shows that the OoD accuracy increases as the data diversity is increased. Fig. 20c shows that the predictive model of the per-orientation accuracy has the same trends as for the rest of the objects except in the cases when rotating on α. This is explained by observing that when rotating on α the *restricted-seen* instances all look exactly the same due to the symmetry of the lamp. Thus, the training set is not really diverse even though the lamps are being rotated.



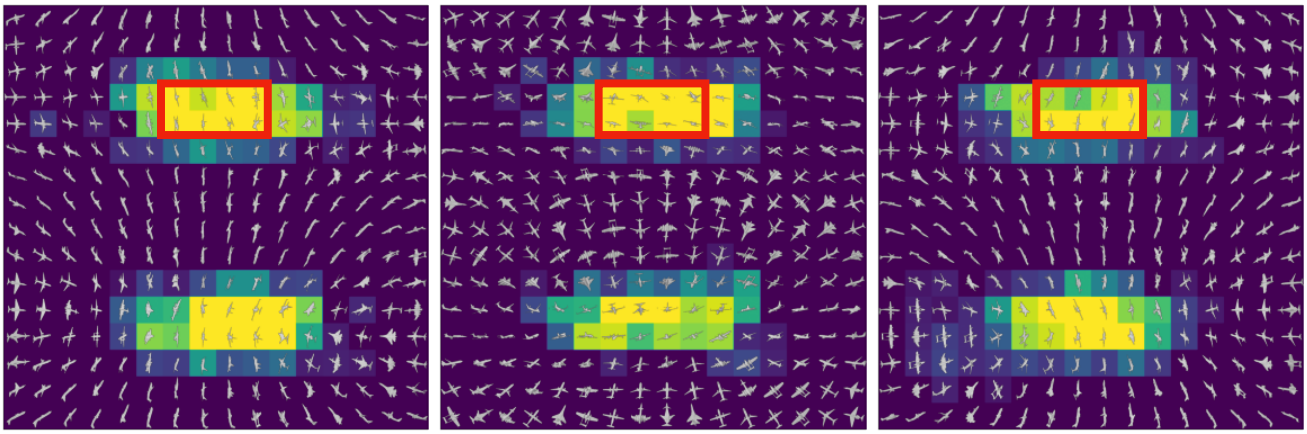
$-1.96 \leq \alpha \leq -1.57$ $0 \leq \alpha \leq 0.39$ $1.96 \leq \alpha \leq 1.57$

Figure 6: **Car Dataset, Rotating on** α



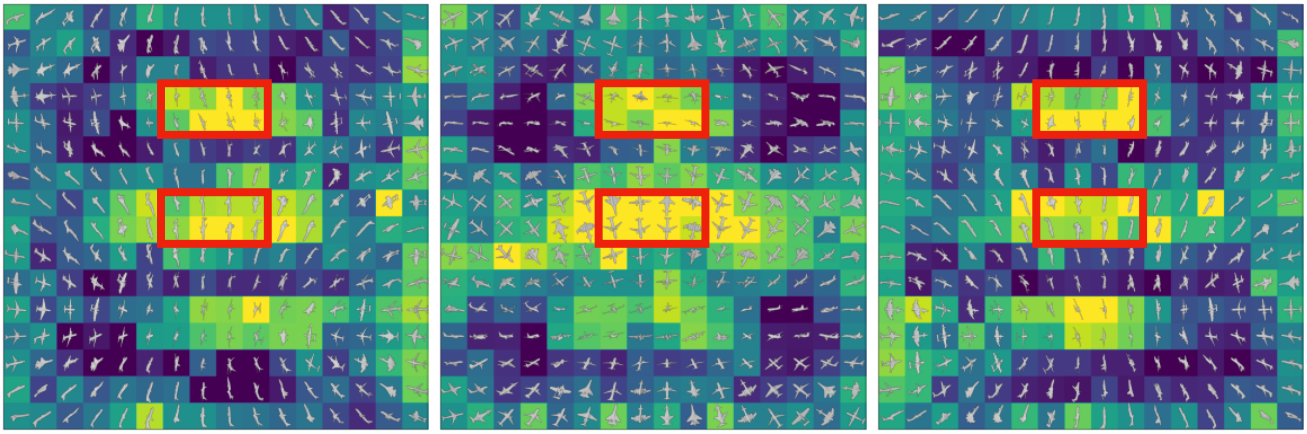
$-1.96 \leq \alpha \leq -1.57$ $0 \leq \alpha \leq 0.39$ $1.96 \leq \alpha \leq 1.57$

Figure 7: ***Shepard & Metzler* Dataset, Rotating on** α



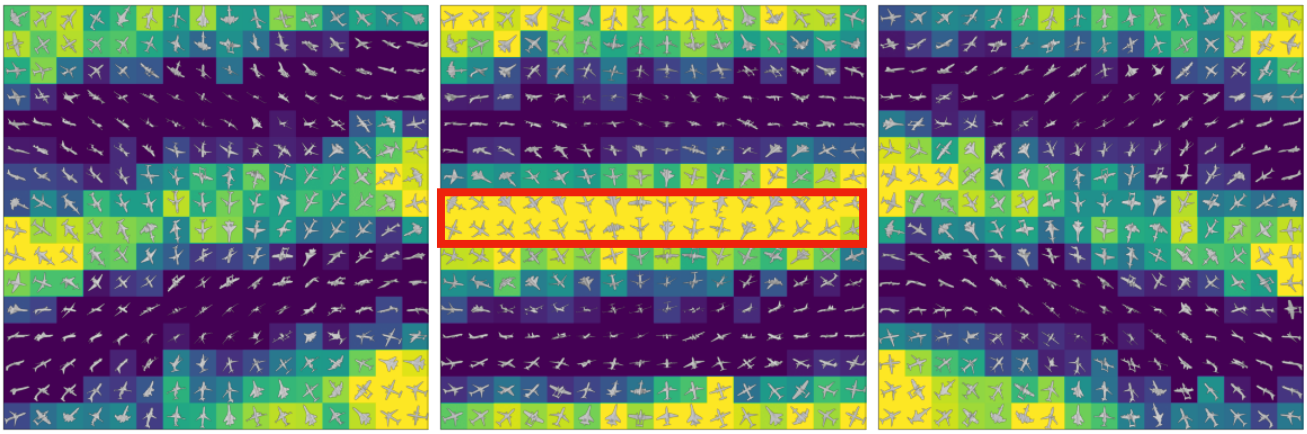
$-1.96 \leq \alpha \leq -1.57$ $0 \leq \alpha \leq 0.39$ $1.96 \leq \alpha \leq 1.57$

Figure 8: **Airplane Dataset, Rotating on** $\alpha, (\gamma \approx \pi/2)$



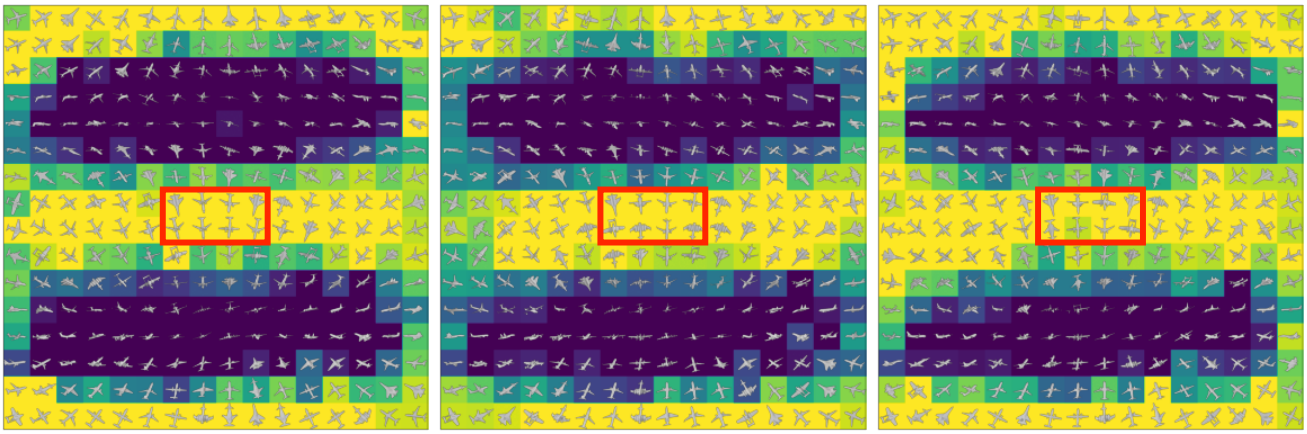
$-1.96 \leq \alpha \leq -1.57$ $0 \leq \alpha \leq 0.39$ $1.96 \leq \alpha \leq 1.57$

Figure 9: **Airplane Dataset, Rotating on** $\alpha, (\gamma \approx \pi/2 \cup \gamma \approx 0)$



$-2.75 \leq \alpha \leq -2.36$ $0 \leq \alpha \leq 0.39$ $2.36 \leq \alpha \leq 2.75$

Figure 10: **Airplane Dataset, Rotating on** β



$N = 20$ $N = 30$ $N = 40$

Figure 11: **Airplane Dataset, Rotating on** α**, Data Augmentation Applied During Training**

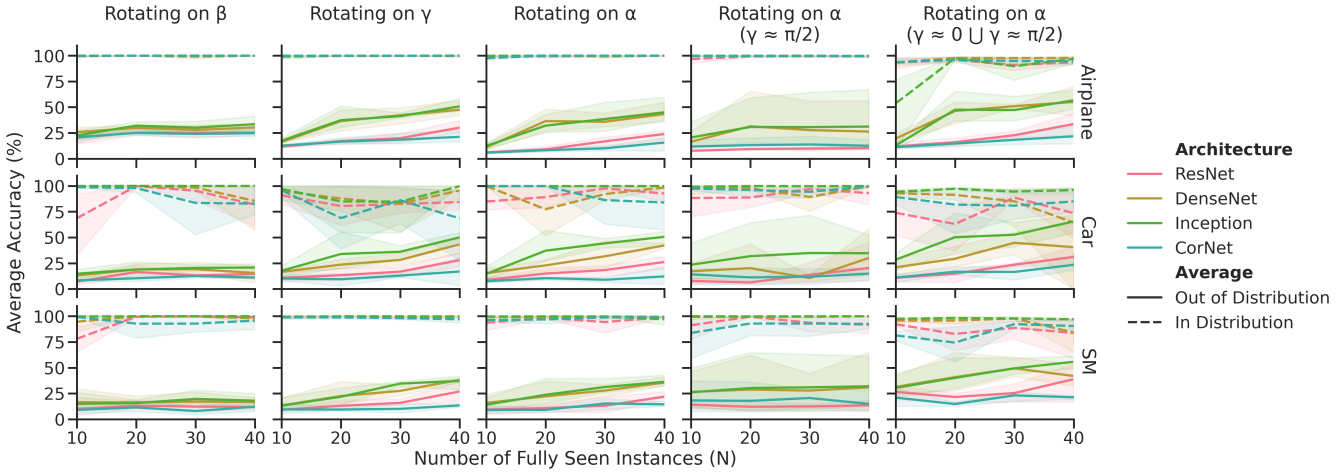


Figure 12: **Average Accuracy for Various Architectures**

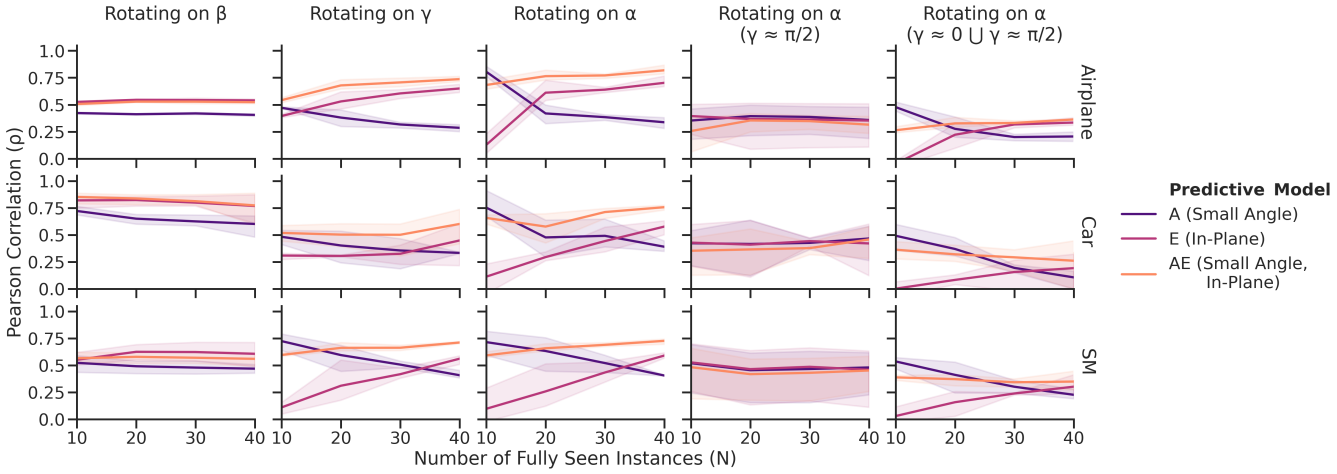


Figure 13: **Correlation Between Proposed Predictive Models and Experimental Results for DenseNet**

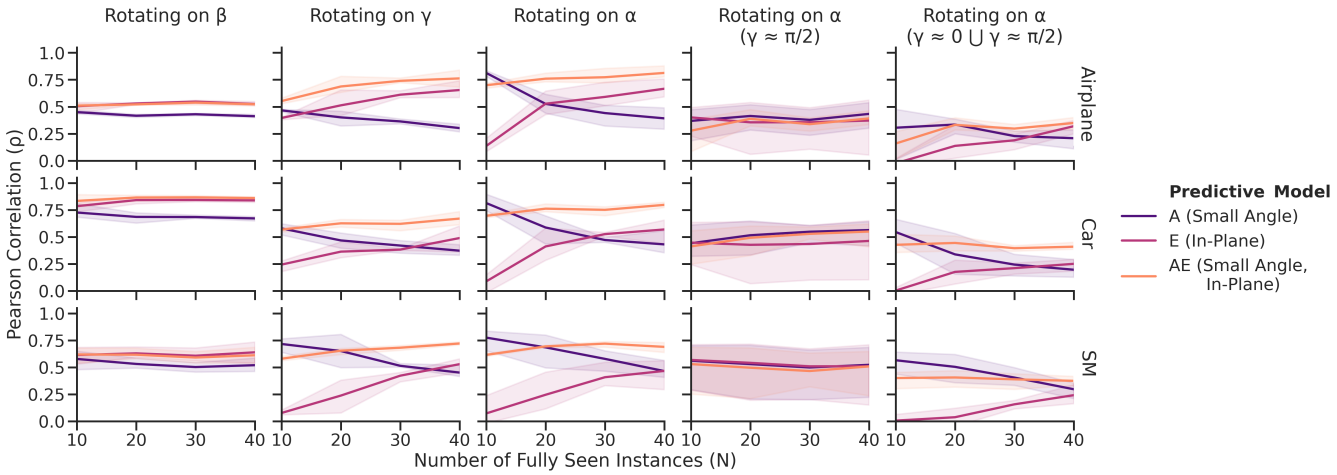


Figure 14: **Correlation Between Proposed Predictive Models and Experimental Results for Inception**

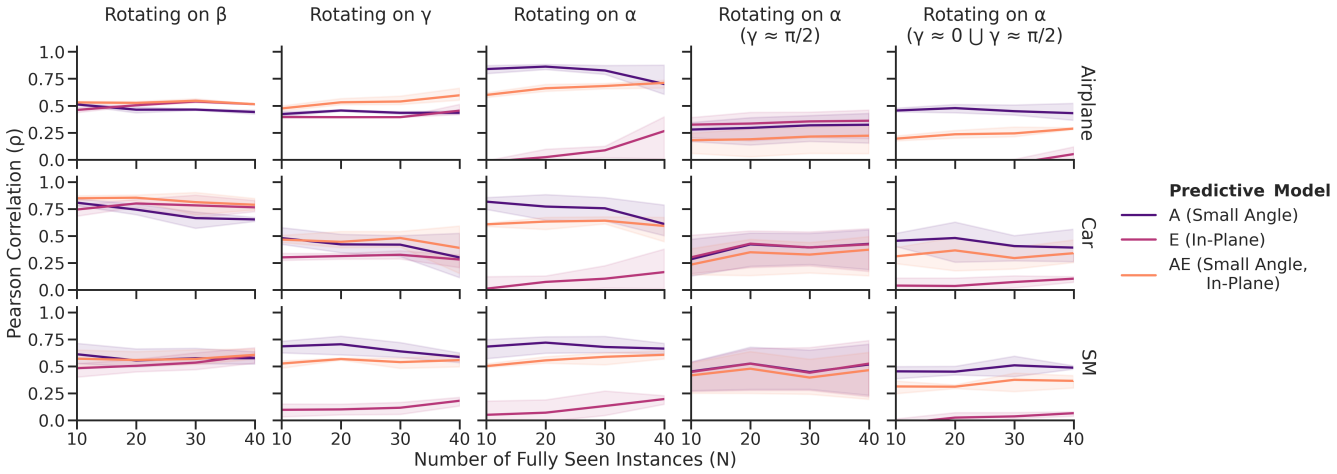


Figure 15: **Correlation Between Proposed Predictive Models and Experimental Results for CORnet**

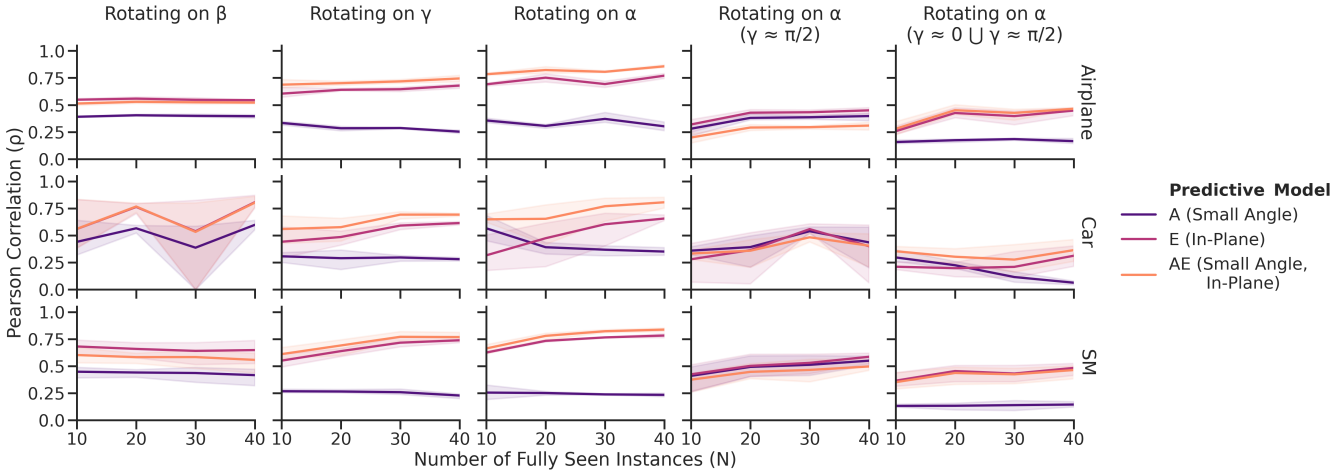


Figure 16: **Correlation Between Proposed Predictive Models and Experimental Results for ResNet, with Data Augmentation During Training**

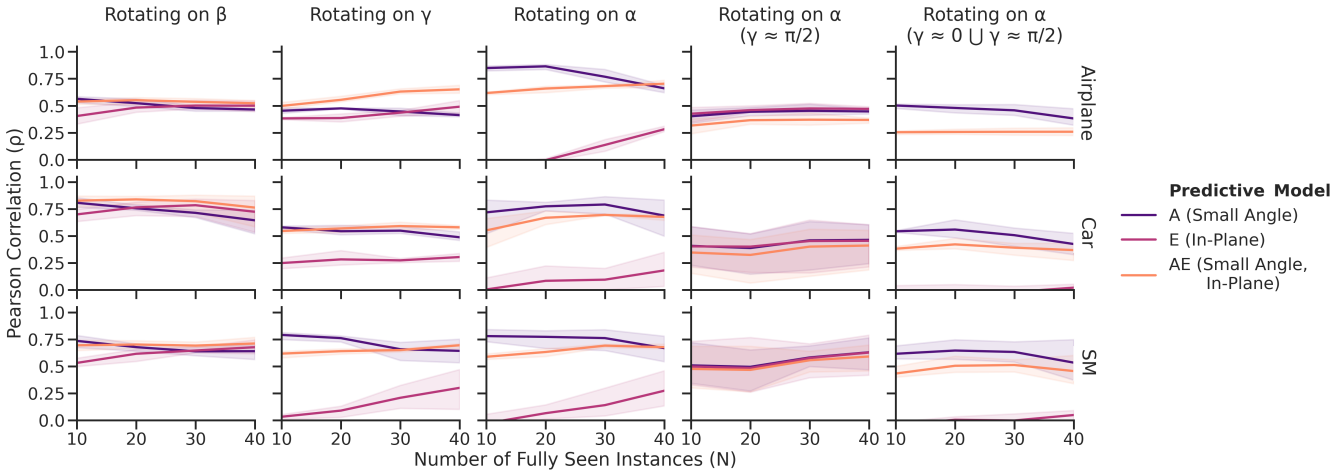


Figure 17: **Correlation Between Proposed Predictive Models and Experimental Results for Pretrained ResNet**

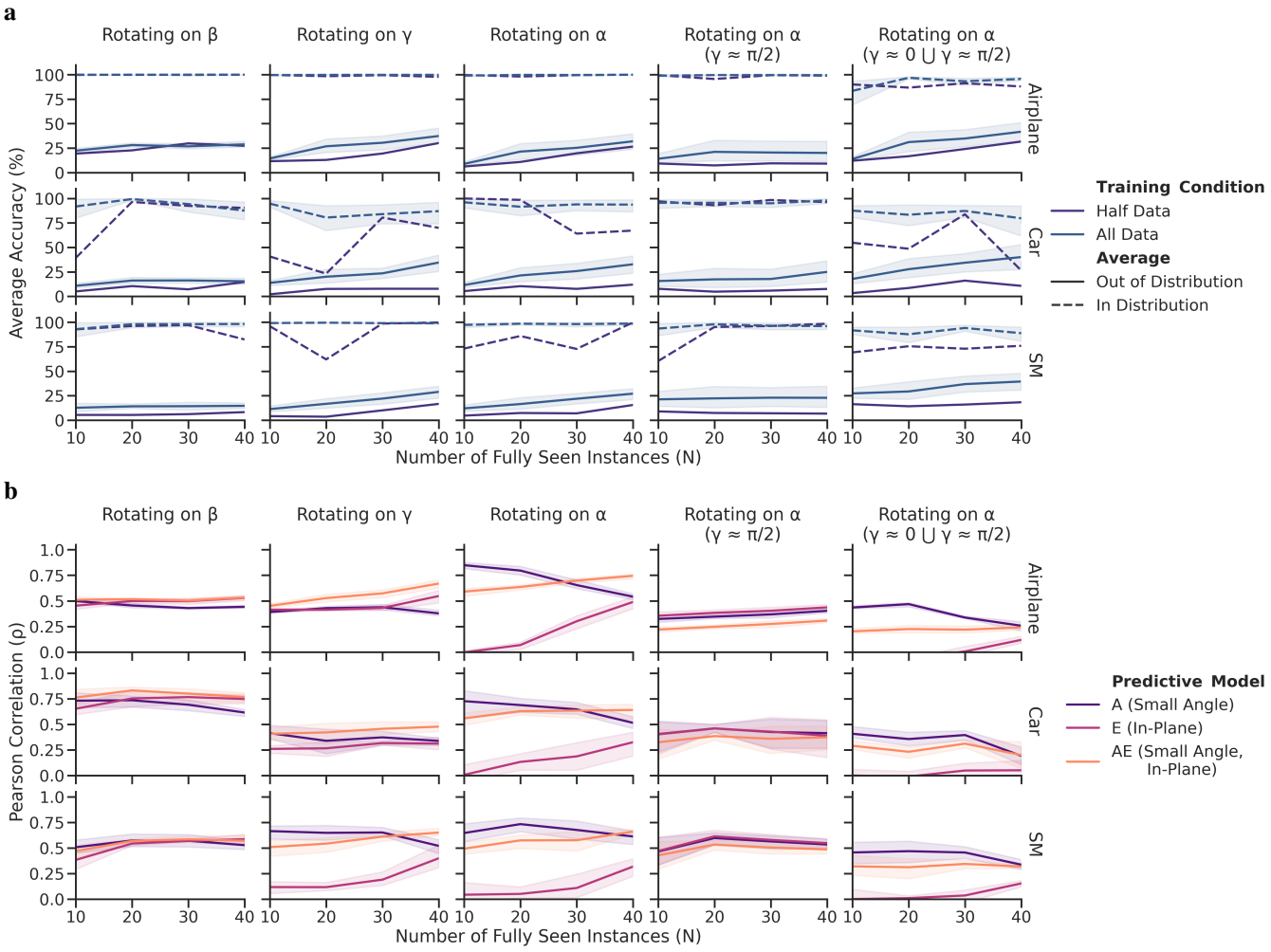


Figure 18: **ResNet with 1/2 Training Data: (a) Average Accuracy (b) Correlation Between Proposed Predictive Models and Network Trained with 1/2 of the Data**

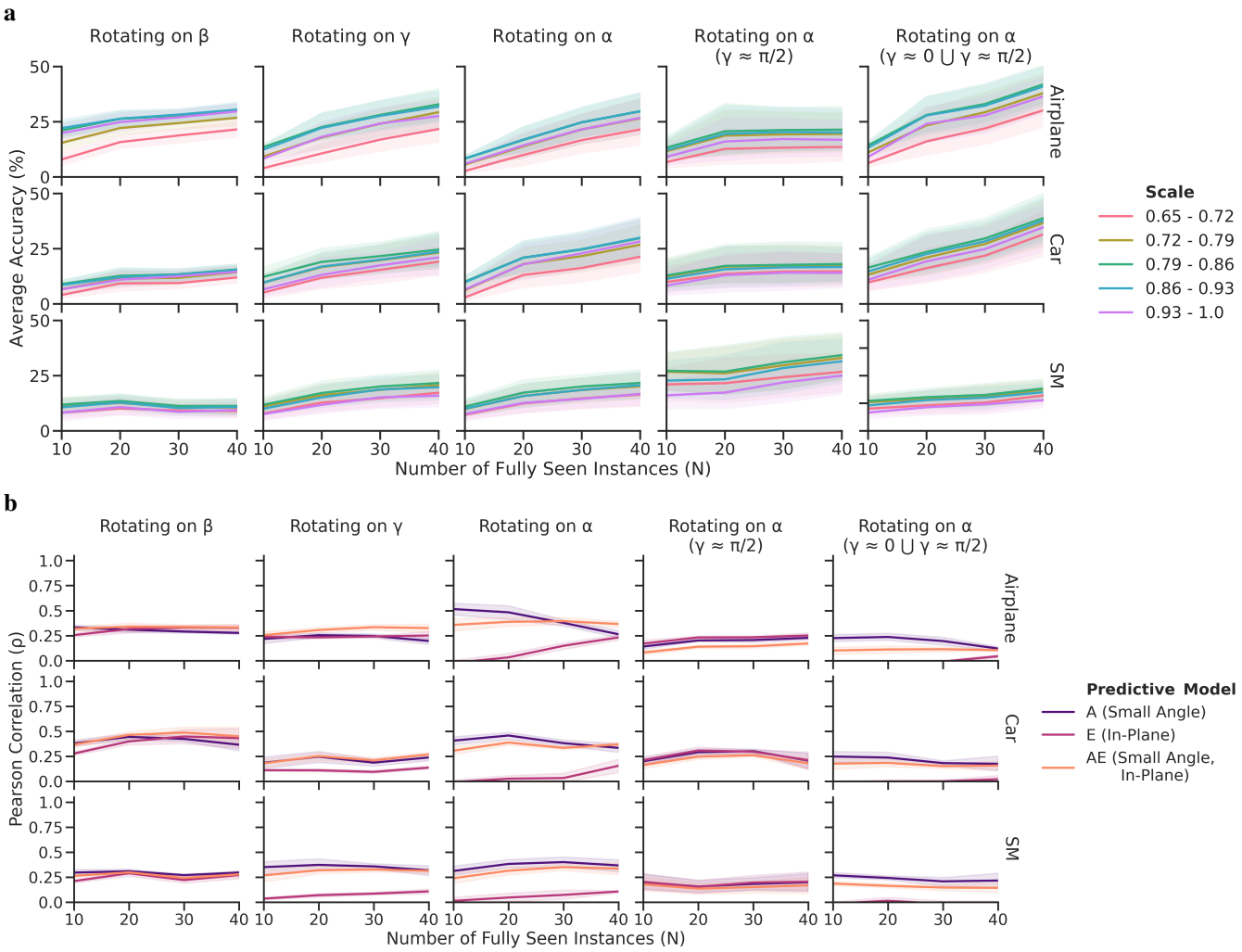


Figure 19: **ResNet with Random Scaling During Training: (a) Average Accuracy (b) Correlation Between Proposed Predictive Models and Experimental Results**

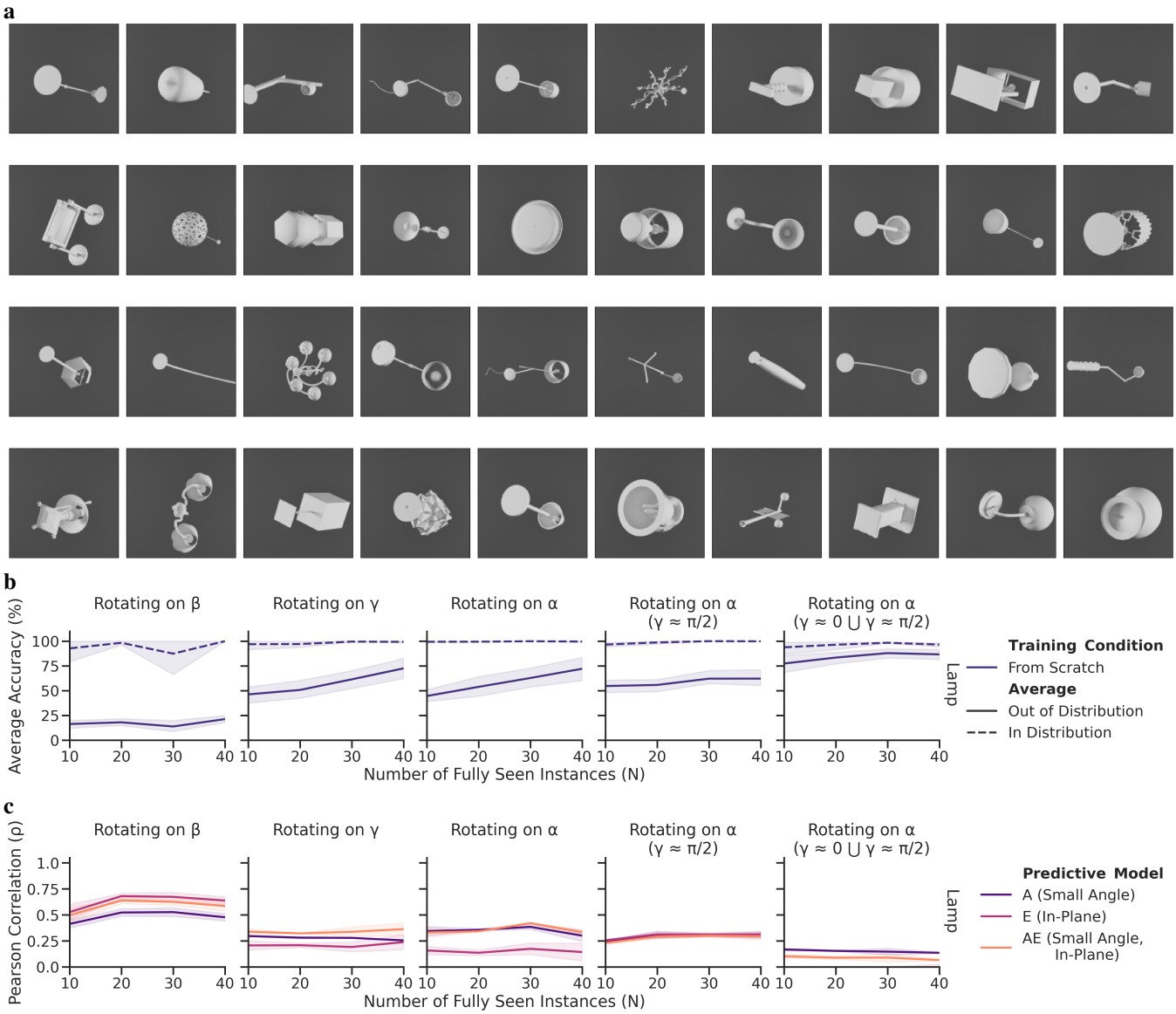


Figure 20: **Lamps Dataset, ResNet: (a) Examples of lamp instances (b) Average Accuracy (c) Correlation Between Proposed Predictive Models and Experimental Results**