Robustness to Transformations Across Categories: Is Robustness Driven by Invariant Neural Representations?

Hojin Jang
jangh@mit.edu
Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139, U.S.A.

Syed Suleman Abbas Zaidi
ssazaidi@mit.edu
Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139, U.S.A.; and Department of Electrical and Computer Engineering, Technical University of Munich, 80333 Munich, Germany

Xavier Boix
xboix@fujitsu.com
Department of Brain and Cognitive Sciences and Center for Brains, Minds and Machines, MIT, Cambridge, MA 02139, U.S.A.; and Fujitsu Research of America, Sunnyvale, CA 94085, U.S.A.

Neeraj Prasad
nprasad@mit.edu
Sharon Gilad-Gutnick
sharongu@mit.edu
Shlomit Ben-Ami
shlomit@mit.edu
Pawan Sinha
psinha@mit.edu
Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139, U.S.A.

Deep convolutional neural networks (DCNNs) have demonstrated impressive robustness to recognize objects under transformations (e.g., blur or noise) when these transformations are included in the training set. A hypothesis to explain such robustness is that DCNNs develop invariant neural representations that remain unaltered when the image is transformed. However, to what extent this hypothesis holds true is an outstanding question, as robustness to transformations could be achieved with properties different from invariance; for example, parts of the network could be specialized to recognize either transformed or...
nontransformed images. This article investigates the conditions under which invariant neural representations emerge by leveraging that they facilitate robustness to transformations beyond the training distribution. Concretely, we analyze a training paradigm in which only some object categories are seen transformed during training and evaluate whether the DCNN is robust to transformations across categories not seen transformed. Our results with state-of-the-art DCNNs indicate that invariant neural representations do not always drive robustness to transformations, as networks show robustness for categories seen transformed during training even in the absence of invariant neural representations. Invariance emerges only as the number of transformed categories in the training set is increased. This phenomenon is much more prominent with local transformations such as blurring and high-pass filtering than geometric transformations such as rotation and thinning, which entail changes in the spatial arrangement of the object. Our results contribute to a better understanding of invariant neural representations in deep learning and the conditions under which it spontaneously emerges.

1 Introduction

A widely known strategy to gain robustness to image and object transformations in deep convolutional neural networks (DCNNs) is to include the transformation in the training set (Wang & Perez, 2017; Schmidt et al., 2018; Geirhos et al., 2018, 2019). Geirhos et al. (2018) recently demonstrated that this strategy is successful in achieving robustness to local transformations, such as blur and noise, and also outperforms human-level recognition accuracy in transformed images. In a subsequent paper, Geirhos et al. (2019) show that including some transformations in the training set that emphasize the shape of the object rather than the texture leads to robustness to transformations beyond the ones seen during training.

Despite these recent results demonstrating effective ways of achieving robustness to transformations, the field still lacks a good understanding of the underlying neural mechanisms that enable such robustness. Decades of research at the intersection of neuroscience and computer vision have led to the hypothesis that neural networks may develop invariant neural representations to achieve robustness to transformations (Riesenhuber & Poggio, 1998; Quiroga et al., 2005; Goodfellow et al., 2009; Achille & Soatto, 2018; Poggio & Anselmi, 2016), that is, the internal neural representations may remain mostly unaltered when the image is transformed. Yet, to what

---

1 We use the term robustness to a transformation (or simply robustness) to denote that DCNN classification accuracy does not deteriorate in transformed images, whereas we use the term invariance to a transformation (or invariance) to indicate that the internal neural representations of the DCNN are not altered by the transformation.
extent this hypothesis holds true is an outstanding question. It is possible that to achieve robustness, neural representations develop properties different from invariance when transformations are included in the training set; for example, parts of the network may specialize in recognizing either transformed or nontransformed images.

To understand the emergence of invariant neural representations in DCNNs trained with transformations, an important question to address is to what extent and under what conditions such representations arise. Measuring the degree of invariance in neural representations can offer some hints but falls short of providing a comprehensive answer. A high degree of neural representations does not guarantee robustness to transformations, as even a small change in the representations can potentially result in poor robustness (e.g., adversarial examples). Only when the neural representations exhibit perfect invariance to image degradations can we ensure robustness to transformations. In this article, we aim to study the potential relationship between invariance and robustness to transformations by examining the impact of invariant neural representations on robustness beyond the training distribution. Invariant neural representations imply that if the network correctly classifies images from a category, it will correctly classify transformed versions of the same image as well. This fundamental property of invariant neural representations enables robustness to object categories that the network has not seen with transformations during training (Poggio & Anselmi, 2016). Thus, we analyze a training paradigm in which only some object categories are seen transformed during training and evaluate robustness to transformations for object categories that the network has not seen transformed. We denote this form of robustness as across-category robustness, while we use the term within-category robustness to designate robustness to categories that are seen transformed during training. In this way, we assess the emergence of invariant neural representations by evaluating across-category robustness. To strengthen the evidence obtained with such evaluation, we also define a metric to directly measure the degree of invariance in the neural representations. This metric, together with the across-category robustness evaluation, provides the relationship between invariant neural representations and robustness to transformations.

A series of experiments with state-of-the-art DCNNs in FaceScrub (Ng & Winkler, 2014) and ImageNet (Deng et al., 2009) reveal that robustness is not always related to invariant neural representations, as networks show robustness to categories seen transformed during training even in the absence of invariant neural representations. Invariance only emerges as the number of transformed categories in the training set is increased. Moreover, our observations indicate that different categories of transformations exhibit varying degrees of invariance, with local transformations (e.g., blurring and high-pass filtering) exhibiting an increase in invariance roughly two times greater than that of geometric transformations (e.g., rotations and
thinning). Our analysis uncovers new evidence that invariance emerges at the individual neural level and demonstrates that it facilitates robustness to categories that the network has not seen transformed during training.

These results are the first milestone in understanding the neural representations that contribute to robustness to transformations. Furthermore, answering what representations emerge in the network could guide research to develop more robust models, and investigating across-category robustness could help to reduce the number of training examples, as not all categories need to be seen transformed to achieve robustness.

2 Signatures of Invariant Neural Representations

In this section, we introduce our methodology for investigating to what extent invariant neural representations contribute to robustness to transformations. For simplicity, we consider DCNNs trained on a set of images with only one transformation, as well as a set of nontransformed images. This yields robustness to the trained transformation but is unlikely to result in robustness to other transformations that the DCNN was not trained on Geirhos et al. (2018).

A neural network is completely invariant when the neural representations in the DCNN are identical for the transformed and nontransformed images. While this is ideal, it is important to note that it is not a prerequisite for robustness: in practice, a certain amount of invariance may emerge, even without attaining identical neural activity, and result in robustness, that is, perfect invariance may not be needed in order to achieve robustness. The opposite may be true as well, as a certain amount of invariance may be unrelated to robustness—that is, high but imperfect degree of invariance may be epiphenomenal. To confirm the relationship between invariance and robustness, we evaluate across-category robustness in conjunction with the degree of invariance, which can offer strong evidence for the role of invariance in achieving robustness. In the following, we first introduce the paradigm to evaluate across-category robustness and then the measure of invariance to transformations.

2.1 Paradigm to Evaluate the Across-Category Robustness. To test across-category robustness, we use the procedure depicted in Figure 1. Images of the subset of categories denoted as seen-transformed appear both transformed and nontransformed in the training set. The remainder of the categories, denoted unseen-transformed, appear in the training set only in their nontransformed version. Note that the sum of the number of categories in each set is equal to the total number of categories in the data set. The network is trained with both seen- and unseen-transformed categories by uniformly randomizing the categories and images during each training step. When an image of the seen-transformed set is used, it is also randomly chosen with equal probability to appear as transformed or not transformed.
Figure 1: Experimental paradigm and image transformations. (a) The DCNN is trained with the transformations in the training set restricted to the *seen-transformed* categories. The accuracy for transformed images is evaluated in categories that are either *seen* or *unseen* transformed. (b) We analyze 12 image transformations from four families: color, convolutional, noise, and geometric.

The network’s task is to predict the category of the input image. We evaluate the accuracy of the network on transformed test images for the seen- and unseen-transformed categories separately, which yields the within- and across-category accuracy, respectively. We train multiple networks with different numbers of seen-transformed categories during training. In all the experiments, the network predicts a category among all categories in the data set—that is, the possible output choices of the network remain the same when evaluating within- and across-category accuracy. Thus, robustness among all the networks can be compared without bias because the only change between them is the number of seen-transformed categories during training. Alternative paradigms are also available in addition to our proposed one, and more details can be found in appendix C.

2.2 Invariance to Transformations at the Individual Neural Level. Understanding DCNNs at the individual neural level has led to promising results that provide granular and simple explanations of the underlying mechanisms behind DCNN generalization (Zeiler & Fergus, 2014; Zhou et al., 2018; Olah et al., 2018). Individual neurons are commonly interpreted as feature detectors that are tuned to image features. A neuron is active (i.e., the neuron’s output value is high) when the feature is present in the image; otherwise the neuron is not active and the output value of the neuron is low. Neurons are tuned to features that are not necessarily interpretable by humans (Bau et al., 2017). We analyze the invariance of a neuron by evaluating the change of its activity when the image is transformed. For example, a neuron may be tuned to detect a dog nose because it is active only in the
presence of a dog nose in the image. If the neuron is invariant, it will continue being active when the dog nose is transformed.

Invariance in neurons that are active may help to achieve robustness to transformations, but this is not the case for inactive neurons. If a neuron is inactive in both the transformed and the nontransformed images, the activity is invariant to the transformation. Yet this is because the feature that the neuron is tuned to is not present in the image, which does not contribute to the network’s classification accuracy. Also, neurons whose activity is constant across all images (i.e., neurons that are not tuned to any feature) are invariant and do not contribute to network accuracy. Thus, invariance is helpful to achieve robustness to transformations only when neurons that are activated by an image are also invariant to the transformation.

Following these intuitions, Goodfellow et al. (2009) introduced a measure of invariance that analyzes the activity in images that generate the top 1% of the neuron’s output value. These images are transformed, and invariance is evaluated by assessing whether the activity remains in the top 1% or not. This metric is effective to discard the cases of invariance that are not related to robustness. Yet it is unclear to what degree the activity generated in a neuron by the rest of the 99% of images that are not analyzed relates to robustness. In the following, we introduce a procedure that controls that the majority of the neural activity related to robustness is taken into account by the invariance measurement. We first introduce the procedure to determine when a neuron is active and then address the measure of invariance of active neurons.

2.2.1 Active Neurons. Let \( \mathcal{X} \) be the images of the test set, and let \( x \in \mathcal{X} \) be one of its images. We refer to the transformed version of the image as \( T(x) \). For notation simplicity, we consider \( T(x) \) to be the transformation that is included in the training set without specifying the type transformation. Let \( \hat{f}_k(x) \) be the activity of the neuron indexed by \( k \) when the network’s input is \( x \in \mathcal{X} \). Note that \( \hat{f}_k(x) \) can be multiplied by a factor without affecting the performance of the network if the postsynaptic weights are divided by that factor. Thus, without the postsynaptic weights, it is not possible to determine the relative importance of each neuron. Thus, we consider the order of magnitude of \( \hat{f}_k(x) \) not relevant, and we discard it by normalizing \( \hat{f}_k(x) \). We define \( f_k(x) \) as the normalized activity of neuron \( k \), such that \( f_k(x) \) is equal to 1 for the image that generates the maximum activity among transformed and not-transformed images. Note that the minimum value that \( f_k(x) \) can take is zero because the neurons use ReLU activation (Krizhevsky et al., 2012).

We define the neuronal tuning of a neuron as the set of images that generate an activity above a threshold. Let \( \tau \) be such a threshold, and let \( A_k \) be the set of images that generate an output value higher than \( \tau \). Thus, \( A_k \) represents the neuronal tuning of neuron \( k \) as it contains the set of images that yield the neuron active. \( A_k \) is the following subset of \( \mathcal{X} \):
$A_k = \{ x \in X \mid (f_k(x) > \tau) \lor (f_k(T(x)) > \tau) \}$, in which $\lor$ is the “logic or” operator. Note that the activity generated by both transformed and nontransformed images is considered.

In order to determine the value of $\tau$, we need to take into account a compromise between the following two factors. On one hand, the value of $\tau$ should be as high as possible, such that the neuronal tuning is specific and invariance due to lack of activity is discarded. On the other hand, $\tau$ should be made as low as possible in order to include in $A_k$ all the activity that is relevant for the network’s robustness. To do so, we validate that the network’s robustness is not affected by ablating all neurons except the active, given an image. Also, we validate that the network’s accuracy is at chance when all the active neurons are ablated. These tests provide reassurance that the active neurons drive the network’s robustness. In the experiments in section 3.2, we show that neurons are active on only a small set of images, and their activity is almost entirely responsible for the network’s robustness.

2.2.2 Measure of Invariance in Active Neurons. In order to measure the invariance of a neuron, only the images in the set $A_k$ will be used. This is because, as previously discussed, the invariance of neurons that are not active for a given image is not related to robustness. We define $h_k(x)$ as the invariance measure of neuron $k$ given an image $x \in X$. We use the following expression based on the normalized difference between the neurons’ activity of the transformed and nontransformed images:

$$h_k(x) = 1 - \frac{|f_k(T(x)) - f_k(x)|}{|f_k(T(x)) + f_k(x)|}. \quad (2.1)$$

We can see by analyzing equation 2.1 that $h_k(x)$ is equal to 1 when the neuron is perfectly invariant (i.e., $f_k(T(x)) = f_k(x)$) and is equal to 0 when the neuron’s activity varies maximally, such that $f_k(T(x)) = 0$ and $f_k(x) = 1$ and vice versa. This invariance measure takes values between 0 and 1 depending on how much the activity of $x$ and $T(x)$ varies. We define $I_k$ as the invariance coefficient of neuron $k$. $I_k$ is obtained by averaging $h_k(x)$ over the set of images that the neuron is active, $A_k$, that is, $I_k = \frac{1}{\#A_k} \sum_{x \in A_k} h_k(x)$, in which $\#A_k$ is the cardinality of the set $A_k$ (the number of images that neuron $k$ is active for). In the experiments, we analyze the distribution of $I_k$ among neurons in a layer.

3 Results

We evaluate robustness to 12 image transformations shown in Figure 1 (see appendix A for details). These are grouped in four transformation types: color (gray-scale and color rotation), geometric (thinning and rotation), convolutional (blur, high pass, and horizontal, vertical filtering, and contrast
inversion), and noise (white and salt and pepper). Our choice of transformations was motivated by studies showing that DCNNs are robust to them when they are included in the training set (Geirhos et al., 2018).

To ensure that our results are not data set- or architecture dependent, we evaluate AlexNet (Krizhevsky et al., 2012) in the FaceScrub data set (Ng & Winkler, 2014) and ResNet-18 (He et al., 2016) in the ImageNet data set (Deng et al., 2009). In ImageNet we evaluate only four transformations (blur, high-pass filtering, white noise, and salt and pepper noise) as some other transformations are already present in the training set and cannot be excluded (note that in ImageNet, objects appear at different orientations, sizes, contrasts, and colors). We use the loss function, regularizers, optimizer, and other hyperparameter choices reported in the literature, except for the learning rates that are tuned for each experiment (see appendix B for more details).

3.1 Results on Across-Category Robustness. In the following, we report the accuracy of the network using the paradigm introduced in section 2.1. For both FaceScrub and ImageNet, we report top-1 accuracy (%).

3.1.1 Reproducing Previous Results on Robustness to Transformations. Figure 2a depicts the accuracy of networks trained with each transformation and tested on all transformations. As expected, on nontransformed images, the networks perform similar to what is reported in the literature (Vogelsang et al., 2018; He et al., 2016) independent of which transformations are included in the training set. On transformed images, networks are robust to the transformations included in the training but not robust to other transformations, a finding consistent with Geirhos et al. (2018). Note that color transformations are an exception, as all networks have some degree of robustness to them. This may be because the color transformations are in part already included in the data set as a consequence of different illumination conditions. Additionally, the network trained with blur transformation is slightly robust to the noise transformations, likely because training with blur yields large receptive fields in the first layer that may help to filter out the noise—a finding consistent with previous studies (Vogelsang et al., 2018).

3.1.2 Across-Category Robustness Improves as the Number of Seen-transformed Categories is Increased. In Figures 2b and 2c, the within- and across-category accuracy are evaluated for different numbers of seen-transformed categories. The within-category accuracy (at the top of the figure) and the across-category accuracy (at the bottom) are reported. We can observe that the across-category accuracy increases as more categories are seen transformed. This indicates that invariant neural representations may have emerged in the network (we cannot confirm this without analyzing the neural activity, as across-category robustness is a consequence of
3.1.3 Color, Convolutional and Noise Transformations Lead to Higher Across-Category Robustness than Geometric Transformations. The increase of the invariance but not the reverse). Also, observe that when the network is trained with few seen-transformed categories, the robustness for these categories cannot be driven by invariance because the network is not across-category robust. Understanding the mechanisms for robustness when there is lack of invariance is an interesting open question that will be analyzed in future work.

Note also that the within-category accuracy decreases as more categories are seen transformed, and it reaches the values previously reported in Figure 2a when all categories are seen transformed. Observe that the within-category accuracy is close to 100% when the number of seen-transformed categories is close to 1. This suggests that the network uses the transformation as a feature to identify the few seen-transformed categories. This can be validated with the confusion matrices between seen- and unseen-transformed categories (see appendix D), which show that transformed images from unseen-transformed categories are mainly confused with seen-transformed categories.
across-category accuracy with the number of seen-transformed categories is common among transformations. Yet, Figures 2b and 2c show that the rate at which the across-category varies is different depending on the transformation. Observe that for color transformations, just few seen-transformed categories are sufficient to achieve a high across-category accuracy. This may be because the network is already quite robust to color transformations even training only with nontransformed images, as shown in Figure 2a. For convolutional transformations, the increase of the across-category accuracy is not as high as for color transformations, but it is remarkable. The across-category accuracy of the noise transformations is on par with the convolutional in ImageNet but not in FaceScrub. Note that the within-category accuracy for noise transformations in FaceScrub is 10% to 20% lower than for other transformations. Thus, in FaceScrub, the amount of noise impairs both within- and across-category robustness, which suggests that the low across-category accuracy is a characteristic not of the noise transformation but of the overall accuracy in FaceScrub.

For geometric transformations, the across-category accuracy is lower than other transformations, especially for rotations (the increase is only of a few percent when all categories are seen transformed). This may be because spatial rearrangement of the object is involved in geometric transformations but not in the rest. These rearrangements are larger for rotations of 90 or 180 degrees than for thinning, and rotations have much lower across-category accuracy than thinning. Also, to achieve across-category robustness for spatial rearrangements requires capturing the long-range dependencies between object parts, and this has been shown to be difficult for feedforward architectures (Srivastava et al., 2019; Engstrom et al., 2019; Azulay & Weiss, 2019; Jaderberg et al., 2015).

3.1.4 Across-Catgory Robustness May Depend on the Visual Homogeneity of the Object Categories. For convolutional and noise transformations, the across-category generalization in FaceScrub increases linearly with respect to the number of categories seen transformed, while in ImageNet, it increases much more slowly (note that in ImageNet, the number of seen-transformed categories is in logarithmic scale). These differences between data sets may be because the object categories in FaceScrub are more visually homogeneous than in ImageNet, that is, FaceScrub shares common configurations across different facial identities while ImageNet is composed of a varied set of more distinct object categories. Thus, the visual homogeneity of the categories may facilitate across-category generalization (assuming that the invariant features can be more effectively transferred across categories that share comparable visual structures).

3.2 Results on Invariance to Transformations at the Individual Neural Level. Next, we evaluate the amount of invariance using the measure introduced in section 2.2.
3.2.1 Less Than 20% of the Images Activate a Neuron. Recall that the threshold that determines if a neuron is active, $\tau$, is the compromise between two factors: (1) neurons are active for only a few images because they are tuned to specific patterns (i.e., $\tau$ is as high as possible) and (2) only the activity in active neurons matters for robustness (i.e., ablating all the inactive neurons in an image should not drop the accuracy), whereas ablating all the active neurons in an image should bring the accuracy to chance. In practice, we implement a grid search to determine $\tau$ by evaluating values between zero and one in steps of 0.05. We select the highest $\tau$ that leads to a drop of 1% of the accuracy when ablating the inactive neurons. This results in $\tau = 0.1$ for FaceScrub and $\tau = 0.05$ for ImageNet. We validate that neurons are active for only a small subset of images—that is, cardinality of $A_k$ is small for most neurons $k$ and also that each image activates a subset of neurons (i.e., cardinality of $\{k | x \in A_k\}$ is small for most images $x$). In Figure 3a, we show the distributions of both cardinalities using violin plots. Notably, between 1% and 20% of the images activate a neuron, and between 10% and 20% of the neurons are active in an image. Thus, neurons are tuned to patterns that appear in few images, and in most images, only a subset of neurons is active, which fulfills factor 1. In appendix D, we report the network’s accuracy obtained from ablating active and inactive neurons. We observe that when the inactive neurons are ablated (80% to 90% of the neurons per image), the network’s accuracy decreases less than 1% overall. When the active neurons are ablated (10% to 20% of the neurons per image), the network’s accuracy is very low for all transformations. This low but above-chance accuracy suggests that the neural activity of the inactive neurons may play a minor but redundant role toward the network’s robustness, which approximately fulfills factor 2.

In appendix D, we confirm that Goodfellow et al.’s (2009) proposal to measure invariance only in the top 1% of the neural activity per neuron...
Figure 4: Invariance for different numbers of seen-transformed categories. (a, b) Violin plots of the invariance coefficient among neurons at the penultimate layer for FaceScrub and ImageNet, respectively. Different numbers of seen-transformed categories are displayed, and the invariance coefficient is reported separately for seen- and unseen-transformed categories.

is incomplete. The results in the appendix provide evidence that supports the notion that a substantial proportion of accuracy performance is still attributed to neurons that were excluded by the threshold of the top 1% proposed by Goodfellow et al. (2009). Specifically, applying the top 20% threshold allows eliminating a large number of neurons that significantly contribute to the network’s accuracy. This observation justifies using the procedure that we introduced as it is a much more rigorous procedure to filter out neurons that are inactive most of the time.

### 3.2.2 Invariance Increases When the Number of Seen-Transformed Categories Increases.

Recall that the invariance coefficient, $I_k$, is evaluated when the neurons are active. In Figures 3b and 3c, we show the distribution of $I_k$ among neurons in the penultimate layer for networks that are trained with or without transformations. All transformations lead to more invariance than when the training is only with nontransformed images. In appendix D, we show that this invariance builds up across layers, which is in accordance with previous work (Goodfellow et al., 2009; Poggio & Anselmi, 2016).

Next, we aim to explore the relationship between invariance and the across-category robustness. Figure 4 shows the distribution of $I_k$ among neurons in the penultimate layer when the number of seen-transformed categories is increased. The distribution for the seen- and unseen-transformed categories is shown separately (see appendix D for all transformations). We
observe that the amount of invariance remarkably increases when the number of seen-transformed categories is increased. Also, the amount of invariance for the seen- and unseen-transformed categories is nearly identical (except when the number of seen-transformed categories is low because there is almost no invariance). Thus, invariance in the network is not category dependent and emerges alongside across-category robustness. This is strong evidence that invariance contributes to robustness when the number of seen-transformed categories is large.

It is worth noting that while most neurons develop invariance to transformations through training, it is unclear what neural mechanisms emerge when there is no invariance. In the case that invariance does not emerge, networks are not robust to transformations of unseen-transformed categories, but they are still robust to the transformation of the seen-transformed categories. A plausible hypothesis to explain this phenomenon is that some neurons may exhibit specialization toward the transformations such that the transformations can be handled for seen-transformed categories but not for unseen-transformed categories. In order to verify the presence of neurons specialized to the transformation, we modified equation 2.1 to calculate a specialization score, which yields a value of 1 if a neuron is selective for $T(x)$ but not $x$, or $-1$ if a neuron is selective for $x$ but not $T(x)$. In appendix D, our results demonstrate that when the transformations that are included in the training set do not lead to the emergence of invariant representations, there are neurons specialized to $x$ and other neurons to $T(x)$, but there is also a range of neurons in between. This can be observed in Figure 12 (in appendix D), especially for the case of rotations, as the violin plot is spread out across different degrees of specialization. In conclusion, when there is no emergent invariance, the neural mechanisms are related to specialization to a certain degree, and more research is needed to fully understand them.

### 3.2.3 Invariance Explains the Across-Category Robustness

Next, we further analyze the relationship between invariance and across-category robustness by scrutinizing the differences among transformations. Figure 5a shows the amount of invariance for all transformations and data sets as the number of seen-transformed categories is increased. We observe that the trends of invariance follow that of across-category robustness. To quantify whether there is a direct relationship between them, in Figures 5b and 5c, we depict the invariance and robustness for each network trained with a different number of seen-transformed categories. Observe that invariance is not needed to achieve within-category robustness (see Figure 5b), whereas the extent of across-category robustness is correlated with the amount of invariance for all transformations (see Figure 5c). Thus, the network develops properties different from invariance to achieve robustness, but as we increase the number of seen-transformed categories, the amount of invariance increases.
3.2.4 Results Are Consistent across Network Architectures. The achievement of invariance can be affected by the network architecture and capacity chosen for the task, presenting a potential point of concern. Although we attempted to address this concern by experimenting with AlexNet and ResNet-18, the differences in the data sets on which these models were trained leave open the question of whether the selection of different networks could produce distinct patterns. To mitigate this concern, we employed the Inception model (Szegedy et al., 2015) and conducted the same experiment on FaceScrub. Our analysis demonstrates a consistent pattern wherein invariance gradually emerges as the number of transformation-seen categories increases (see appendix D). This finding again suggests that the relationship between invariance and the number of transformation-seen categories during training is a generalizable phenomenon, regardless of the specific network architecture employed. However, it is intriguing to note that Inception exhibits a higher number of transformation-invariant neurons than AlexNet, indicating that network architecture and capacity may influence the attainment of invariance. Despite this, the fundamental observation that invariance arises at the neural level and is closely linked to robust performance remains valid. Future studies exploring the association between network capacity and invariance will be of great interest.
4 Discussions and Conclusion

We have demonstrated that increasing the number of seen-transformed categories results in an increase of two network properties: the amount of invariance to transformations and the robustness to transformations of unseen-transformed categories. This suggests that invariance drives robustness to transformations depending on the number of seen-transformed categories. Furthermore, we have shown that invariance emerges at the individual neural level, which adds to the growing body of literature that uses individual neurons as the elemental building blocks to analyze DCNNs (Zeiler & Fergus, 2014; Zhou et al., 2018; Olah et al., 2018). Key open questions derived from our results that will be tackled in future work are understanding why invariant representations emerge for some transformations and what the neural mechanisms for robustness are when there is a lack of invariance (e.g., for geometric transformations and also for small numbers of seen-transformed categories).

One potential limitation of our study is that it examines how networks attain invariance for individual types of transformations independently, whereas multiple types of transformations often co-occur in real-world scenarios. It is worth exploring how individual neurons develop invariance in the presence of multiple transformation types. We anticipate that neurons would form clusters based on relevant transformations and demonstrate distinct patterns across different levels of visual processing. As this exploration would necessitate additional experimental complexity, we concede that it remains an open question that needs to be addressed in future research.

Another promising avenue for future research is to explore how individual neurons respond to transformations that fall outside the range of the training transformations. While our training approach focused on generalizing invariance across different object categories, it remains an open question whether our findings hold under such conditions. The literature suggests that neural network models face significant challenges when expanding their knowledge to unseen conditions beyond the training regime (Montero et al., 2021; Schott et al., 2021). We encourage further investigation of this topic in future research.

Furthermore, in building bridges between machine learning and neuroscience, the current work adds to the existing body of literature that seeks to understand how visual experience drives the development of robust recognition systems (Le Grand et al., 2001; Wood & Wood, 2018; Jang et al., 2021). More specifically, our results suggest that robust recognition behavior across categories can be achieved as individual neurons become tuned to be invariant to experienced transformations. This enriches our understanding of how biological visual systems may develop invariant representations of objects while maintaining their sensitivity to object-specific features, one of
the central questions in computational neuroscience (Palmeri & Gauthier, 2004; Peissig & Tarr, 2007; Tsao & Livingstone, 2008; DiCarlo et al., 2012; Tacchetti et al., 2018). Our computational findings also agree with previous experimental observations made in studies of single neurons. For example, the prefrontal cortex neurons have been found to become invariant to noise after learning to recognize noisy images (Rainer & Miller, 2000). Additionally, our training paradigm may account for the observation that a large number of cells in the macaque showed viewpoint invariance to objects that the monkey had never seen before (Bao et al., 2020).

Our study motivates future endeavors in determining other contributing factors to invariance. We found that the network’s ability to learn invariance was significantly lower when geometric transformations were evaluated. One potential explanation is that our training scheme was tested only within a class of feedforward networks, while the brain heavily relies on recurrent computations (Felleman & Van Essen, 1991; Lamme & Roelfsema, 2000). Recurrent processing has been shown to confer a critical advantage in performing robust object recognition (Wyatte et al., 2012; O’Reilly et al., 2013; Spoerer et al., 2017; Kar et al., 2019). It has been also suggested that recurrent processing is actively engaged when human observers perform mental rotation (Shepard & Metzler, 1971; Roelfsema & de Lange, 2016), which may give a clue to explain why the tested networks in our study failed to achieve invariance to geometric transformations. It will be of considerable interest for future studies to investigate the efficacy of recurrent processing in acquiring invariance to various types of transformation.

Appendix A: Transformations

The transformations were applied to the data set as an input layer in the GPU, such that the transformations were computed at runtime with every forward pass.

For color transformation, the hue was rotated 180° in the hue space. The grayscale operation was performed using built-in TensorFlow functions.

For blur transformation, a gaussian filter was convolved with the input image. The standard deviation of the gaussian was selected such that the top-1 accuracy of a network trained on the untransformed FaceScrub data set was less than 10%. To do so, we increased the standard deviation in steps of 0.5 until the accuracy was less than 10%, and then we stopped. The horizontal and vertical filters have a filter size of 3 × 3 of form [−1, 0, 1]. The high-pass filter has a filter size of 5 × 5 of form [−1, 2, 4, 2, −1] in both directions.

Noise-based transformations were applied by adding white noise and salt and pepper noise to the original images. The standard deviation of noise was selected with the same procedure as the blur transformation. For the white noise, the standard deviation was increased in steps of 25. The salt and pepper noise randomly sets half of the pixels to either 0 or 255.
Geometrical transformations rotated the input images counter clockwise with angles 90° and 180°, respectively. For thinning, the image width was reduced to half and padded with zeros on both sides.

In ImageNet, the transformations were applied with parameters such that the top-1 accuracy was under 15%, using the procedure we have noted to set the parameters of the transformations.

Appendix B: Data Sets and Networks

B.1 Data Sets. We use two data sets: FaceScrub (Ng & Winkler, 2014) and ImageNet (Deng et al., 2009). FaceScrub is a face recognition data set containing over 50,000 images of 388 individuals. ImageNet is an object recognition data set with around 1.4 million images of 1000 object categories.

When increasing the number of seen-transformed categories, we use the same seen-transformed categories before the increase and an additional set of categories randomly selected. The smallest numbers of seen-transformed categories are randomly selected. This procedure facilitates comparing results between different number of seen-transformed categories, as it minimizes the dependency of the selected seen-transformed categories.

B.2 Networks. For FaceScrub, we used a modified AlexNet (Krizhevsky et al., 2012) architecture with five convolutional layers followed by two fully connected layers. The modifications included removal of the max-pooling layers after the first two convolutional layers in order to preserve necessary resolution for face recognition. The rest of the network was exactly as AlexNet (local response normalization, dropout, and initialization parameters were the same as in AlexNet). The networks were trained for a maximum of 45 epochs (sufficient for convergence in all cases) with a learning rate of 1e-4 and a weight decay of 5e-4 for all transformations. These values were selected using grid search over 25 possible combinations of learning rates and weight decay, evaluated on 10% of a held-out images of the training set (all transformations lead to the same hyperparameters). The optimization algorithm was stochastic gradient descent with a momentum of 0.9 and a batch size of 32.

For ImageNet, we trained Resnet18s (He et al., 2016) using the official tensor flow implementation and with the hyperparameters that come by default.

B.3 Preprocessing. All input images were standardized before the first hidden layer. This did not change the accuracy of the network, and it was particularly important to facilitate robustness to transformations without the need to adjust the mean and standard deviation parameters of the batch normalization to the transformation. Before the standardization, the
transformations were applied to the image, whose pixel values were between 0 and 255.

Appendix C: Alternative Training Paradigms

A possible alternative to our paradigm could be considered based on training the network in two steps: first, training the seen-transformed and then the unseen-transformed categories. Yet this has limitations to analyze invariance because lack of across-category robustness in a two-step paradigm can be due to other reasons unrelated to invariance, that is, the two-step paradigm introduces confounding factors. Namely, if the training of the unseen-transformed categories is done by fine-tuning the network after training the network with the seen-transformed categories, lack of robustness could be due to so-called catastrophic forgetting (Kirkpatrick et al., 2017) of the transformation rather than lack of invariance. Otherwise, if the training of the unseen-transformed categories is done without fine-tuning—for example, by training a classifier on the representations learned for the seen-transformed categories—lack of robustness could be because these representations may not be generalizing well to the unseen-transformed, rather than lack of invariance. Training both sets at the same time, as in our proposed paradigm, controls these confounding factors.

Another potential alternative approach could be to use a meta-learning methodology (Vinyals et al., 2016; Snell et al., 2017), in which we can investigate the generalization of invariance learned from the training set to a novel set. Since networks are presumed to acquire the ability to learn more in the concept space, meta-learning could provide a beneficial environment to evaluate invariance less influenced by category bias and could potentially serve as an appropriate platform for testing generalization ability. The extent to which the results from our current paradigm align with those obtained from a meta-learning framework remains an open question.

Appendix D: Additional Results

D.1 Confusion Matrices. In Figure 6, we show the confusion matrix between seen- and unseen-transformed categories. The figure reports the percentage of images of one set of categories classified as the same or another set.

D.2 Active Neurons. To validate that the activity of the active neurons—those with activity higher than $\tau$—captures all the activity relevant for the network’s robustness, we performed two types of ablation experiments: set all activity below the threshold to zero and set all the activity above the threshold to zero. The performance of the DCNN in these two conditions is shown in Figure 7.
Figure 6: Confusion matrix between seen- and unseen-transformed categories for FaceScrubs for all transformations at different numbers of categories transformed during training. Quadrant 1: Percentage of predictions for unseen-transformed categories misclassified as seen-transformed. Quadrant 2: Percentage of predictions for seen-transformed categories correctly classified as seen-transformed. Quadrant 3: Percentage of predictions for seen-transformed categories misclassified as unseen-transformed. Quadrant 4: Percentage of predictions for unseen-transformed categories correctly classified as unseen-transformed.

Figure 7: Accuracy of ablating active and inactive neurons. Within-category accuracy after ablating inactive (left) and active (right) neurons for FaceScrub and ImageNet, respectively.
Figure 8: Accuracy of ablating active and inactive neurons. Within-category accuracy after ablating inactive (left) and active (right) neurons for FaceScrub by changing the threshold of the proposed method by Goodfellow et al. (2009).
Figure 9: Invariance for different numbers of seen-transformed categories. Violin plots of the invariance coefficient among neurons at the penultimate layer for FaceScrub. Different numbers of seen-transformed categories are displayed and the invariance coefficient is reported separately for seen- and unseen-transformed categories.

D.3 Evaluation of the Method by Goodfellow et al. (2009). To assess the effectiveness of Goodfellow et al.’s (2009) method, we varied the firing threshold from 0.5% to 1% (as reported in the original study), 5%, 10%, 20%, and 40%, and measured the accuracies when ablating active and inactive neurons. The resulting accuracies are presented in Figure 8.

D.4 Invariance. In Figures 9 and 10, we display the invariance for the different transformations and number of categories seen-transformed in FaceScrub and ImageNet, respectively. Figure 11 shows the amount of invariance at each layer.
Figure 10: Invariance for different numbers of seen-transformed categories. Violin plots of the invariance coefficient among neurons at the penultimate layer for ImageNet. Different numbers of seen-transformed categories are displayed, and the invariance coefficient is reported separately for seen- and unseen-transformed categories.

Figure 11: Layer-wise invariance. Mean amount of invariance for different numbers of seen-transformed categories and transformations. Each row reports separately seen- and unseen-transformed categories, or both.

**D.5 Specialization.** To evaluate the selectivity of individual neurons toward transformed or non-transformed images, we modified equation 2.1 to compute specialization scores as \( \frac{f_k(T(x)) - f_k(x)}{f_k(T(x)) + f_k(x)} \). The resulting score indicates the degree to which a neuron is selective toward either transformed or nontransformed images. A score of +1 indicates high selectivity for the transformed images, whereas a score of −1 indicates high
Figure 12: Specialization in FaceScrub. Violin plots of the specialization coefficient among neurons in the penultimate layer. The specialization to each transformation is shown for networks trained with the transformation applied to all categories (indicated with $x + T(x)$) and for networks trained only with non-transformed images ($x$).

selectivity for the nontransformed images. The specialization scores of individual neurons are shown in Figure 12.

D.6 Inception. In Figures 13 and 14, we replicate the previously reported findings of AlexNet for FaceScrub using the Inception model. We
Figure 14: Invariance coefficients of individual neurons in FaceScrub using the Inception model. Replication of Figures 3 to 5 by the Inception model.

...observe a consistent pattern of results that supports our initial assertion that invariance is achieved gradually as the number of transformation-seen categories increases and occurs at the level of individual neurons.
Acknowledgments

We are grateful to Tomaso Poggio for his insightful advice and warm encouragement. This work is supported by the Center for Brains, Minds and Machines (funded by NSF STC award CCF-1231216), Fujitsu Laboratories Ltd. (contracts 40008401 and 40008819), the MIT-Sensetime Alliance on Artificial Intelligence, and the R01EY020517 grant from the National Eye Institute (NIH).

Code Availability

Code to reproduce the results is available in the following public GitHub repository: https://github.com/xboix/InvarianceRobustness.

References

Achille, A., & Soatto, S. (2018). Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research, 19*(1), 1947–1980.

Azulay, A., & Weiss, Y. (2019). Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research, 20*, 1–25.

Bao, P., She, L., McGill, M., & Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature, 583*(7814), 103–108. 10.1038/s41586-020-2350-5

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the Conference on Computer Vision and Pattern Recognition* (pp. 3319–3327).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron, 73*(3), 415–434. 10.1016/j.neuron.2012.01.010

Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., & Madry, A. (2019). Exploring the landscape of spatial robustness. In *Proceedings of the International Conference on Machine Learning*.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1*(1), 1–47. 10.1093/cercor/1.1.1

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proceedings of the International Conference on Learning Representations*.

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Gartnett (Eds.), *Advances in neural information processing systems, 31*. Curran.
Goodfellow, I., Lee, H., Le, Q. V., Saxe, A., & Ng, A. Y. (2009). Measuring invariances in deep networks. In Y. Benjio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), Advances in neural information processing systems, 22. Curran.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the Conference on Computer Vision and Pattern Recognition.

Jaderberg, M., Simonyan, K., Zisserman, A., & Kavukcuoglu, K. (2015). Spatial transformer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), Advances in neural information processing systems, 28. Curran.

Jang, H., McCormack, D., & Tong, F. (2021). Noise-trained deep neural networks effectively predict human vision and its neural responses to challenging images. PLOS Biology, 19(12), e3001418. 10.1371/journal.pbio.3001418

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream’s execution of core object recognition behavior. Nature Neuroscience, 22(6), 974–983. 10.1038/s41593-019-0392-5

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., . . . Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114 (13), 3521–3526. 10.1073/pnas.1611835114

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. Burges, L. Bottou, & K. Q. Weinberger (Eds.), Advances in neural information processing systems, 25. Curran.

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. Trends in Neurosciences, 23(11), 571–579. 10.1016/S0166-2236(00)01657-X

Le Grand, R., Mondloch, C. J., Maurer, D., & Brent, H. P. (2001). Early visual experience and face processing. Nature, 410(6831), 890–890. 10.1038/35073749

Montero, M. L., Ludwig, C. J., Costa, R. P., Malhotra, G., & Bowers, J. (2021). The role of disentanglement in generalisation. In Proceedings of the International Conference on Learning Representations.

Ng, H.-W., & Winkler, S. (2014). A data-driven approach to cleaning large face datasets. In Proceedings of the IEEE International Conference on Image Processing.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The building blocks of interpretability. Distill.

O’Reilly, R. C., Wyatte, D., Herd, S., Mingus, B., & Jilk, D. J. (2013). Recurrent processing during object recognition. Frontiers in Psychology, 4, 124.

Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. Nature Reviews Neuroscience, 5(4), 291–303. 10.1038/nrn1364

Peissig, J. J., & Tarr, M. J. (2007). Visual object recognition: Do we know more now than we did 20 years ago? Annual Review of Psychology, 58, 75–96. 10.1146/annurev.psych.58.102904.190114

Poggio, T., & Anselmi, F. (2016). Visual cortex and deep networks: Learning invariant representations. MIT Press.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. Nature, 435, 1102–1107. 10.1038/nature03687
Rainer, G., & Miller, E. K. (2000). Effects of visual experience on the representation of objects in the prefrontal cortex. *Neuron*, 27(1), 179–189. 10.1016/S0896-6273(00)00019-2

Riesenhuber, M., & Poggio, T. (1998). Just one view: Invariances in inferotemporal cell tuning. In G. Tesauro, D. Toretsky, & T. Leen (Eds.), *Advances in neural information processing systems*, 7. MIT Press.

Roelfsema, P. R., & de Lange, F. P. (2016). Early visual cortex as a multiscale cognitive blackboard. *Annual Review of Vision Science*, 2, 131–151. 10.1146/annurev-vision-111815-114443

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., & Madry, A. (2018). Adversarially robust generalization requires more data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems*, 31. Curran.

Schott, L., von Kügelgen, J., Traüble, F., Gehler, P., Russell, C., Bethge, M., . . . Brendel, W. (2021). *Visual representation learning does not generalize strongly within the same domain*. arXiv:abs/2107.08221.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, 171(3972), 701–703. 10.1126/science.171.3972.701

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. In I. Guyon, Y. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems*, 30. Curran.

Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: A better model of biological object recognition. *Frontiers in Psychology*, 8, 1551. 10.3389/fpsyg.2017.01551

Srivastava, S., Ben-Yosef, G., & Boix, X. (2019). Minimal images in deep neural networks: Fragile object recognition in natural images. In *Proceedings of the International Conference on Learning Representations*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–9).

Tacchetti, A., Isik, L., & Poggio, T. A. (2018). Invariant recognition shapes neural representations of visual input. *Annual Review of Vision Science*, 4, 403–422. 10.1146/annurev-vision-091517-034103

Tsao, D. Y., & Livingstone, M. S. (2008). Mechanisms of face perception. *Annual Review of Neuroscience*, 31, 411–437. 10.1146/annurev.neuro.30.051606.094238

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., & Wierstra, D. (2016). Matching networks for one shot learning. D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems*, 29. Curran.

Vogelsang, L., Gilad-Gutnick, S., Ehrenberg, E., Yonas, A., Diamond, S., Held, R., & Sinha, P. (2018). Potential downside of high initial visual acuity. *Proceedings of the National Academy of Sciences*, 115(44), 11333–11338. 10.1073/pnas.1800901115

Wang, J., & Perez, L. (2017). The effectiveness of data augmentation in image classification using deep learning. arXiv:1712.04621.

Wood, J. N., & Wood, S. M. (2018). The development of invariant object recognition requires visual experience with temporally smooth objects. *Cognitive Science*, 42(4), 1391–1406. 10.1111/cogs.12595
Wyatte, D., Curran, T., & O’Reilly, R. (2012). The limits of feedforward vision: Recurrent processing promotes robust object recognition when objects are degraded. *Journal of Cognitive Neuroscience, 24*(11), 2248–2261. 10.1162/jocn_a_00282

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision.*

Zhou, B., Sun, Y., Bau, D., & Torralba, A. (2018). Revisiting the importance of individual units in CNNs via ablation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Gartnett (Eds.), *Advances in neural information processing systems, 31.* Curran.

Received July 19, 2021; accepted August 7, 2023.