Learning online visual invariances for novel objects via supervised and self-supervised training

Valerio Biscione *, Jeffrey S. Bowers

Department of Psychology, University of Bristol, Bristol, BS8 1TL, United Kingdom

ABSTRACT

Humans can identify objects following various spatial transformations such as scale and viewpoint. This extends to novel objects, after a single presentation at a single pose, sometimes referred to as online invariance. CNNs have been proposed as a compelling model of human vision, but their ability to identify objects across transformations is typically tested on held-out samples of trained categories after extensive data augmentation. This paper assesses whether standard CNNs can support human-like online invariance by training models to recognize images of synthetic 3D objects that undergo several transformations: rotation, scaling, translation, brightness, contrast, and viewpoint. Through the analysis of models' internal representations, we show that standard supervised CNNs trained on transformed objects can acquire strong invariances on *novel* classes even when trained with as few as 50 objects taken from 10 classes. This extended to a different dataset of photographs of real objects. We also show that these invariances can be acquired in a self-supervised way, through solving the same/different task. We suggest that this latter approach may be similar to how humans acquire invariances.

Crown Copyright © 2022 Published by Elsevier Ltd. All rights reserved.

1. Introduction

Humans can identify objects despite the variable images they project on our retina, including variation in image size, orientation, illumination, and position (Tanaka, 1996). Critically, this invariance is computed 'online': a transformed, unfamiliar object can often be recognized after a single presentation at a given pose. For example, after identifying a novel object that is projected at a given retinal location we can immediately identify the object across a wide range of retinal locations (Blything et al., 2020, 2021). How the visual system succeeds under these conditions is still poorly understood, but we know that this task is solved through hierarchical processing along the ventral stream that ends at the inferotemporal cortex, where the activation of neural populations are largely independent to object transformation (Bar & Biederman, 1999).

Recently, Convolutional Neural Networks (CNNs) have been proposed as a model for the human visual system (Khaligh-Razavi & Kriegeskorte, 2014; Schrimpf et al., 2018; Yamins & DiCarlo, 2016). This is currently a topic of intense debate. On one hand, CNNs' internal activations are predictive of brain recordings in response to images taken from classic benchmark datasets in the mid and high levels of the ventral visual pathway (Cichy et al., 2016; Yamins et al., 2013, 2014). On the other hand, CNNs show

some behavioral discrepancies that challenge their plausibility as a model of the visual system: they are susceptible to adversarial attacks (Dujmović et al., 2020; Szegedy et al., 2013); they are highly susceptible to low amount of image degradation (Geirhos et al., 2018); they often classify images based on textures instead of shape (Geirhos et al., 2018) and local instead of global features (Baker et al., 2018; Malhotra et al., 2021), do not account for humans' similarity judgments of 3D shapes (German & Jacobs, 2020), and do not exhibit human-like Gestalt properties (Biscione & Bowers, 2022). Within the framework of this debate, we aim to answer the question of whether CNNs can learn to support online invariance for a wide variety of transformations commonly experienced in the human visual environment. This would help us better understand the limits and the possibilities of using CNNs as a model of the human visual system.

When comparing CNNs to human vision, it is important to distinguish between *online* transformation invariance and *trained* transformation invariance (Fig. 1). Trained invariance refers to the ability to classify novel exemplars of objects from *trained* classes in trained locations (e.g., invariance to the pose of a specific image of a dog after exposure to multiple poses of other exemplars of dogs). This is the standard approach in training CNNs, and because training is rarely performed through the lens of human cognition, it often includes augmentations that are unlikely to be experienced by humans (e.g. color jittering) as well as exclude common ones (e.g. change in viewpoint). Accordingly, the trained invariance commonly reported in CNNs is qualitatively

---

* Corresponding author.
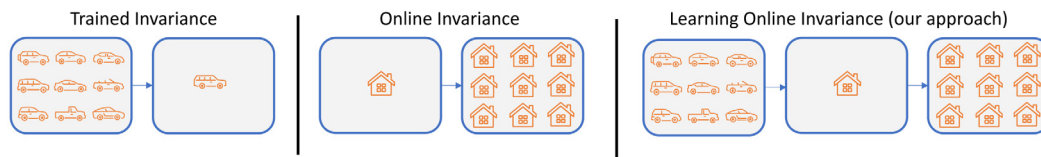*E-mail address:* valerio.biscione@bristol.ac.uk (V. Biscione).

Fig. 1. Illustration of trained and online invariance and our proposal for learning online invariances. In this figure, the transformation we depict is translation, but the general logic applies to any transformations. **Left**: In trained invariance, after a model is trained on some classes of objects across different locations it can recognize novel instances from those same classes across different locations. This property is commonly obtained in CNNs through extensive data augmentation. **Middle**: In online invariance, a novel object can be instantly recognized across several transformations, after have experienced it at one location. Neural networks fail to show this property (see Section 1.1). **Right**: In our approach, a model is trained on one dataset of transformed objects in order to acquire online invariance to novel classes.

different from human *online* invariance that extends to *novel objects from unseen classes* following exposure to only one or a few poses (Bowers et al., 2016). The degree to which CNNs are invariant to online transformation is unclear.

In this work we aim to close this gap by testing, with a variety of methods, whether CNNs support online invariance for novel objects from novel classes even when the novel classes come from a new dataset. We only trained and tested models using naturalistic transformations, namely, scale, rotation, translation, brightness, contrast and change in viewpoint. Particular attention is given to the latter as it is not a simple transformation of a 2D image and thus is not typically used when performing data augmentation. We show that common CNNs models can acquire, to a strong degree, all the above invariances, even when trained with as few as 50 images of objects taken from 10 categories. Critically, these models not only support invariance for trained object classes, but also for untrained object classes. We find this following supervised learning, and also in a self-supervised manner by classifying pairs of objects as the same or different. We argue that this latter approach is more psychologically plausible, and discuss the implication in terms of human psychology in Section 4.2.

1.1. Related work

The distinction between trained and online invariance has been the focus of recent work on translation invariance (recognizing an object at novel retinal locations after having seen it at one location). Behavioral studies have highlighted the extent to which humans possess online translation invariance (Blything et al., 2020, 2021), and it is often claimed that convolutional and pooling operations endow CNNs a similar degree of online invariance to translation (LeCun & Bengio, 1995; Marcus, 2018). However, several experiments have shown that this is not the case, and indeed, CNNs not only fail to support online invariance to translation, but also fail to support online invariance to scale, rotation, and flipping along the vertical and horizontal axis (Blything et al., 2021; Chen et al., 2017; Gong et al., 2014; Kauderer-Abrams, 2017). One approach to achieving online invariance in CNNs is through architectural modification: for example, adding a Global Average Pooling layer to the end of the convolutional block results in complete translation invariance (Blything et al., 2021). Additional architectural modifications have been introduced in order to support other types of transformation. This has resulted in a proliferation of modified models that accounted for either individual transformations (e.g. Han et al., 2020; Xu et al., 2014 for scale; Kim et al., 2020; Marcos et al., 2016 for rotation) or multiple transformations at the same time (Cohen & Welling, 2016). Although these models may prove valuable in terms of technological advancements in the field of object recognition, the architectural modifications are not guided by any concern of biologically plausibility. And more importantly, a key assumption of these approaches is that standard CNNs architectures are incapable of supporting online invariances. If

it turns out that some training regimes allow standard CNNs to support online invariances, the introduction of additional architectural (innate) mechanisms may not be required either for engineering or psychological considerations.

Biscione and Bowers (2020) and Blything et al. (2021) first reported examples of a classic CNN model (VGG16), without any architectural changes, exhibiting complete online invariance to translation when the model was pretrained on the ImageNet dataset. That is, a network pretrained on ImageNet supported translation invariance to other, very different datasets (e.g. MNIST, EMNIST, and others). A key feature of the pretraining was that it consisted of resizing and crop augmentations, which indirectly resulted in translated versions of the same images. That is, these findings lend support the hypothesis online invariance can be trained by using a dataset of translated samples. Biscione and Bowers (2021) further investigated this approach by pretraining several convolutional architectures on a variety of datasets that differed in the complexity of the translated stimuli. They observed that training on simple translated datasets was often enough to support online translation invariance on other, *novel* classes, sometimes very different from the trained ones.

Overall, a review of the literature points to the following facts: CNNs are not, by design, architecturally invariant to any tested transformation (scale, rotation, translation). Many *ad hoc* architectural changes can be employed to provide invariance for some transformations. Only recently it has been found that pretraining standard CNNs (without any architectural modification) with translated objects endows them with online translation invariance (that is, translation invariance for classes *different* than the pretrained ones). However, it has not yet been explored to what extend this technique can be employed across transformations other than translations and across different networks architectures. The main goal of the present work is to fill this gap by testing a large set of transformations (in fact, a set covering all possible transformations in 3D space) on many networks, both supervised and a self-supervised.

1.2. Outline of the current work

In this work, we extend the work of Blything et al. (2021) and Biscione and Bowers (2020, 2021), and show that: (1) online invariance to many different types of transformations can be acquired with no architectural change by pretraining the network on the same transformations assessed at test; (2) many different supervised networks acquire this property, as does a self-supervised network that solves the same/different task. We argue that the latter task provides a more psychologically plausible approach to learning invariances in infants.

We tested 7 supervised networks (AlexNet, VGG11, VGG19, Resnet-18, Resnet-50, GoogLeNet, Densenet-201) and one self-supervised network (detailed in Section 2.3) on 6 invariances: translation, rotation, scale, brightness, contrast, and viewpoint. Although this final transformation does not belong to standard data augmentation technique (in that it is not a transformation

of a 2D image), it is a psychologically relevant transformation, normally experienced by an observer in a 3D world. In order to obtain different viewpoints samples, we generated a new dataset based on ShapeNet (Section 2.2).

Our general approach consisted in training each network on 10 classes from the ShapeNet dataset, for each transformation separately, and without any transformation. We then tested on 10 novel classes either from ShapeNet or from a dataset of photographed objects, ETH-80 (Leibe & Schiele, 2003). A network trained with a certain transformation was tested on the corresponding transformation (e.g. a network trained with translated ShapeNet objects was tested on translated objects from novel classes). The networks trained without any transformation were tested across all transformations. With the self-supervised network, network performance on novel classes can be tested directly without having to retrain the model (Section 3.1). For the supervised networks, assessing the amount of online invariance on novel classes through the classification performance requires retraining on the novel classes. This in turn could result in catastrophically forgetting the acquired invariances (Biscione & Bowers, 2021 as discussed in more detail in Section 3.2). Therefore we assessed the amount of online invariance with two approaches that did not require retraining: a 5-alternative forced choice task, in which the internal representation was used to identify which transformed object, among five candidates, was the same as the target object (Section 3.2.1); a direct measure of the internal representation similarity between transformed and un-transformed samples (Section 3.2.3). These tests were always performed on novel classes. In Section 3.2.2, we also ran a cross-transformation analysis, in which we tested whether pretraining on one transformation afforded invariance to any other transformation. In Section 3.3 we investigated how the number of trained objects impacted the acquisition of invariances to transformations on novel classes of objects. A summary of the result is provided in Section 3.4. Finally, the limitation of the current method and the implications of our findings for Machine Learning and Psychology are discussed in Section 4.

2. Methodological overview

2.1. Transformation scheme

We explore whether each model can learn each of the 6 invariances by applying a corresponding stochastic transformation scheme $T_{t,\theta}$ to each object, where t identifies a transformation with parameters θ (e.g. $t = rotation$ and $\theta = [-180, 180]$ indicates a random rotation on the image plane between $-180°$ and $180°$. When $t \neq viewpoint$, the object is presented at an inclination of $80°$ and azimuth of $36°$. The parameters θ for each transformation $t = \{rotation, scale, translation, brightness, contrast, viewpoint\}$ are shown in Fig. 2B.

2.2. Multi-viewpoint dataset

In order to investigate whether invariance to viewpoint can be acquired we generated a novel 2D dataset based on ShapeNet (Chang et al., 2015), a dataset containing 50300 3D objects across 55 categories. Each 3D object was rendered textureless on a uniform black background as a $128 \times 128 \times 3$ image. To obtain variation in viewpoint, we placed a camera on a sphere at an inclination ranging from $30°$ to $110°$ ($10°$ intervals) and an azimuth covering the whole sphere ($36°$ intervals, Fig. 2C), generating 90 different viewpoints per object. In the following sections the term "object" refers to any image of a specific 3D model regardless of viewpoint or 2D transformations applied to it. Notice that this approach differs greatly from other approaches in which ShapeNet's 3D objects are used in networks that directly consume point clouds (e.g. PointNet, Qi et al., 2016).

2.3. Same/Different task approach

Consider a set of objects $O = \{o_n\}_{n=1}^N$ subjected to a transformation scheme $T_{t,\theta}(o_n)$ which stochastically produces an image x_n^i. We define $X_n = \{x_n^i\}_{i=1}^M$ as a set of images resulting from applying M times the transformation scheme $T_{t,\theta}$ to the object o_n. We furthermore define an embedding module $g_\psi(\cdot)$ which produces a compact representation z_n^i of an input image x_n^i. Pairs of inputs $\{x_n^i, x_m^j\}$ are build such that either they are from the same set X_n and thus $n = m$ (they represent same object o_n) with probability p or they are from two different sets so that $n \neq m$ (they represent different objects) with probability $1 - p$. The pair embeddings $\{z_n^i, z_m^j\}$ are aggregated through $a(\cdot, \cdot)$ and fed into $h_\phi(\cdot)$, as a non-linear function approximator parameterized by a learnable weights ϕ which returns a dissimilarity score $\hat{y}$. We use mean square error (MSE) for the loss, matching the dissimilarity score with object identity (Sung et al., 2018):

$$\mathcal{L}(\hat{y}_{n,m}) = \sum_{}^{N} (\hat{y}_{n,m} - \mathbf{1}\{n \neq m\})^2$$

This is minimum when the $\hat{y} = 1$ for different objects and $\hat{y} = 0$ for the same objects (see Fig. 3).

In particular, the embedding module g_θ is a copy of VGG11 in which the last fully connected layer (the classifier) is replaced by another fully connected layer which output is an embedding of 512 units + Sigmoid activation function. As an aggregation function we use $a(z_n^i, z_m^j) = |z_n^i - z_m^j|$. The dissimilarity module h_ϕ is a single fully connected layer with a single output unit + Sigmoid. The probability p of a pair being composed of the same objects was set at 0.5.

Our approach is related to self-supervised learning, in which a system is trained with unlabeled data on a surrogate task to acquire features that can be used on a downstream task (Jing & Tian, 2019). For example, SimCLR, (Chen et al., 2020), and Relational Reasoning Network (Patacchiola & Storkey, 2020) use mini-batches of augmented samples and map together representations of different augmentations of the same image and separate the representations different samples. One of the main differences between our self-supervised approach on one hand, and SimCLR and Relational Reasoning Network on the other, is that we use a 3D model instead of a 2D image to generate data augmentations, and thus we can consider augmentations that are normally excluded from other models, that is viewpoint. This could be expanded to changes in lightning, textures, and other material/environmental properties of the 3D object, allowing us to train models on much more naturalistic data. However, our aim is not to compare our implementation with similar contrastive approaches or to obtain a state-of-the-art accuracy, but to show as a proof of concept that a simple self-supervised network can indeed acquire these invariances.

2.4. Supervised approach

Supervised training was arranged as usual: considered a set of categories $C = \{c_k\}_{k=1}^K$ each containing a number of objects $O_k = \{o_{n,k}\}_{n=1}^N$ subjected to a transformation scheme $T_{t,\theta}(o_{n,k})$ which stochastically produces an image $x_{k,n}^i$. In a supervised approach, the identity of an object n is discarded and only the class k is considered during training. A non-linear function approximator $f_\phi(x_k^i)$ parameterized by learnable weights ϕ is trained to minimize the error in predicting the output class k. For f_ϕ we consider a variety of Convolutional Neural Network: Alexnet, VGG11, VGG19, Resnet-18, Resnet-50, GoogLeNet, Densenet-201, covering a wide range of feedforward architectures. Each network was trained through stochastic gradient descent on the cross-entropy loss.
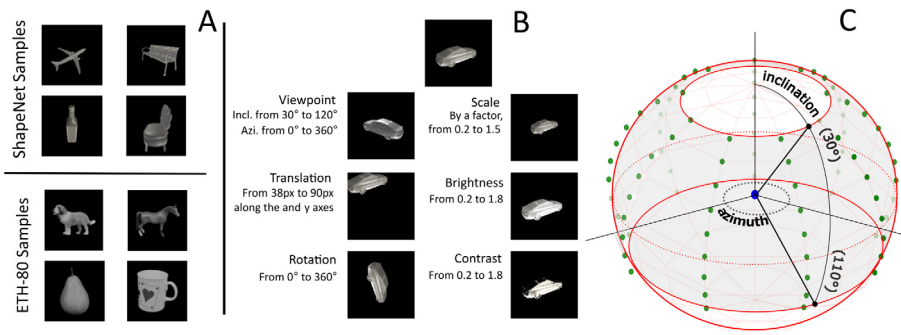
Fig. 2. A: Some samples from the ShapeNet and ETH-80 datasets (from one viewpoint and without any 2D transformations). **B**: Example of different samples resulting from stochastically applying a transformation to the 3D object, and the parameters used for each transformation. **C**: Representation of the method used for generating different objects' viewpoints. The cameras (dots) were placed at different locations around a sphere. Each camera pointed at the center of the sphere, where the 3D object was placed (central dot). The cameras were only placed between an inclination of 30° to 110°, and covered the whole sphere longitudinally (shaded area).
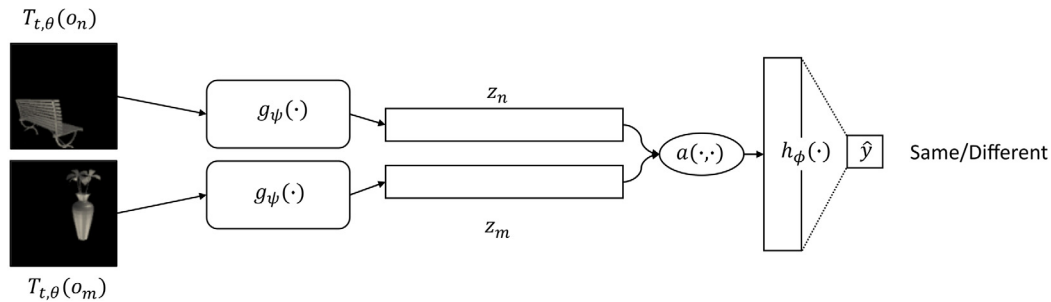

Fig. 3. Overview of the proposed Same/Different Network. See text for details.

3. Experiments and results

We assessed trained and online invariance in six different supervised networks and in the self-supervised same/different network. We trained all networks on the same samples obtained from the ShapeNet objects dataset: we randomly sampled 250 3D objects from 10 classes, totaling 2500 objects. The classes were randomly selected to be: bench, bottle, bus, clock, faucet, jar, knife, laptop, rifle, table. Each model was trained on two conditions: (1) with each transformation (e.g., translation, rotation, etc.) applied separately, and (2) without transformations. This results in models in this latter condition being trained on less samples (as the objects used in this condition were not transformed) but this does not affect the results, as the analysis in Section 3.3 shows. We used Adam optimizer with learning rate 10^{-3}, batch size of 64. We trained until convergence, stopping when the exponential moving average of the loss (with $\alpha = 0.1$) computed on the training set did not decrease by at least 0.01 for 250 mini-batch iterations. We normalized the input images between -1 and 1 by pre-computing the mean and standard deviation for a large subset of the ShapeNet/ETH-80 datasets. Each experiment is repeated over 3 random seeds.

To measure the degree of online invariance we selected 3D models from 10 novel classes from ShapeNet: airplane, bathtub, car, mobile-phone, chair, guitar, lamp, pot, sofa, vessel. We also tested online invariance on a new dataset, ETH-80, which includes photographs of 80 real life objects along 8 classes (apple, car, cow, cup, dog, horse, pear, tomato), each object photographed from 41 viewpoints, which made it ideal for our viewpoints tests (Fig. 2). The background was removed from these images, and they were converted to grayscale values. We excluded the class "tomato", as it was almost indistinguishable from the class "apple" when converted to grayscale.

3.1. Invariance in the Same/Different network

In the case of the Same/Different Network, invariance is manifest when the model outputs a low dissimilarity score to transformations of the same object and a high score to different objects, and online invariance can be directly measured by presenting novel classes. Each version of the same/different network was tested on the same transformation it was trained on (apart for the networks trained without any transformations, which were tested on each transformation separately). With both novel ShapeNet classes and ETH objects, the performances of the networks trained on transformations was consistently better than the network trained on untransformed objects, with performance > 90% in all conditions other than viewpoint (Table 1). Testing on the ETH dataset resulted in a drop in accuracy of around 4% for all but the viewpoint transformation, which dropped by $\sim$ 16%. This is still noteworthy considering the difference in appearance between the trained ShapeNet objects and these ETH objects and the low variability within ETH classes (e.g. when a pair is made of two different objects from the same class, such as two mugs, the low inter-class variability makes it more difficult to classify the objects as different). Overall, the network shows strong online invariance for novel classes and novel datasets.

3.2. Invariance across all models

Assessing online invariance on supervised networks presents special challenges. One approach consists of retraining the networks on un-transformed samples from novel classes, and then measuring the classification performance on the transformed version of the same classes. For example, to assess translation invariance, a network could be trained on samples from novel classes, always presented at the center of the canvas, and then tested on the same classes presented at different locations. The resulting

Table 1
Accuracy for the Same/Different Network on novel ShapeNet classes and ETH-80 classes.

		Transformations on test					
		Viewpoint	Translation	Rotation	Scale	Brightness	Contrast
ShapeNet → ShapeNet	Trained with transformations	82.65 ± 0.75	95.02 ± 0.88	96.61 ± 0.49	95.06 ± 0.35	98.54 ± 0.36	97.85 ± 0.39
	Trained without transformations	50.85 ± 0.59	50.56 ± 1.35	50.58 ± 1.41	50.68 ± 1.37	50.01 ± 1.19	52.17 ± 1.19
ShapeNet → ETH-80	Trained with transformations	66.67 ± 3.56	93.33 ± 1.78	92.85 ± 3.08	92.85 ± 4.04	94.76 ± 3.56	93.33 ± 3.56
	Trained without transformations	53.81 ± 2.43	51.42 ± 5.83	51.90 ± 5.87	53.33 ± 5.51	54.76 ± 4.71	54.28 ± 4.04

classification accuracy would correspond to the degree of online invariance to translation. The problem with this approach is that it requires retraining on novel classes, which might result in the catastrophic forgetting of the acquired invariances (French, 1999). This means that even if the networks did acquire the invariance from the transformed dataset, we would not be able to observe it through the classification performance alone (Biscione & Bowers, 2021). This point will be further discussed in Section 4.1.

Following Biscione and Bowers (2021) we opted to assess online invariance by measuring the similarity of internal representation between transformed and untransformed versions of the same objects in both the same/different and supervised networks using cosine similarity:

$$C^l(\mathbf{a}, \mathbf{b}) = \frac{d^l(\mathbf{a}) \cdot d^l(\mathbf{b})}{\left\| d^l(\mathbf{a}) \right\| \left\| d^l(\mathbf{b}) \right\|},$$

where $d^l(\mathbf{a})$ is the activation at layer l given input vector $\mathbf{a}$. We then assessed invariance in two ways, namely, in a choice task in which similarity scores were used to classify objects, and a normalized cosine similarity score that accounted for within vs. between objects cosine similarity scores. We consider these two measures in turn.

3.2.1. 5-Alternative forced choice task

We tested each network on 5-alternative choice task (5AFC). The cosine similarity between a target object and 5 candidates was computed, setting one candidate to be the same as the target (Fig. 4, top). All objects were subjected to random transformation, and each network was tested on the same transformation it was trained on. Objects were randomly sampled from the unseen classes. We ran 100 trials for each condition, and defined as "correct" a response trial in which the cosine similarity at the last layer was higher for the transformed version of the target object. For the Same/Different task, the cosine similarity was computed on the embedding vector. The results are shown in Fig. 4, bottom-left, together with the results of the same test applied on a network trained on un-transformed samples (black circles in the plot). The results indicate high performance in all transformation conditions, with viewpoint being the most challenging. That is, all models acquired an impressive degree of online invariance. The 5AFC test repeated on the ETH objects resulted in a slight drop in accuracy (by ∼ 6%), with the viewpoint transformation suffered the most extreme drop in accuracy (lowered to ∼ 60%, and thus still remarkably higher than chance) compared to the ShapeNet objects. Therefore, even with this new dataset, accuracy was always significantly higher with network trained on transformed samples than pretrained on un-transformed samples (see Appendix A).

3.2.2. Cross-transformation tests

One possible interpretation of these findings is that the online invariance for a given transformation type emerges in response to training on the corresponding transformation, and another is that training on any invariance will improve all forms of online invariance. To test this, we performed cross-transformations tests: a network trained with a transformation t was subjected to a choice-test with a different transformation applied. For the viewpoint, translation, rotation, and scale, the obtained performance is similar to the performance of an un-transformed network (Fig. 4, bottom-right panel). On the other hand, brightness and contrast generalized to each other, and only to each other: for example, training translation did not increase invariance on either brightness or contrast.

3.2.3. Internal representations analysis

As a further metric of transformation invariance, we directly explored how the internal representations of an object changed when transformed compared to a base view. Consider $\tau_t^\theta(o_n)$ as a specific transformation t with parameter θ of object o_n, (e.g. rotation of the object o_n by 60°). Let us define a set of baseline parameters α for each transformation t: 1 for scale, brightness, and contrast; 0° for rotation, (64, 64) pixels for translation (center of the canvas), inclination = 80° and azimuth = 36° for the viewpoint transformation. Therefore, for each network trained on a transformation $T_{t,\theta}$, and a set of randomly sampled objects $\{o_1, \ldots, o_R\}$ from novel classes, we defined $I_t(\theta)$ (for *Invariance*) as the average similarity of representations between the objects base-view and their transformations τ_t^θ. A network could obtain a high Invariance score by simply collapsing all representations together, regardless of variability across objects or class identity. The model could be trivially invariant to object transformation by being invariant to any object feature. We accounted for this possibility by measuring the variability across *different* objects: we define $U_t(\theta)$ (for *Uniformity*) as the between-objects invariance, that is the total average similarity across different randomly sampled objects $u = \{u_1, \ldots, u_n\}$ and $v = \{v_1, \ldots, v_n\}$:

$$I_t(\theta) = \frac{1}{R} \sum_r^R C^l(\tau_t^\alpha(o_r), \tau_t^\theta(o_r)) \quad \text{and}$$

$$U_t(\theta) = \frac{1}{N} \sum_n^N C^l(\tau_t^\alpha(u_n), \tau_t^\theta(v_n)).$$

We used $U_t(\theta)$ as a baseline invariance across objects: if a model has learnt to be non-trivially invariant to transformation t, $I_t(\theta)$ should be higher than $U_t(\theta)$. Therefore we defined the Adjusted Invariance Metric as

$$\tilde{I}_t(\theta) = \frac{I_t(\theta) - U_t(\theta)}{1 - U_t(\theta)},$$

which returns the adjusted invariance for each transformation by parameter θ. A value of 1 indicates identical representation between transformed version of the same objects (same internal representation across transformed versions of the same objects, and different internal representations of different objects); a value of 0 indicates that two different transformations are as different as two different objects.

All objects used in this analysis belong to novel, untrained classes from ShapeNet. The results (Fig. 5) clearly show that, when trained and tested on corresponding transformations, each model can acquire a strong invariance to that transformation, even for novel objects and classes. This is not the case with
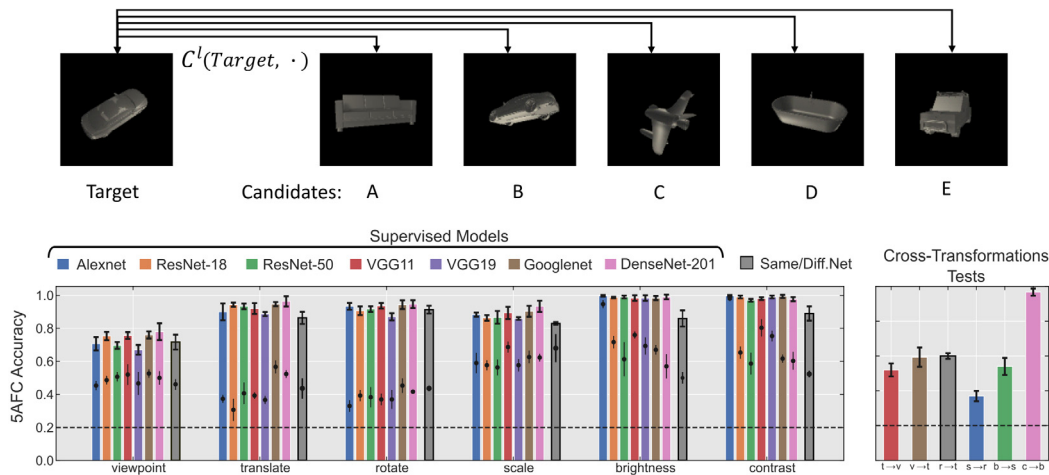
Fig. 4. Top: Illustration of the 5-alternative forced choice task. One object is compared with 5 alternatives in which one of them is the same object (in the picture, Candidate B). All samples are subjected to a transformation (that varied by condition). We used the cosine similarity of the internal representations across pairs of Target and Candidates as a metric of choice. **Bottom-left**: Results in terms of accuracy for each transformation and each model. The bar represents the accuracy for a network trained on the same transformation it is been tested on. Compare this with the black circle inside each bar which represents the accuracy for a network trained on un-transformed objects. **Bottom-right**: accuracy when a network trained on one transformation is tested on a different transformation. Apart from brightness and contrast, training on one transformation did not generalize to other transformations. Overall, these plots show a high degree of learnt online invariance to a transformation when the network is trained on the same transformation (the classes used for the 5AFC were not used during training).
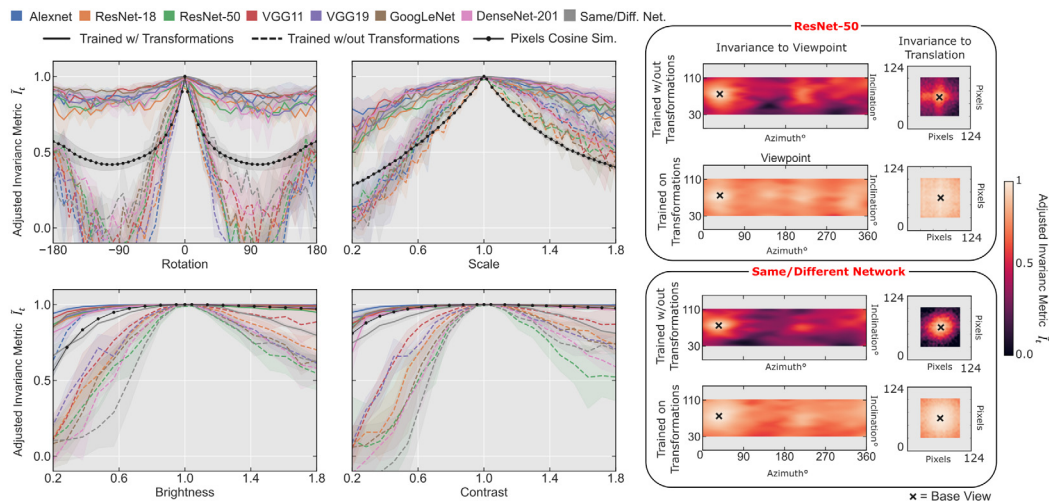

Fig. 5. Invariance Representation Metric between a base view of an object and its transformed versions. The networks trained with transformations (continuous lines) have a high invariance across most levels of each transformation, whereas for networks trained on un-transformed objects (dashed lines) the similarity sharply decreases even at low levels of transformations. The results in the four panels on the **left** are averaged across 200 random samples from untrained classes. The shaded area represents the standard deviation across the 3 seeds for each network. The black line with circular markers represents the cosine similarity at the pixel level at different transformations. The **right** panels show the results of viewpoint and translation invariance for two networks (other networks are shown in Appendix B). The results were consistent across all tested networks.

networks not trained on transformed samples, as their representations strongly depend on the object pose. There does not seem to be any relevant difference amongst models, either supervised or self-supervised. The analysis of $U_t(\theta)$ (in Appendix B.1) also shows that the networks not trained on transformed samples collapse representations of different, transformed objects together. For these models, a car rotated 90° is very similar to a horse rotated 5° (however, un-transformed objects were still clearly separated from different, un-transformed objects).

3.3. Influence of number of objects trained

In the preceding experiments, each network reached a high level of invariance following training on 250 objects per class. We

explored how invariance changed as a function of the number of trained objects. For each network, we repeated the previous training with $N = \{5, 50, 100, 500\}$ where N is the number of objects per each category, and computed the results on the 5AFC task. Surprisingly, there does not seem to be any effect on the number of trained objects with the transformed networks (Fig. 6): as few as 5 objects per classes (with a total of 50 transformed objects trained on) were enough to acquire strong invariances. The plot also shows that the invariance acquired was not due to the higher number of samples experienced by the network trained with transformations: increasing the number of objects (and thus the number of different samples) did not improve online invariance on networks trained on un-transformed samples. In fact, it appeared to decrease it in some cases.
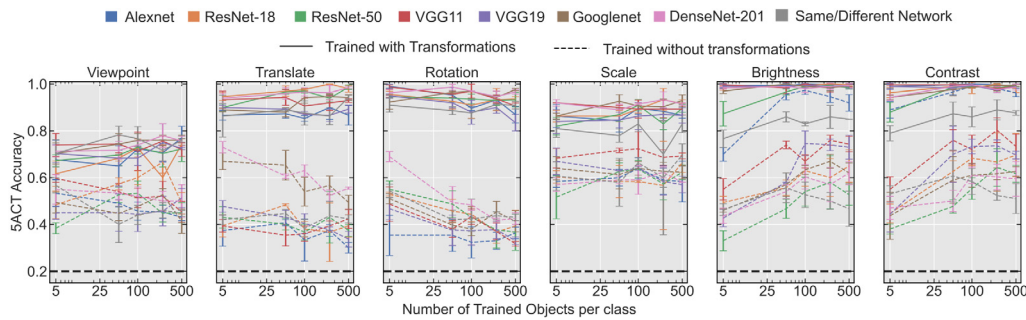
Fig. 6. Effect of training on different number of objects on the learnt invariance. There is no effect on the network trained on transformed objects. However, there seems to be some effect on network trained on un-transformed objects, but the direction of this effect is not consistent across transformations.

3.4. Summary of the results

CNNs architectures can learn to support online invariance for scale, rotation, translation, brightness, contrast, and, to a lower amount, viewpoint, when trained on an appropriately transformed dataset. The findings are robust across a variety of architectures, learning setups (supervised and self-supervised), and training size (trained on as few as 50 objects taken from 10 classes).

In the 5 alternative forces choice task (Section 3.2.1) all networks performed at ∼ 90% accuracy for all transformations but viewpoint (∼ 70% across all networks), when tested on novel classes from ShapeNet. When tested on a novel dataset (ETH-80), accuracy dropped only slightly for all conditions but viewpoint invariance, which dropped to ∼ 60%, still remarkably higher than chance. Overall, viewpoint invariance appeared to be the most difficult transformation to generalize across novel classes. Similar results are obtained by analyzing the networks' internal representations (Section 3.2.3). We also verified that, in order to acquire invariance to a certain transformation, the networks needed to be trained specifically on that transformation, and that training on other transformations will not suffice (with the exception of contrast and brightness, which would generalize to each other, Section 3.2.2). Similar results are obtained when varying the number of trained objects, and networks were able to acquire all invariances following training on as few as 50 objects taken from 10 classes (Section 3.3).

Overall, we found that training a network on a transformation on one set of objects allowed CNNs to support the corresponding transformation for novel objects taken from novel classes. For example, we showed that training rotation invariance on images of synthetic 3D buses supports rotation invariance on images of synthetic 3D chairs. We show that this extends to novel classes from the same synthetic dataset and on a different dataset of real, photographed objects (ETH-80).

4. Discussion

The finding that a range of visual invariances can be acquired using a general learning mechanism is significant both for the machine learning community and for the psychological literature. We discuss both points in turn.

4.1. Significance for machine learning

As outlined in Section 1.1, the common approach to achieve strong invariance to various transformations is to apply architectural modifications to standard CNNs. Some of these modifications afford invariance to only one transformation (e.g. Han

et al., 2020; Xu et al., 2014 for scale; Kim et al., 2020; Marcos et al., 2016 for rotation), others to multiple transformations (Cohen & Welling, 2016). Our work shows that these modifications may not be necessary to acquire invariances. Indeed, we found through measuring internal representation that standard CNNs show strong invariances to a range of transformations following the appropriate training. Furthermore, the current approach allows the possibility of acquiring invariances that are not limited to 2D affine transformation (as in the architectural approach) but can be extended to all types of variations: we have shown here viewpoint variation, but it is possible to design a dataset to induce invariance to background, clutter, texture, etc. It is not clear how to modify a network's architecture to innately capture these invariances.

However, it is important to emphasize that these learnt latent invariances will not always manifest themselves in performance. For example, imagine that we want to perform classification on some classes, e.g. different breeds of dogs, and for computational reasons we only want to train with the dogs at one location and wish the network to be able to classify the dog at any location (that is, we want the network to possess online invariance to translation). With our approach, we would need to pretrain the network on a translated dataset, and then retrain the network on our dataset of dogs. The problem of this approach is that the re-training session could result in catastrophic forgetting (French, 1999; McCloskey & Cohen, 1989), a phenomenon in which acquired capability rapidly degrade as the system learns a new task. In our cases, it means that even though CNNs can acquire online invariances, and these can be generalized on object from *novel* classes (as we have shown through internal representation analysis), *retraining* to perform classification could result in losing the learnt invariance.

Still, there are reasons to think this limitation can be overcome. Biscione and Bowers (2021), in the context of translation invariance, found that learnt invariances have higher chance to be retained if the re-training dataset is less complex than the pretraining one. Furthermore, there is ongoing work in overcoming catastrophic interference through interleaved training (Schaul et al., 2015), selective plasticity (Beaulieu et al., 2020; Fernando et al., 2017; Rusu et al., 2016), and by incentivizing sparse or disjoint representations (French, 1999; Liu et al., 2018). Even though the current applicability of this approach *as-is* in classification is limited, we suggest that promising advances in the field of catastrophic forgetting could provide a possible path for making use of the learnt invariances and thus complement or replace the current architectural approach. Our work highlights how the solution to catastrophic interference may have implications for how CNNs can solve various online invariances.

We also note that the current results appear to be inconsistent with two recent findings. Hernandez-Garcia and Konig (2019) found low similarities of the latent embeddings of transformed objects for a network trained on augmented CIFAR10 (with random affine transformations, brightness and contrast). We believe that this is due to their reliance of Euclidean distance which underestimated the degree of scale invariance and subsequently (since the transformations were aggregated) the strength of all invariances. Also Xu and Vaziri-Pashkam (2021) found inconsistent embeddings across classes of translated objects, which we believe to be due to the different training setup they employed. We expand on this in Appendix C.

4.2. Significance for human psychology

An important observation regarding invariances in humans is that they exist from an early age: rotation (Schwartz et al., 1979) and scale (Day & McKenzie, 1981) invariance was found in infants at 3–4 months; viewpoint invariance seems to be present in infants as young as 3 months old (Bornstein et al., 1986; Kraebel & Gerhardstein, 2006).

How these invariances are acquired, however, is not clear. Our work raises two points. Firstly, the proliferation of architectural modifications to obtain online invariances in CNNs (see Section 1.1), even though developed without any psychological considerations, suggests that "innate" mechanisms for invariances are needed in this context, which further suggests that a form of innate architecture is needed in human vision. In fact, we show that special purpose architectures are unnecessary, and learning invariances to object transformation can be achieved by a generic mechanism that performs feature learning on 2D samples. This may also apply to humans.

Secondly, our results with the same/different network raise the possibility that the learning mechanism responsible for acquiring these invariances could be based on the evaluation of the "sameness" of objects following changes in their retinal projections due to transformation. This is consistent with the analogical reasoning research literature that suggests the ability to form abstract relationship amongst objects is linked to the ability to solve the same/different tasks (Gentner et al., 2021; Premack, 1983). This skill is acquired during the first months, fully developed in 7-months old infants, and strongly linked to higher form of abstractions (Hespos et al., 2021). Importantly, this approach does not require the complex process of object categorization through supervised learning that does not apply at these early ages. However, as shown by Puebla and Bowers (2021), modelling the same/different task with CNNs will require a deeper understanding of the conditions under which this ability can be extended to unseen visual environments.

Of course, the same/different task still requires the information of whether the two objects are in fact the same or not, and this might be provided by a variety of mechanisms that exploit the temporal continuity of a sequence of retinal projections from objects in the world following spontaneous movements (e.g. eyes or head movement) and, later in the developmental stage, the active manipulation of objects while viewing them. It is important to acknowledge that our training regime sidesteps some important issues, notably, the specific mechanisms for telling the network that two images are the same or different. Nevertheless, our findings show that the same/different signal is sufficient for CNNs to support online invariance to transformations, and this should motivate more research into how the infant brain might extract this signal in order to learn online invariances to various transformations.

Using CNNs as a models for the human brain is an ambitious project. Thus far, the claim that CNNs provide a promising model of human vision has largely been supported by statistical measures of similarity between unit activations in models and neural populations in visual cortex (Zhuang et al., 2020). We believe that, for CNNs to be useful models of the visual stream, they also need to possess a set of fundamental functional properties that characterize human vision. Our high level contribution is to show that one of these properties, invariance to object transformations, that is not present in the system by default, can be acquired, even in a self-supervised manner.

5. Conclusion

Overall, the results presented in this work suggest that it is possible to acquire a wide set of online invariances by pretraining on an appropriate dataset, with as few as 50 objects. Invariance was acquired for scale, rotation, translation, brightness, contrast, and, to lower amount, viewpoint. Critically, online invariance was obtained both with supervised and self-supervised training. Even though these invariances observed in the latent space will not always manifest themselves in classification performance at present due to catastrophic interference, recent progress in addressing catastrophic interference in other contexts may eventually overcome these difficulties. From a psychological viewpoint, our findings suggest that online invariances may by the product of learning rather than an innate endowment. From a machine learning perspectives, our findings highlight how recent architectural modifications of CNNs to support these invariances may be unnecessary. Rather, the focus should be on training models on the relevant datasets.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. 5AFT for ETH-80

We present here the result of the 5-alternative forced choice task (5AFC) run on the ETH-80 dataset. The details of the test were identical for the ShapeNet test. The results, Fig. A.7, top show an overall drop in accuracy averaged across networks of $6.5 \pm 7.8\%$. The variability in generalization across datasets varied between the different transformations and networks, but we could not find any pattern. Notice that even with the sharpest drop (with viewpoint transformation), the accuracy remained far above chance (which with 5AFT is 20%). We directly show the difference in 5AFC accuracy between the ETH-80 and the ShapeNet results in Fig. A.7, bottom. The most striking results may be the drop in accuracy in the contrast transformation for the Same/Different Network. At the same time, the Same/Different network seems to generalize consistently well (across the 3 seeds) on new viewpoints.

Appendix B. Invariant representations analysis

We provide here the full results of the invariant representation analysis with the Adjusted Invariance Metric $\bar{I}_t(\theta)$ for the transla-
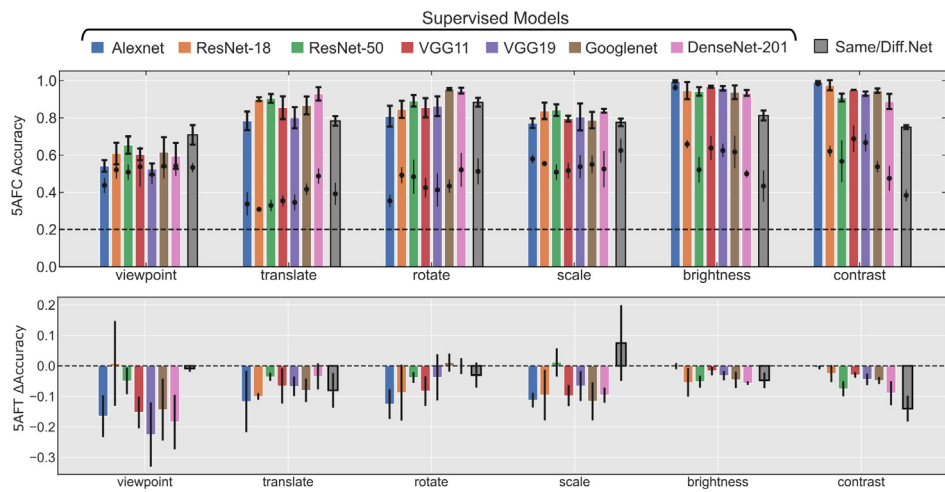
Fig. A.7. Top: Results for the 5AFT task applied on the ETH-80, a novel dataset of photographed objects. The results are slightly worse than the tests on novel ShapeNet classes (Fig. 4), which is expected given the different appearance of the samples, but the models are still showing a high degree of invariance. **Bottom**: We plot the differential accuracy between the 5AFT on the ETH-80 dataset and on the ShapeNet dataset. The difference is computed across pairs of seeds.
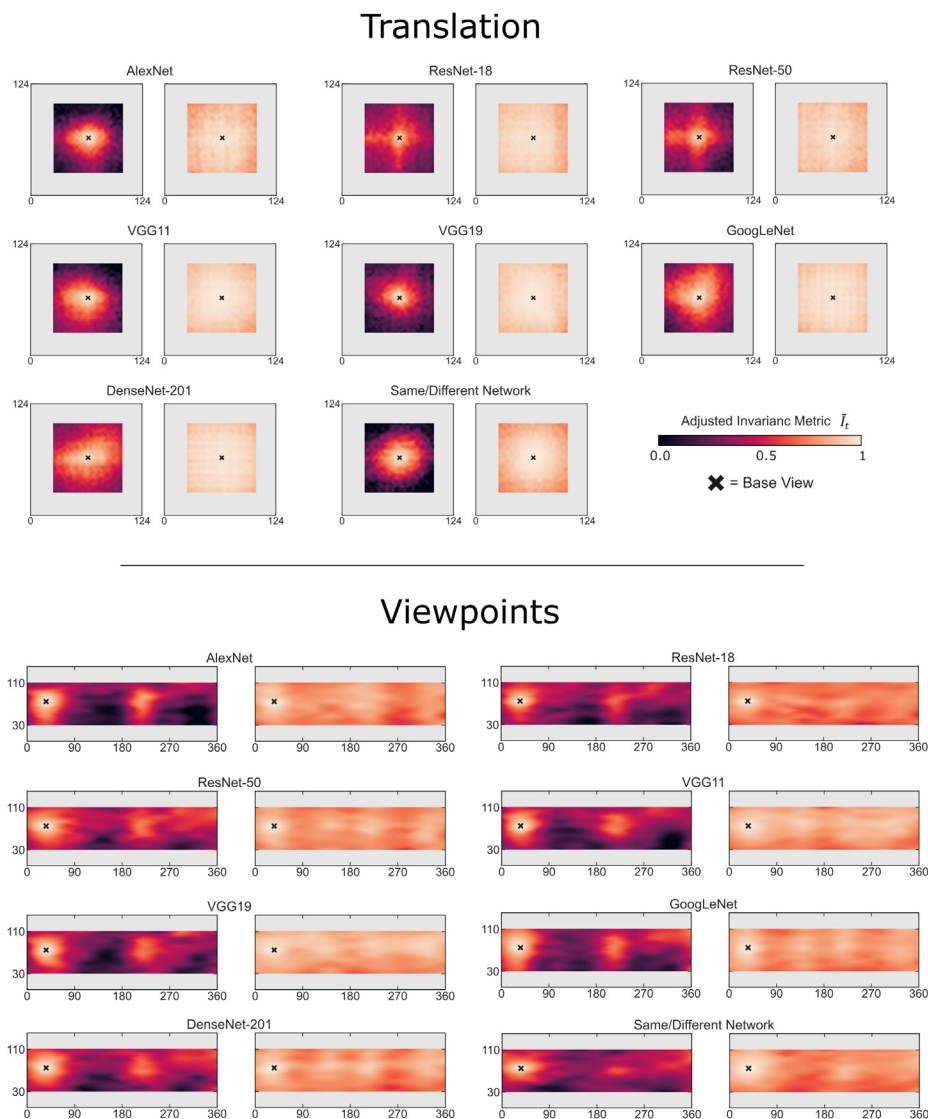
Translation

Viewpoints

Fig. B.8. The value of the Adjusted Invariance Metric for the Translation and Viewpoint transformation conditions, for all networks. Transformation of objects across novel classes are compared to a base view (the × in the plot), and the invariance representation across those poses is computed. For each network, we show **on the left** the result for the network trained on un-transformed objects, and **on the right** results for the network trained on transformed objects (either translation or viewpoints).
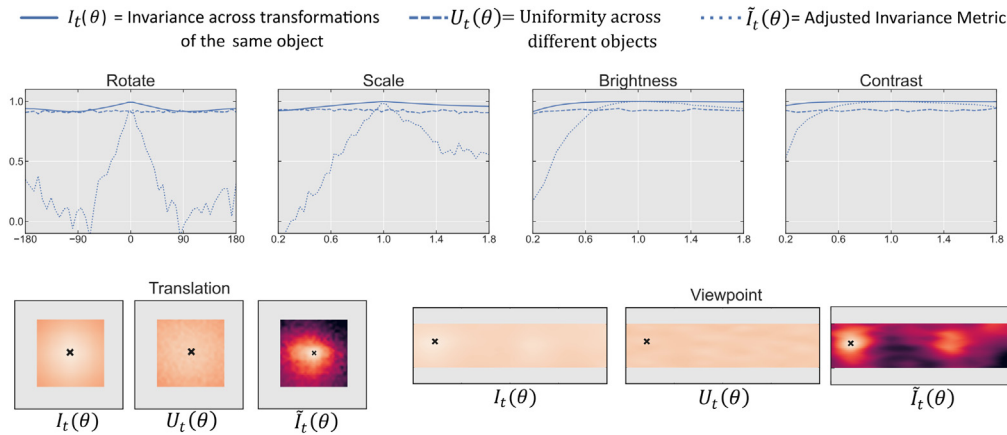
Fig. B.9. Invariance $I_t(\theta)$, Uniformity $U_t(\theta)$ and Adjusted Invariance Metric $\tilde{I}_t(\theta)$ for a vanilla (untrained) VGG11. This is a clear case where the network shows invariance to transformation by trivially collapsing *all* representations together, regardless of them belonging to the same or different objects. The metric $\tilde{I}_t(\theta)$ correctly adjusts the values to show that the network does not possess human-like invariance.

tion and viewpoint transformations (which were presented only for two networks in the main text in Fig. 5) in Fig. B.8. The results are averaged across 3 seeds. There is very little difference across networks, which similar patterns across un-transformed and transformed networks. The translation invariance results are consistent with Biscione and Bowers (2021), which tested for this transformation across different datasets.

B.1. Analysis of $I_t(\theta)$ and $U_t(\theta)$

In the main text, we defined $\tilde{I}_t(\theta)$ as the Adjusted Invariance Metric, a metric that uses the cosine similarity at the last layer to measure the invariance to object transformations and also accounts for the similarity across objects. To do that, we computed $I_t(\theta)$, the representation similarity across objects' transformations, and $U_t(\theta)$, the similarity across different objects. These two metrics are interesting on their own, as they reflect two different properties that the networks can acquire: $I_t(\theta)$ is *Invariance* of representations under transformations, $U_t(\theta)$ can be seen as *Uniformity* of representations across objects. Intuitively, a good model of the human visual system should possess a high Invariance within transformations and a low Uniformity between (different) objects. A striking example where both metrics need to be considered is a vanilla (untrained) network, which possess a high Invariance but also a high Uniformity (Fig. B.9 for VGG11). A vanilla network tends to collapse all representations to the same output (thus returning the same class for any input) regardless of objects' transformation, identity, or category. It is, in some way, perfectly invariant to transformation, but it achieves that by trivially being invariant to everything, and it is thus a poor model of human object representation. Fig. B.9 also shows how the Adjusted Invariance Metric $\tilde{I}_t(\theta)$ accounts for this.

We show these two metrics separately for the networks used in the main text in Figs. B.10, B.11, and B.12. The network trained on transformed objects appears to have a high Invariance combined with a low Uniformity (Figs. B.10, B.11, and B.12, bottom): transformed versions of the same objects are represented similarly, and different objects are represented separately. The pattern for networks trained on un-transformed objects was more peculiar (Figs. B.10, B.11, and B.12, top). In most cases, the Invariance and Uniformity were both low, meaning that the network incorrectly represented different transformation of the same objects as different objects, but also that the network correctly separated different objects. However, in few cases, the Invariance was high (e.g. the translation and Viewpoint Invariance for the Same/Different Network or for GoogLeNet in Fig. B.12, top), but

the Uniformity was also fairly high, indicating that the network collapsed together representations of *different,* transformed objects. For these networks, *different* objects seen from different viewpoints have a *similar* internal representation. Therefore, by analyzing $I_t(\theta)$ and $U_t(\theta)$ separately, we can investigate different ways in which a network could fail to possess human-like invariance to transformations (that is invariance to transformation that is not the result of collapsing all representations together).

**Appendix C. Discrepancies with previous experiments**

In a somewhat similar experiment, Hernandez-Garcia and Konig (2019) found a low similarity across transformed samples, which is in direct contradiction with our findings. This can be accounted by the usage of different metrics: Hernandez-Garcia and Konig (2019) used Euclidean distance, whereas we used cosine similarity.

When using the Euclidean distance, change in the magnitude of the feature maps' activations (e.g. when using transformations such as brightness or contrast) will result in a low similarity score. The cosine similarity is, instead, normalized by magnitude, and is therefore not affected by an overall scaling of the activation values. We show that, similarly to Hernandez-Garcia and Konig (2019), we also obtained lower invariance when using the Euclidean distance in computing the Adjusted Invariance Metric $\tilde{I}_t(\theta)$ in place of cosine similarity (Fig. C.13, compare this figure with Fig. 5). Note that there is still a fairly strong separation between similarity in networks trained on transformed objects and networks trained on un-transformed objects, which was not measured in Hernandez-Garcia and Konig (2019). The results obtained in the present work show that cosine similarity is a better suited metric for capturing invariance to transformations.

Xu and Vaziri-Pashkam (2021) similarly used the Euclidean distance to compute the dissimilarity across categories (building a Representation Dissimilarity Matrix, RDM) at two levels of a translation transformations (with objects either at the top or at the bottom of a canvas). They then computed the Representation Similarity Analysis (RSA) across the RDMs for a wide variety of CNNs and found low correlation across transformation. Their analysis is, therefore, very different to our approach, as they were focused on *consistency* of representations across transformations, not *invariance* of representations. Their finding of low consistency is nevertheless puzzling, and could be explained by several differences between ours experiments: firstly, like in Hernandez-Garcia and Konig (2019), they relied on the Euclidean distance which has the problem discussed above. Moreover, they used networks

Networks Trained on Un-transformed Objects
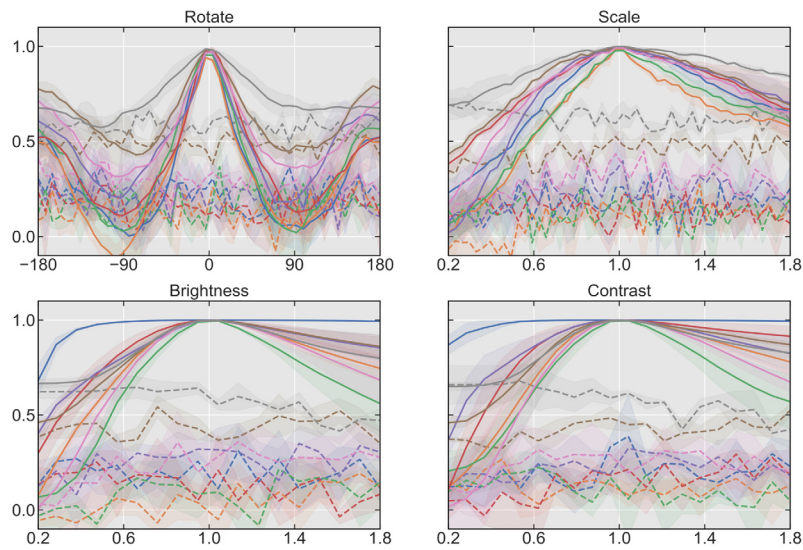
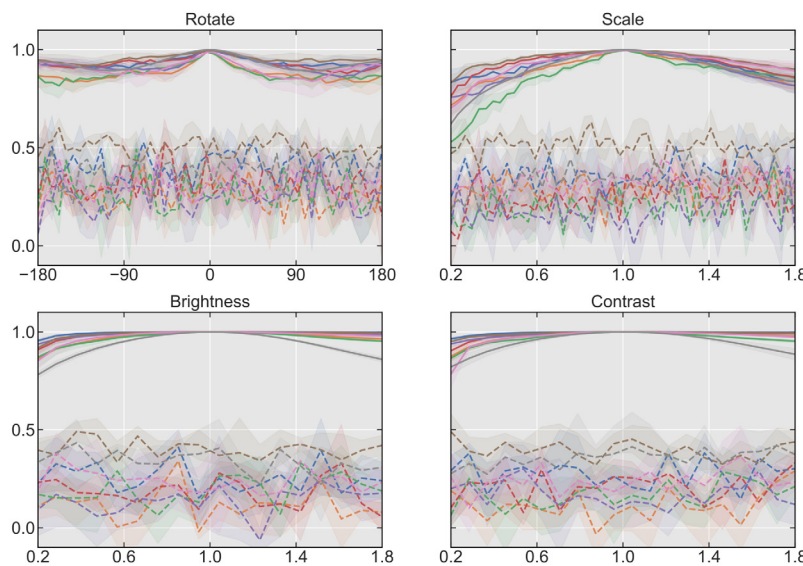## Networks Trained on Transformed Objects

Fig. B.10. Invariance ($I_t(\theta)$, continuous lines) and Uniformity ($U_t(\theta)$, dashed lines) for all networks, for rotation, scale, brightness and contrast transformations.

pretrained on ImageNet. Biscione and Bowers (2020) showed that ImageNet, when trained with specific augmentation (that is, "Random Crop"), could provide a certain degree of translation invariance, but it is not clear whether this augmentation was actually used and whether it would be enough to generalize to

such different samples. We hypothesize that pretraining the networks on a fully translate dataset, and using the cosine similarity instead of Euclidean distance, could possibly result in a high RSA, showing not only invariance, but consistency of representation across transformations.
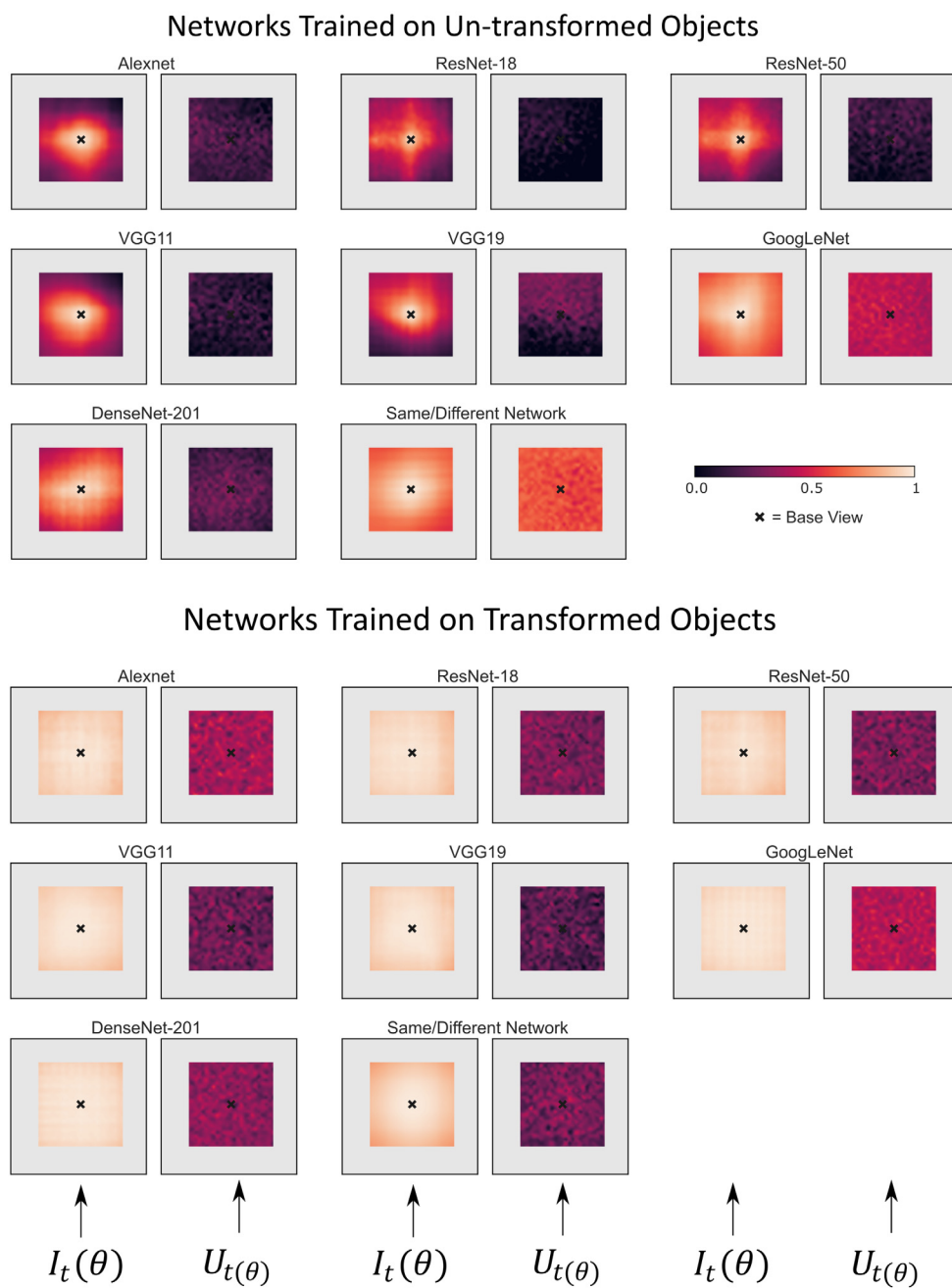
Fig. B.11. Invariance ($I_t(\theta)$, on the left of each pair of heatmaps) and Uniformity ($U_t(\theta)$, on the right of each pair of heatmaps) for translation.
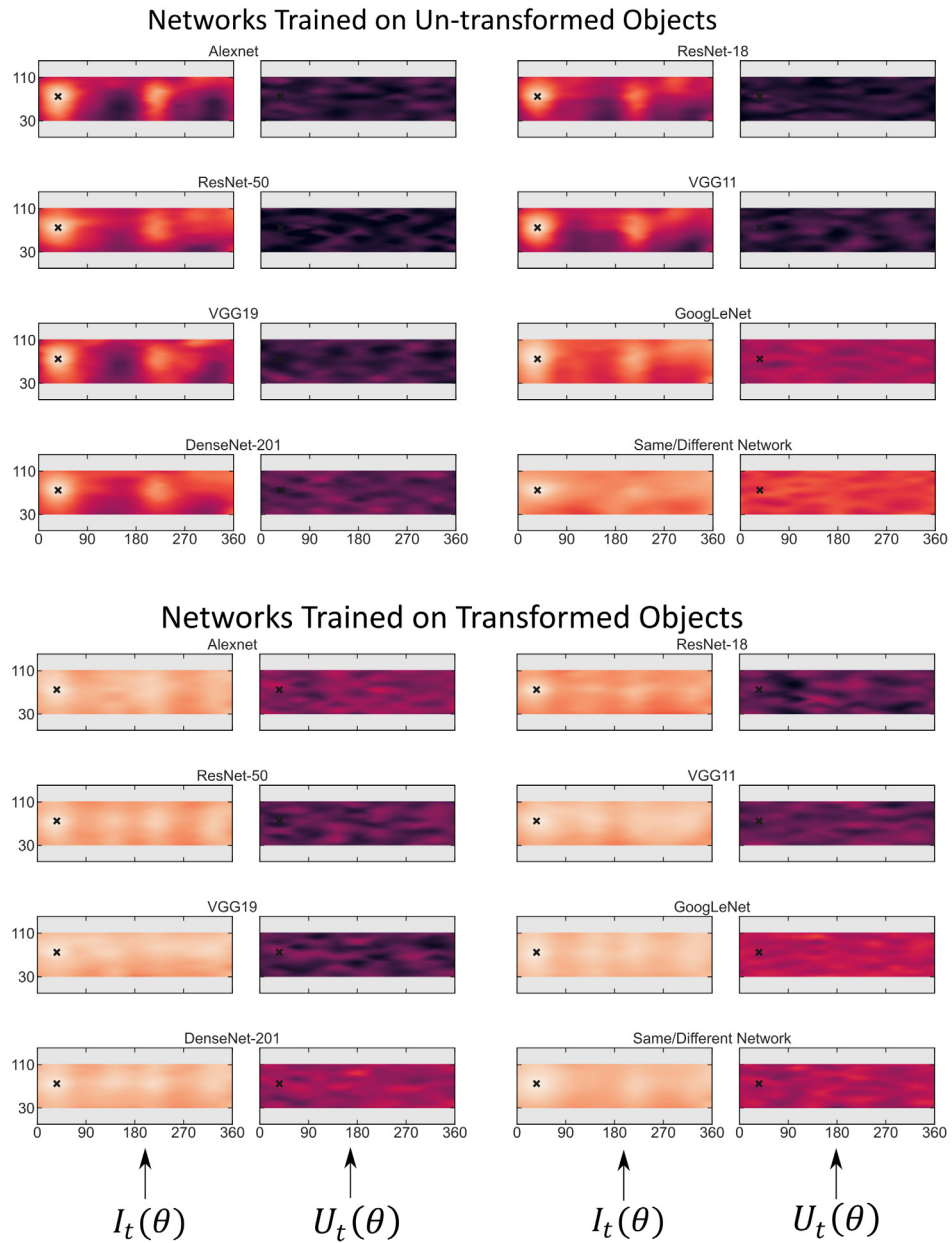
Fig. B.12. Invariance ($I_t(\theta)$, **on the left** of each pair of heatmaps) and Uniformity ($U_t(\theta)$, **on the right** of each pair of heatmaps) for change in viewpoint. Overall, networks trained on transformed objects (**bottom**) showed high Invariance and low Uniformity. Network trained on un-transformed objects (**top**) showed for most cases low Invariance and low Uniformity. In some cases, however, they showed fairly high Invariance, but also high Uniformity, which would produce unnatural internal representation in which different objects would be represented through the same activation pattern.
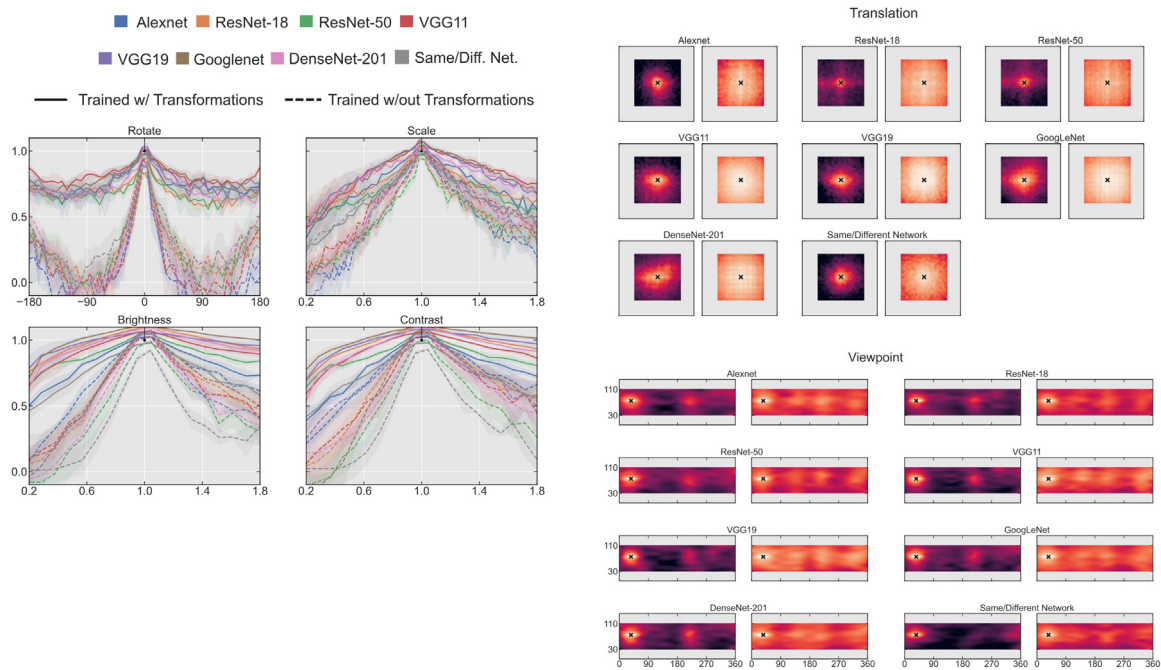
Fig. C.13. Adjusted Invariance Metric $\tilde{I}_t(\theta)$ with Euclidean Distance in place of cosine similarity for all transformations and all networks. Euclidean distance is sensitive to absolute values of layers' activations, which results in underestimating the amount of invariance across transformations. For the panels on the right side, the pair of heatmaps for each network represents the invariance for network trained on un-transformed objects (**on the left**) and network trained on transformed objects (**on the right**).

References

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology*, *14*(12), 1–43. http://dx.doi.org/10.1371/journal.pcbi.1006613.

Bar, M., & Biederman, I. (1999). Localizing the cortical region mediating visual awareness of object identity. *Proceedings of the National Academy of Sciences*, *96*(4), 1790–1793. http://dx.doi.org/10.1073/PNAS.96.4.1790, URL: https://www.pnas.org/content/96/4/1790 https://www.pnas.org/content/96/4/1790.abstract.

Beaulieu, S., Frati, L., Miconi, T., Lehman, J., Stanley, K. O., Clune, J., & Cheney, N. (2020). Learning to continually learn. *Frontiers in Artificial Intelligence and Applications*, *325*, 992–1001. http://dx.doi.org/10.3233/FAIA200193, URL: http://arxiv.org/abs/2002.09571.

Biscione, V., & Bowers, J. (2020). Learning translation invariance in CNNs. In *NeurIPS 2020, 2nd workshop on shared visual representations in human and machine intelligence (SVRHM)*. URL: http://arxiv.org/abs/2011.11757.

Biscione, V., & Bowers, J. S. (2021). Convolutional neural networks are not invariant to translation, but they can learn to be. *Journal of Machine Learning Research*, *22*(229), 1–28, URL: http://jmlr.org/papers/v22/21-0019.html.

Biscione, V., & Bowers, J. S. (2022). Do dnns trained on natural images acquire gestalt properties?. *Journal of Machine Learning Research*, arXiv preprint arXiv:22203.07302 URL: http://arxiv.org/abs/2203.07302.

Blything, R., Biscione, V., & Bowers, J. (2020). A case for robust translation tolerance in humans and cnns. a commentary on han et al. arXiv prepring arXiv:2012.05950 URL: http://arxiv.org/abs/2012.05950.

Blything, R., Biscione, V., Vankov, I. I., Ludwig, C. J. H., & Bowers, J. S. (2021). The human visual system and CNNs can both support robust online translation tolerance following extreme displacements. *Journal of Vision*, *21*(2), 1–16. http://dx.doi.org/10.1167/jov.21.2.9.

Bornstein, M. H., Krinsky, S. J., & Benasich, A. A. (1986). Fine orientation discrimination and shape constancy in young infants. *Journal of Experimental Child Psychology*, *41*(1), 49–60.

Bowers, J. S., Vankov, I. I., & Ludwig, C. J. (2016). The visual system supports online translation invariance for object identification. *Psychonomic Bulletin and Review*, *23*(2), 432–438. URL: https://link.springer.com/article/10.3758/s13423-015-0916-2.

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., & Yu, F. (2015). ShapeNet: AN information-rich 3D model repository. arXiv prepring arXiv:1512.03012 URL: http://arxiv.org/abs/1512.03012.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. (pp. 1575–1585). arXiv preprint arXiv:2002.05709 PartF16814. URL: http://arxiv.org/abs/2002.05709.

Chen, F. X., Roig, G., Isik, L., Boix, X., & Poggio, T. (2017). *Eccentricity dependent deep neural networks: Modeling invariance in human vision*: AAAI spring symposium - technical report SS-17-01 -, (pp. 541–546).

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), 1–13. http://dx.doi.org/10.1038/srep27755, URL: https://www.nature.com/articles/srep27755.

Cohen, T. S., & Welling, M. (2016). Group equivariant convolutional networks. In *33rd international conference on machine learning, ICML 2016, Vol. 6* (pp. 4375–4386). International Machine Learning Society (IMLS), URL: http://arxiv.org/abs/1602.07576.

Day, R. H., & McKenzie, B. E. (1981). Infant perception of the invariant size of approaching and receding objects. *Developmental Psychology*, *17*(5), 670–677. http://dx.doi.org/10.1037/0012-1649.17.5.670, URL: /record/1981-32490-001.

Dujmović, M., Malhotra, G., & Bowers, J. S. (2020). What do adversarial images tell us about human vision? *ELife*, *9*, 1–29. http://dx.doi.org/10.7554/ELIFE.55978.

Fernando, C., Banarse, D., Blundell, C., Zwols, Y., Ha, D., Rusu, A. A., Pritzel, A., Wierstra Google DeepMind, D., & Brain, G. (2017). PathNet: EVolution channels gradient descent in super neural networks. URL: https://arxiv.org/abs/1701.08734v1.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, *3*(4), 128–135. http://dx.doi.org/10.1016/S1364-6613(99)01294-2.

Geirhos, R., Medina Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., Wichmann, F. A., 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. Garnett (Eds.), *Advances in neural information processing systems, Vol. 31* (pp. 7538–7550). Curran Associates, Inc..

Gentner, D., Shao, R., Simms, N., & Hespos, S. (2021). Learning same and different relations: cross-species comparisons. *Current Opinion in Behavioral Sciences*, *37*(April), 84–89. http://dx.doi.org/10.1016/j.cobeha.2020.11.013.

German, J. S., & Jacobs, R. A. (2020). Can machine learning account for human visual object shape similarity judgments? *Vision Research*, *167*(December 2019), 87–99. http://dx.doi.org/10.1016/j.visres.2019.12.001.

Gong, Y., Wang, L., Guo, R., & Lazebnik, S. (2014). Multi-scale orderless pooling of deep convolutional activation features. In *LNCS: Vol. 8695*, *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (pp. 392–407). Springer Verlag, URL: http://arxiv.org/abs/1403.1840.

Han, Y., Roig, G., Geiger, G., & Poggio, T. (2020). Scale and translation-invariance for novel objects in human vision. *Scientific Reports*, *10*(1), 1–13. http://dx.doi.org/10.1038/s41598-019-57261-6.

Hernandez-Garcia, A., & Konig, P. (2019). Learning representational invariance instead of categorization. In *Proceedings - 2019 international conference on computer vision workshop, ICCVW 2019* (pp. 4587–4590). http://dx.doi.org/10.1109/ICCVW.2019.00561.

Hespos, S., Gentner, D., Anderson, E., & Shivaram, A. (2021). The origins of same/different discrimination in human infants. *Current Opinion in Behavioral Sciences*, *37*, 69–74. http://dx.doi.org/10.1016/j.cobeha.2020.10.013.

Jing, L., & Tian, Y. (2019). Self-supervised visual feature learning with deep neural networks: A survey. (pp. 1–22). http://dx.doi.org/10.1109/tpami.2020.2992393, arXiv prepring arxiv:1902.06162 8828.

Kauderer-Abrams, E. (2017). Quantifying translation-invariance in convolutional neural networks. arXiv preprint arXiv:1801.01450v1 URL: http://arxiv.org/abs/1801.01450.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*(11), Article e1003915. http://dx.doi.org/10.1371/journal.pcbi.1003915.

Kim, J., Jung, W., Kim, H., & Lee, J. (2020). Cycnn: A rotation invariant CNN using polar mapping and cylindrical convolution layers. URL: https://arxiv.org/abs/2007.10588v1.

Kraebel, K. S., & Gerhardstein, P. C. (2006). Three-month-old infants' object recognition across changes in viewpoint using an operant learning procedure. *Infant Behavior and Development*, *29*(1), 11–23. http://dx.doi.org/10.1016/j.infbeh.2005.10.002.

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time-series. *The Handbook of Brain Theory and Neural Networks*, *3361*.

Leibe, B., & Schiele, B. (2003). Analyzing appearance and contour based methods for object categorization. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. Vol. 2* (pp. II–409). IEEE.

Liu, X., Masana, M., Herranz, L., Van De Weijer, J., Lopez, A. M., & Bagdanov, A. D. (2018). Rotate your networks: Better weight consolidation and less catastrophic forgetting. In *Proceedings - International conference on pattern recognition, Vol. 2018-Augus* (pp. 2262–2268). Institute of Electrical and Electronics Engineers Inc., http://dx.doi.org/10.1109/ICPR.2018.8545895, URL: https://arxiv.org/abs/1802.02950v4.

Malhotra, G., Dujmović, M., & Bowers, J. S. (2021). Feature blindness: a challenge for understanding and modelling visual object recognition. *BioRxiv*, Article 2021.10.20.465074. http://dx.doi.org/10.1101/2021.10.20.465074, URL: https://www.biorxiv.org/content/10.1101/2021.10.20.465074v2 https://www.biorxiv.org/content/10.1101/2021.10.20.465074v2.abstract.

Marcos, D., Volpi, M., & Tuia, D. (2016). Learning rotation invariant convolutional filters for texture classification. In *Proceedings - International conference on pattern recognition* (pp. 2012–2017). Institute of Electrical and Electronics Engineers Inc., http://dx.doi.org/10.1109/ICPR.2016.7899932, URL: http://arxiv.org/abs/1604.06720 http://www.outex.oulu.fi.

Marcus, G. (2018). Deep learning: A critical appraisal. arXiv preprint arXiv:1801.00631 URL: http://arxiv.org/abs/1801.00631.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation - Advances in Research and Theory*, *24*(C), 109–165. http://dx.doi.org/10.1016/S0079-7421(08)60536-8.

Patacchiola, M., & Storkey, A. (2020). Self-supervised relational reasoning for representation learning. In *Advances in Neural Information Processing Systems*.

Premack, D. (1983). The codes of man and beasts. *Behavioral and Brain Sciences*, *6*(1), 125–136. http://dx.doi.org/10.1017/S0140525X00015077.

Puebla, G., & Bowers, J. S. (2021). Can deep convolutional neural networks support relational reasoning in the same-different task?. *bioRxiv*, 2021.09.03.458919. http://dx.doi.org/10.1101/2021.09.03.458919, https://www.biorxiv.org/content/10.1101/2021.09.03.458919v1, https://www.biorxiv.org/content/10.1101/2021.09.03.458919v1.abstract.

Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2016). Pointnet: deep learning on point sets for 3d classification and segmentation. *arXiv preprint arXiv:1612.00593*.

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., & Hadsell, R. (2016). Progressive neural networks. URL: https://arxiv.org/abs/1606.04671v3.

Schaul, T., Quan, J., Antonoglou, I., & Silver, D. (2015). Prioritized experience replay. In *4th international conference on learning representations, ICLR 2016 - conference track proceedings*. URL: https://arxiv.org/abs/1511.05952v4.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, Article 407007. http://dx.doi.org/10.1101/407007, URL: https://www.biorxiv.org/content/10.1101/407007v1.

Schwartz, M., Day, R. H., & Cohen, L. B. (1979). Visual shape perception in early infancy. http://dx.doi.org/10.2307/1165963, URL: https://www.jstor.org/stable/1165963.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., & Hospedales, T. M. (2018). Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 1199–1208). http://dx.doi.org/10.1109/CVPR.2018.00131.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. In *2nd international conference on learning representations, ICLR 2014 - conference track proceedings*. URL: https://arxiv.org/abs/1312.6199v4.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, *19*(1), 109–139. http://dx.doi.org/10.1146/annurev.ne.19.030196.000545.

Xu, Y., & Vaziri-Pashkam, M. (2021). Examining the coding strength of object identity and nonidentity features in human occipito-temporal cortex and convolutional neural networks. *The Journal of Neuroscience*, *41*(19), 4234–4252. http://dx.doi.org/10.1523/jneurosci.1993-20.2021, URL: https://pubmed.ncbi.nlm.nih.gov/33789916/.

Xu, Y., Xiao, T., Zhang, J., Yang, K., & Zhang, Z. (2014). Scale-invariant convolutional neural networks. arXiv preprint arXiv:1411.6369 URL: http://arxiv.org/abs/1411.6369.

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. http://dx.doi.org/10.1038/nn.4244, URL: https://www.nature.com/articles/nn.4244.

Yamins, D. L. K., Hong, H., Cadieu, C., & DiCarlo, J. (2013). Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream. *DiCarlo Via Courtney Crummett*, URL: https://dspace.mit.edu/handle/1721.1/95910.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. http://dx.doi.org/10.1073/PNAS.1403112111, URL https://www.pnas.org/content/111/23/8619 https://www.pnas.org/content/111/23/8619.abstract.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. K. (2020). Unsupervised neural network models of the ventral visual stream. *BioRxiv*, Article 2020.06.16.155556. http://dx.doi.org/10.1101/2020.06.16.155556, URL: https://www.biorxiv.org/content/10.1101/2020.06.16.155556v1.