The human visual system and CNNs can both support robust online translation tolerance following extreme displacements

Ryan Blything School of Psychological Science, University of Bristol, Bristol, UK

Valerio Biscione School of Psychological Science, University of Bristol, Bristol, UK

Ivan I. Vankov Department of Cognitive Science and Psychology, Sofia, New Bulgarian University, Bulgaria

Casimir J. H. Ludwig School of Psychological Science, University of Bristol, Bristol, UK

Jeffrey S. Bowers School of Psychological Science, University of Bristol, Bristol, UK

Visual translation tolerance refers to our capacity to recognize objects over a wide range of different retinal locations. Although translation is perhaps the simplest spatial transform that the visual system needs to cope with, the extent to which the human visual system can identify objects at previously unseen locations is unclear, with some studies reporting near complete invariance over 10 degrees and other reporting zero invariance at 4 degrees of visual angle. Similarly, there is confusion regarding the extent of translation tolerance in computational models of vision, as well as the degree of match between human and model performance. Here, we report a series of eye-tracking studies (total $N = 70$) demonstrating that novel objects trained at one retinal location can be recognized at high accuracy rates following translations up to 18 degrees. We also show that standard deep convolutional neural networks (DCNNs) support our findings when pretrained to classify another set of stimuli across a range of locations, or when a global average pooling (GAP) layer is added to produce larger receptive fields. Our findings provide a strong constraint for theories of human vision and help explain inconsistent findings previously reported with convolutional neural networks (CNNs).

Introduction

We can identify familiar objects despite the variable images they project on our retina, including variation in image size, orientation, illumination, and position. How the visual system succeeds under these conditions is still poorly understood. Here, we consider the case of variation across position and the extent to which the visual system and artificial neural networks can identify objects at previously unseen locations.

Although translation is perhaps the simplest spatial transform that the visual system needs to cope with, there is nevertheless confusion in the literature regarding the extent of translation tolerance in human vision, the extent of tolerance in computational models of vision, as well as the degree of match between humans and models. There are both theoretical and methodological reasons for this confusion. With regard to theory, researchers often fail to distinguish between on-line tolerance and trained tolerance (see Bowers, Vankov, & Ludwig, 2016). In the case of on-line tolerance, learning to identify an object at one location immediately affords the capacity to identify that object at multiple other retinal locations even when no members of that category have ever been seen at those other locations. For example, if a person has only seen dogs projected at one retinal location, they may nevertheless be able to identify dogs when projected to other retinal locations. At one extreme, the visual system can immediately generalize to all locations (within the limits of visual acuity), what might be called on-line translation invariance; at the other extreme, there is no generalization to untrained locations. Trained tolerance, by contrast, refers to the hypothesis that we

Citation: 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):9, 1–16, https://doi.org/10.1167/jov.21.2.9.

https://doi.org/10.1167/jov.21.2.9 Received March 24, 2020; published February 23, 2021 ISSN 1534-7362

can identify a novel exemplar of an object category across a range of retinal locations by training the visual system to identify other exemplars of this category across a broad range of retinal locations. For instance, learning to identify multiple images of dogs at multiple retinal locations allows the visual system to identify a new image of a dog across multiple locations. This distinction is often ignored in discussion of translation tolerance in humans and computational models of vision, leading to a wide range of different conclusions regarding the extent of translation tolerance. Here, we are concerned with on-line translation tolerance.

Furthermore, when behavioral studies were specifically designed to assess on-line translation tolerance, a variety of methodological differences has led to a wide range of findings, ranging from no on-line tolerance at 4 degrees (Cox & DiCarlo, 2008) to near complete on-line tolerance at 13 degrees (Bowers et al., 2016). Similarly, different computational models of object classification support varying degrees of online translation tolerance, from near zero tolerance (Chen et al., 2017) to complete invariance (Han et al., 2020). These mixed outcomes have led to contrasting conclusions, with many researchers emphasizing the importance of trained rather than on-line tolerance in both biological and computational models of vision (e.g. Cox & DiCarlo, 2008; Dandurand et al., 2013; Di Bono & Zorzi, 2013; Edelman & Intrator, 2003; Elliffe et al., 2002; Serre, 2019), and others proposing theories that support on-line translation invariance (Biederman, 1987; Hummel & Biederman, 1992). Researchers have also modified the architectures of standard convolutional neural networks (CNNs) in order to explain the lack of translation tolerance beyond 4 degrees (Chen et al. 2017), or alternatively, modified CNNs in order to explain the near complete translation tolerance in some conditions and limited translation tolerance in others (Han et al., 2020). In the current work, we show that on-line translation tolerance for images of novel 3D objects is greater in human vision than previously demonstrated even when the images are flashed for 100 ms and masked at an untrained retinal position at test. In addition, our simulations highlight several conditions in which CNNs demonstrate extreme translation tolerance and display human-level like performance.

Brief review of on-line translation tolerance in biological and artificial visual systems

In the case of biological vision, early behavioral studies provided evidence for robust translation tolerance following 10 degrees of translation based on long-term priming studies (Biederman & Cooper, 1991; Cooper, Biederman, & Hummel, 1992; Ellis, Allport, Humphreys, & Collis, 1989; Fiser & Biederman, 2001; Stankiewicz & Hummel, 2002; Stankiewicz, Hummel, & Cooper, 1998). Similarly, in single-cell neurophysiological studies, researchers have identified neurons in inferior-temporal cortex (IT) with extremely large receptive fields (up to 26 degrees; Op De Beeck & Vogels, 2000) that are thought to provide the neural underpinning of translation tolerance. However, these findings were obtained with familiar stimuli, and this has led researchers to argue that robust priming and large receptive fields might reflect trained rather than on-line translation tolerance (e.g. Kravitz, Vinson, & Baker, 2008). That is, although the specific test stimuli in these experiments may have only been experienced at one location, exemplars of these object categories would have been experienced at a wide variety of retinal locations through every day experience, and this may have been necessary in order to support translation tolerance for the stimuli used in these experiments.

As illustrated in Figure 1, this led to a number of studies that assessed on-line translation tolerance for a range of unfamiliar stimuli, with highly mixed results. In many cases, these behavioral and physiological studies revealed that on-line translation tolerance was much reduced following just a few degrees of translation (see Figures 1A–D). At the same time, near complete on-line translation tolerance has been observed following shifts of 8 degrees (Dill & Edelman, 2001; see Figure 1E). Han et al. (2020) also observed near complete on-line translation tolerance when stimuli were first presented at fixation and then shifted 7 degrees (see Figure 1F), but tolerance was much more limited when the stimuli were first presented in peripheral vision and then shifted 7 degrees to fixation (see Figure 1G). Bowers et al. (2016) demonstrated on-line translation tolerance at the most distal locations to date, with high performance at displacements up to 13 degrees (see Figure 1H). One notable feature of many previous studies is that they used novel stimuli that are very unlike real objects and often the stimuli differed from one another in only fine perceptual detail, which may force the visual system to rely on low-level visual representations that are retinotopically constrained (e.g. see Figure 1A–C). This might be relevant to explaining the range of findings given that greater on-line translation tolerance has been observed for novel stimuli that were structurally more similar to real objects (see Figure 1E) or which were designed to differ from one another in their configurational properties rather than fine details (see Figures 1F–H).

With regard to computational modeling, most researchers have only considered trained translation tolerance. For example, a number of models of visual word identification support robust tolerance after training each word at each location (Dandurand et al., 2013; Di Bono & Zorzi, 2013). This is also the case with CNNs widely used in computer science that are

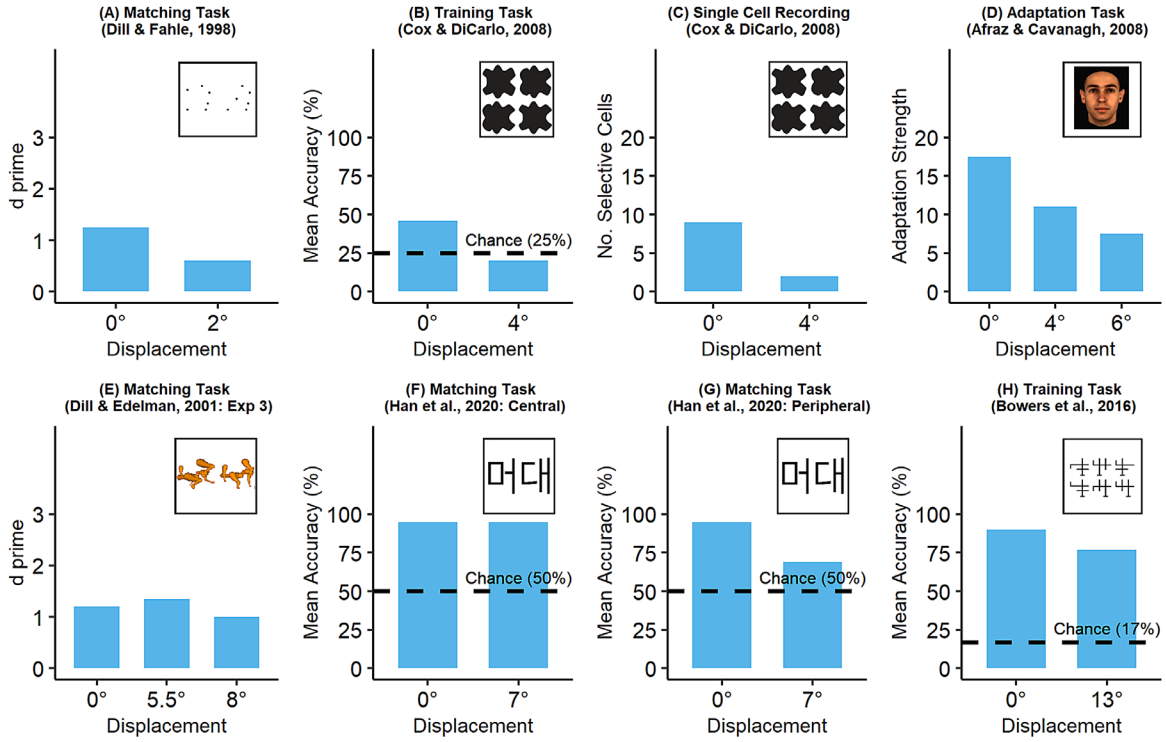

Figure 1. Behavioral investigations of translation tolerance, adapted from Kravitz et al. (2008). (**A**) In a "same-different" matching task, it was significantly easier to discriminate a test-image as "same" or "different" to a probe image when these images were presented at matched locations compared with when the test image was displaced by just 2 degrees (Dill & Fahle, 1998). (**B, C**) In a training task, Cox and DiCarlo (2008) demonstrated that an adult rhesus monkey performed at chance-levels (25%) when required to identify 4 novel objects that were displaced just 4 degrees from the trained location, concluding that the "behavioral failure to position-generalize is caused by the monkey's reliance on a non-position tolerant visual neuronal representation" (page 10053). In a single-cell recording study, Cox and DiCarlo (2008) also detected significantly more selective cells when novel objects were presented at the trained position compared to the displaced position. (**D**) In an adaptation task (Afraz & Cavanagh, 2008), adaptation effects (i.e. when exposure to a face alters the perception of a subsequently presented face) were inhibited after the image was displaced by 4 degrees, and more-so when displaced by 6 degrees. (**E**) Using stimuli composed of scrambled animal parts, Dill and Edelman (2001) showed on-line translation tolerance over displacements of 8 degrees. (**G, H**) Han et al. (2020) showed on-line translation tolerance over displacements of 7 degrees although performance reduced when stimuli were trained in peripheral vision (panel **H**) as opposed to at fixation (panel **G**). (**I**) In Bowers et al. (2016), participants performed well above chance at the largest displacements to date, up to 13 degrees.

often described as the best current theory of object recognition in humans (e.g. Kubilius, Kar, Schmidt, & DiCarlo, 2018). Although the design of CNNs (both the convolution and the pooling layers) are claimed to support translation invariance (for introduction see O'Shea & Nash, 2015), these models are generally trained to categorize images by training multiple exemplars of each image category at multiple spatial locations. Indeed, the standard training procedure for CNNs is to present each image across a range of positions, scales, and poses, a procedure called "data augmentation."

We are only aware of a few cases in which modelers have assessed on-line translation tolerance, and, in most cases, only a limited degree of tolerance was observed. A biologically inspired neural network model called VisNet showed 100% accuracy at untrained locations for simple stimuli (Elliffe et al., 2002), but only when each stimulus was trained at multiple other spatial locations (after training in 7 locations, the model generalized to an 8th and 9th location), and the authors only tested small translations (8 pixels in a 128 × 128 "retina"). This small degree of on-line tolerance was thought to provide a reasonable description of human on-line translation tolerance. The HMAX (Riesenhuber & Poggio, 1999) model showed 100% accuracy for translations up to 4 degrees from the trained location, but its performance more than halved when objects were translated by distances equivalent to 7 degrees and the model was not tested beyond that point. The authors also claimed that HMAX showed "the same scale and position invariance properties as the view-tuned IT neurons described by Logothetis et al. (1995) using the same stimuli" (Riesenhuber &

Poggio, 2002, p.163). There are also a few cases in which CNNs were trained on images at one retinal location and tested at another, and, in most cases, highly limited on-line translation tolerance was observed (Chen et al., 2017; Furukawa, 2017; Kauderer-Abrams, 2017; Qi, 2018). However, Han et al. (2020) observed near perfect translation invariance over 7 degrees for Korean letters in a standard CNN (see Modelling translation tolerance in CNNs for some more details regarding on-line tolerance in CNNs). Clearly, both the behavioral and modeling results are mixed.

Here, we report a series of behavioral studies that demonstrate more extreme on-line tolerance compared with previous research and a set of simulations that examine the capacity of CNNs to support on-line translation tolerance. We show that a standard CNN (VGG16; Simonyan & Zisserman, 2014) only supports robust on-line tolerance for novel stimuli when pretrained on another set of stimuli presented at multiple retinal locations. That is, trained-tolerance for one set of stimuli led to on-line tolerance for another set of stimuli. We also show that robust on-line translation tolerance can be achieved without any pretraining by modifying the architecture of a CNN (by adding a global average pooling (GAP) layer to the network that generates larger receptive fields). Our findings challenge the common claim that human on-line translation tolerance is highly limited, help explain the mixed set of on-line translation tolerance results reported in CNNs, and show that standard CNNs can account for human on-line translation tolerance.

Psychophysical studies: Assessing on-line translation tolerance in the human visual system

Four gaze-contingent eye-tracking studies are reported, and include the following critical design features. First, we used 24 images of naturalistic novel 3D objects organized into pairs composed of similar parts arranged in different global configurations, with one member of each pair was assigned to category A, the other to category B (see Figure 2,A; Leek, Roberts, Oliver, Cristino, & Pegna, 2016). This should encourage participants to learn the complete objects rather than just the parts when categorizing them. Note, previous studies have used a smaller number of novel 2D stimuli that may often have been classified on the basis of local object features (such as those depicted in Figures 1A–C). Second, participants learned to identify the novel objects that were projected one (or two) retinal locations before being tested at novel locations. That is, participants learned new object representations in long-term memory during training, and we assessed whether these long-term object codes supported extensive on-line tolerance. By contrast, many previous studies did not involve learning any new representations, but rather required matching of stimuli presented in quick succession in short-term memory (e.g., Dill & Fahle, 1998, Dill & Edelman, 2001; Han et al., 2020). Third, we included test conditions in which objects were presented for 100 ms durations, reducing the likelihood that participants adopted artificial strategies at test (e.g. slowly searching for a set of features diagnostic of category membership using covert attention). Finally, in order to test the limits of on-line translation tolerance, our experiments used displacements of up to 18 degrees (see Experiment 4), which is larger than any previous experiment. Note that the Bowers et al. (2016) experiments reporting robust on-line translation tolerance at 13 degrees included a smaller number of less realistic 2D objects that were displayed for an extended time at test. Accordingly, the current studies provide a much stronger test of on-line translation tolerance.

General method for psychophysical studies (Experiments 1–4)

Ethics statement

The study was approved by the University of Bristol Faculty of Science Ethics Committee and was in accordance with the provisions of the World Medical Association Declaration of Helsinki. All participants were recruited from the University of Bristol's course credit scheme for psychology students.

Participants

Ten participants completed each experiment (1a, 1b, 1c, 2, 3, 4a, and 4b), giving 70 in total (48 women; median age = 20 years). Across experiments, 17 additional participants were excluded due to failure to complete the training phase within 90 minutes. The sample size was chosen based on previous psychophysical experiments that have used identification tasks to examine translation invariance (i.e. studies reported in Figure 1).

Equipment

Eye movements were monitored using the Eyelink 1000 plus system (SR Research). Stimuli were presented using Psychopy version 1.85.3 (Peirce & MacAskill, 2018; platform: Linux-Ubuntu), and on a Viewpixx 3D Lite monitor running at 120 Hz with a spatial resolution of 1920 × 1080 pixels (screen width = 53 cm), at a distance of 70 cm.

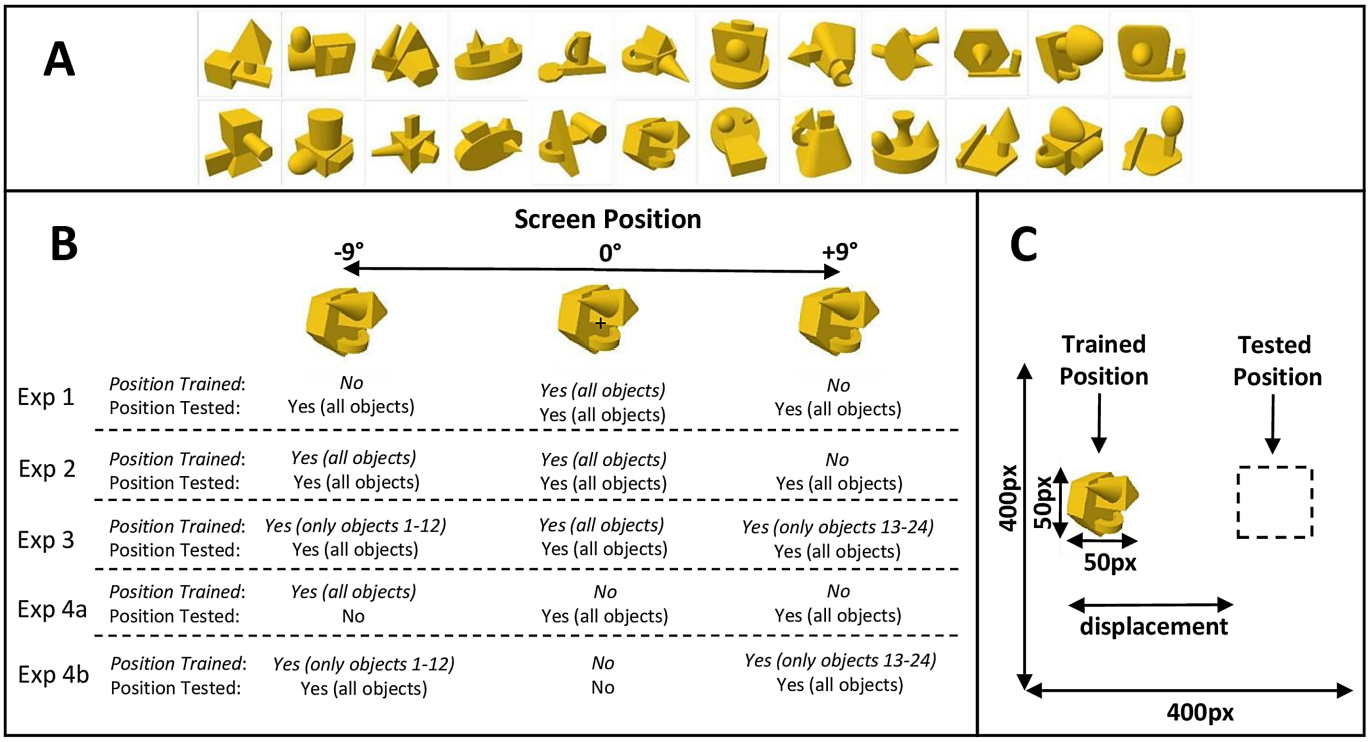

Figure 2. (**A**) Twenty-four novel objects used in behavioral studies and CNN simulations. Each column contains a pair of objects that are matched for similar local features, but which differ in global configuration. One member of each pair was randomly assigned the label "A" and the other was assigned "B." (**B**) Screen positions used at training and test in behavioral experiments (fixation was always at the center). In experiments 1a, 1b, and 1c, all objects were trained at 0 degrees, and tested at 0 degrees, +3 degrees, +6 degrees, and +9 degrees (for space reasons, +3 degrees and +6 degrees are not illustrated). In experiment 2, all objects were trained at -9 degrees and 0 degrees, then tested at -9 degrees, 0 degrees, and +9 degrees. In experiment 3, twelve objects were trained at -9 degrees and 0 degrees, and the other 12 were trained at 0 degrees and +9 degrees; all 24 objects were then tested at -9 degrees, 0 degrees, and +9 degrees. In experiments 4a and 4b, objects were never trained at the 0 degrees position, and, thus, novel test presentations at 9 degrees were displaced by 18 degrees. (**C**) Illustration of procedure used to test on-line translation tolerance with CNN simulations. A CNN (VGG16) (Simonyan & Zisserman, 2014) was trained to classify 224 × 224 pixel images containing 40 × 40 pixel novel objects; at test, the novel objects were displaced from the trained position. The precise training method used for the CNN was manipulated by crossing a number of factors, most notably: (i) Pretraining (VGG16 pretrained on ImageNet versus VGG16 trained only to classify the 24 novel objects), (ii) global pooling (global average pooling versus no global average pooling).

Procedure

In the learning phase of the experiment, participants were trained to categorize the 24 objects as "A" or "B." Each object was presented one-by-one and occupied 5 degrees × 5 degrees of visual angle. Participants were required to maintain their gaze on a centrally located fixation-cross for 1000 ms for an object to appear. If gaze moved 1.5 degrees beyond the fixation-cross, a mask replaced the object. The learning task was split into two phases: (i) familiarization. Each presentation of an object was accompanied by a sound-file indicating its category (A or B). (ii) Training. Each object was presented again but without the sound file and participants pressed a button to indicate each image's category. Audio feedback was then provided. The training phase continued until the participant correctly identified each object consecutively (in most experiments this required 24/24 consecutive correct answers). After the learning phase, participants completed the test phase with the 24 objects presented once in a random order in each test-block with no feedback. Again, if the gaze moved 1.5 degrees beyond the fixation-cross, the object was immediately replaced with a mask. The specific details of the training and test phases in each experiment are provided in the relevant subsection below and summarized in Supplementary Information Table S1). Figure 2B provides an illustration of the training and test positions used in each experiment.

Experiment 1

Procedure

During the familiarization and training phases in experiment 1a and 1b, images remained on the screen until participants responded or for just 100 ms, depending on the learning block (see Supplementary Information Table S1). All 24 objects were trained at the central location (at fixation) and 24 of 24 consecutive correct answers were required to progress to the next learning block or test phase. At test, images remained on the screen until participants responded in experiment 1a, and for 100 ms in experiment 1b in order to reduce possible response strategies. Experiment 1c was the same as experiment 1b except that stimuli were presented only for 100 ms in the familiarization and training phases.

Results

Data for all experiments can be downloaded at https://osf.io/jahm9/. Table 1 shows the accuracy with which participants categorized novel objects at each test position. Performance was excellent at untrained retinal-positions (chance is 50%). Even at the most distal untrained position (9 degrees), objects were recognized with a mean accuracy of 94% when unlimited time was afforded at test (experiment 1a), and although translation tolerance was reduced when stimuli were presented for 100 ms at test (experiments 1b and 1c), accuracy was still at least 80% when at 9 degrees displacement. We carried out Bayesian Hypothesis paired sample *t*-tests, conducted in JASP (JASP Team, 2019),[1] comparing performance at trained versus untrained locations. In all conditions (across all experiments) on-line translation tolerance was robust (all one sample *t*-tests produced Bayes Factors > 1000), but, in most cases, there was evidence for a decrease in accuracy following the most distal translations (all Bayes Factors for the 0 degrees vs. 9 degrees comparison were > 3).[2] In the interest of space, reaction times for each condition are reported in Table S2 of the supplementary section (reaction times tell a largely similar story as the accuracy data reported above, even though our experiments were not designed as reaction time experiments and participants were not explicitly instructed to respond as quickly as possible).

Experiment 2

Experiment 2 investigated two methodological factors from experiment 1 that might explain the lower performance in periphery. First, objects were always trained in foveal vision, and as a consequence, there was a confound with eccentricity and displacement: 0 degrees displacement was tested at fixation where visual acuity is high, whereas 3 degrees, 6 degrees, and 9 degrees displacements were tested in peripheral vision where acuity is lower and within-object crowding by the constituent parts (Martelli et al., 2005) may further impede recognition. Second, the most distal training locations were always presented in the final test blocks, raising possible order effects.

Procedure

In experiment 2 objects were trained 9 degrees left of fixation as well as at fixation (see Figure 2). Three test locations were used: 9 degrees left of fixation, center of fixation (in fovea), and 9 degrees right of fixation, giving three test conditions: "0 degrees P" (0 degrees displacement from peripheral training location), "0 degrees C" (0 degrees displacement from central training location), and "9 degrees" (9 degrees displacement from central training location, on the opposite side to the trained peripheral location). Comparing "0 degrees P" to "9 degrees"provides an assessment of on-line translation tolerance without the confound of eccentricity. Again, 24 of 24 consecutive correct answers were required at each training location to progress to the test phase. To control for possible order effects, the three test locations were randomly interleaved within each of six test-blocks.

Results

Table 1 summarizes the results of experiment 2. Again, robust on-line translation tolerance was observed in all locations, and there was still some reduction in performance between 0 degrees C and the untrained (9 degrees) test locations. Critically, this reduction was still present when comparing 0 degrees P with 9 degrees (Bayes Factor $= 9.68$), indicating that the reduction in performance to novel retinal locations cannot be attributed to the limitations of peripheral vision alone.

Experiment 3

In another attempt to observe more complete on-line translation tolerance we adapted a "double-training" procedure from Xiao et al. (2008) that has been shown to overcome retinal specificity for low-level visual discrimination tasks. Xiao et al. demonstrated that participants who had been trained to discriminate contrasts at location one showed complete transfer of this ability to location two when they had also been trained to discriminate a different stimulus dimension (orientation) at location two (otherwise, enhanced contrast discrimination was location specific). Although it remains unclear why double training leads to position

Experiments	Training locations	Displacement from nearest training location				Reduction in accuracy (Bayesian paired sample *t*-tests)		
		0°	**3°**	**6°**	**9°**	**0° vs. 3°**	**0° vs. 6°**	**0° vs. 9°**
Exp 1a	**Always 0°**	**98%** *(±5)*	**97%** *(±5)*	**96%** *(±7)*	**94%** *(±8)*	0.75	2.51	8.21
Exp 1b	**Always 0°**	**97%** *(±3)*	**94%** *(±4)*	**90%** *(±6)*	**82%** *(±12)*	12.87	43.7	92.17
Exp 1c	**Always 0°**	**93%** *(±6)*	**92%** *(±4)*	**86%** *(±7)*	**80%** *(±8)*	0.43	8.41	165
		0° (P)	**0° (C)**		**9°**	**0° (C) vs. 0° (P)**	**0° (C) vs. 9v**	**0° (P) vs. 9°**
Exp 2	**0° & −9°**	**93%** *(±4%)*	**95%** *(±5%)*		**85%** *(±9%)*	2.27	17.56	9.68
Exp 3	**0° & +/−9°** (double training)	**83%** *(±6%)*	**93%** *(±7%)*		**81%** *(±9%)*	17.4	67.38	0.71
		0°	**9°**		**18°**	**9° vs. 18°**	**0° vs. 18°**	
Exp 4a	**−9°**	Not Tested	**88%** *(±7%)*		**84%** *(±7%)*	1.63	Not Tested	
Exp 4b	**+/−9°** (Double Training)	**97%** *(±4%)*	Not Tested		**89%** *(±9%)*	Not Tested	5.5	

Table 1. Mean (±SD) Accuracy and Bayes Factors in Experiments 1 to 4. The "Training locations" column specifies retinal positions at which stimuli were trained in each experiment (degrees of visual angle from fixation). The "Displacement from nearest training location" column shows the degrees by which test stimuli were displaced from the nearest training location, and mean accuracy (±SD) is indicated below each condition. For experiments 2 and 3, 0 degrees (C) and 0 degrees (P) are shorthand to indicate whether a 0 degrees displacement was in central (C) or peripheral (P) vision, respectively.

tolerance in these low-level perceptual discrimination tasks, it raised the obvious possibility that a similar training regime would lead to improved performance with our stimuli.

Procedure

In experiment 3, we assessed identification of objects at novel test locations when those same locations were used for the training of other objects ("double-training"). As illustrated in Figure 2, all objects were trained in the fovea (until 24/24 consecutive correct answers were provided), then 12 objects were trained at 1 peripheral location, 9 degrees from the central fixation-cross (until 12/12 consecutive answers were provided) and then the remaining 12 objects were trained at a contralateral peripheral location, 9 degrees to the other side of the fixation-cross (until 12/12 consecutive answers were provided). The test phase was identical to that of experiment 2: the key question was whether objects could be recognized as accurately at the peripheral location at which they had not been trained (9 degrees), compared with the peripheral location at which they had been trained (0 degrees P).

Results

The results of experiment 3 are summarized in Table 1. The key finding is that accuracy at 9 degrees (81%) was nearly equivalent to accuracy at 0 degrees P (83%) and Bayes Factor for the paired samples *t*-test was just 0.71, indicating there was no evidence for a difference between conditions even though objects were presented for just 100 ms at test. It is perhaps also worth noting that performance at 0 degrees P (83%) was 10% lower than the equivalent test condition in experiment 2 (93%). This is likely the consequence of our more lenient training criteria used at peripheral locations in double training (12/12 consecutive correct answers required at both 0 degrees P locations as opposed to 24/24 consecutive correct answers at one 0 degrees P location required in previous experiments – see Procedure).

Experiment 4

Experiment 4 examined whether the robust on-line translation reported above could be extended to locations as distal as 18 degrees from the trained location, which is larger than any previous demonstration of on-line translation tolerance (see Introduction). Experiment 4a investigated this question using a paradigm similar to Experiment 1 and 2 (i.e. without double training) and experiment 4b investigated this question using double training (see Figure 2).

Procedure

In order to displace objects by 18 degrees, images were presented at one peripheral location during training, namely, 9 degrees right or left of central fixation. Experiment 4a followed the same training procedure as experiment 2 except that stimuli were trained at one peripheral location only and never at fixation. Two test-blocks were used, one in which test stimuli were presented at fixation (9 degrees displacement), and one which stimuli were tested on the opposite side of fixation (18 degrees displacement). For experiment 4b, we used a double training procedure (similar to experiment 3), with 12 images trained 9 degrees to the right, and the remaining 12 were trained 9 degrees to the left of central fixation. Furthermore, in an attempt to boost performance compared with experiment 3, "left" and "right" presentations were randomly interleaved within a block of 24 presentations (as opposed to being presented in separate blocks of 12) and participants were required to complete 2 separate loops of 24 of 24 consecutive correct answers (as opposed to 12/12). At test, objects were tested at two test locations: 9 degrees left, and 9 degrees right of fixation, giving 2 test conditions: 0 degrees P (0 degrees displacement from peripheral trained location) and "18 degrees" locations (18 degrees displacement from the opposite peripheral location; i.e. 9 degrees from central fixation; see Supplementary Information Table S1).

Results

Results of experiments 4a and 4b are summarized in Table 1. In experiment 4a, participants correctly categorized 84% of objects following a displacement of 18 degrees compared with 88% following a displacement of 9 degrees, with a Bayesian paired samples *t*-test indicating inconclusive evidence of a difference between these test locations. Note, the slightly higher performance following 9 degrees translation may reflect that testing took place at fixation in this condition. In experiment 4b, there was a larger drop of 8% following a displacement of 18 degrees (89%) compared with the 0 degrees P condition (97%). Nevertheless, it is worth noting that 5 of 10 participants performed over 90% following an extreme displacement of 18 degrees, with one participant scoring 96%.

Familiarity ratings of objects used in experiments 1 to 4

Although the experiments outlined above used design constraints to minimize the role of semantics in mediating performance (i.e. rather than using familiar objects, we used novel objects that had similar local

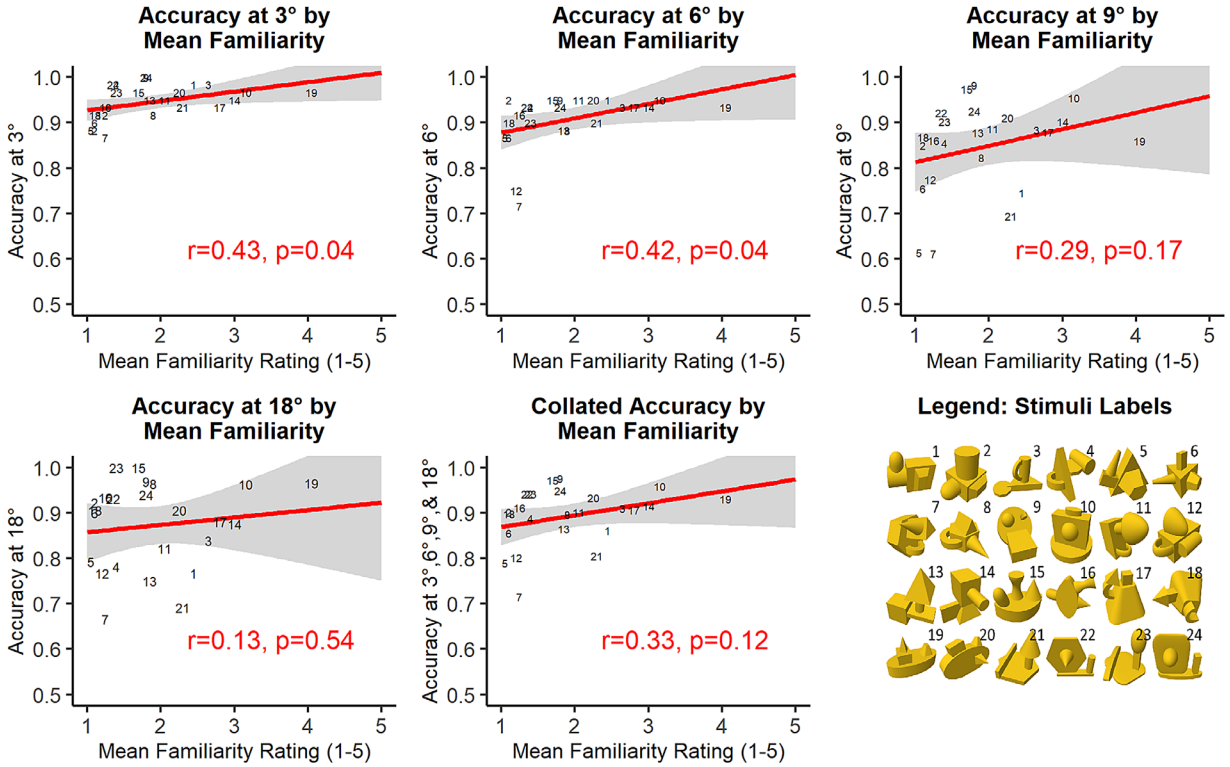

Figure 3. Each scatterplot shows the mean familiarity score of each item (averaged over participants) on the x-axis, and the Mean Accuracy Scores of each item (averaged over participants) on the y-axis. Each plot shows mean accuracy scores at a specific displacement (3 degrees, 6 degrees, 9 degrees, and 18 degrees, collated) and labels each data-point so that the performance of each item can be compared across different displacements (the rank order of items is generally respected (e.g. images 7 and 12 are usually amongst the least accurate, whereas 9 and 15 are usually the most accurate). The shaded grey zone illustrates the 95% confidence interval for the regression line (drawn using the geom_smooth function from the ggplot2 R library) (Wickham 2016).

features but different global configurations, and were presented for just 100 ms at test – and at training in the case of experiment 1c), it is difficult to rule this out completely. Indeed, O'Regan and Nazir (1990) note that even dot-patterns (which were not used in the current study given how unnaturalistic and difficult to differentiate they are) are susceptible to semantic strategies: "when asking the subjects afterwards what the nature of the target was, some of them gave a global description of the target as being like a chessman having something round on its head, or a bizarre telephone" (page 99).

To test the extent to which the current set of stimuli were considered novel, a new group of 20 participants were recruited via www.prolific.ac (Palan & Schitter, 2018), and completed an on-line questionnaire using www.gorilla.sc. Participants were instructed to "rate the extent to which each novel object resembles a familiar object on a 5-point scale (i.e., does the novel object remind you of a particular known-object in any way?)" where "1" indicated no resemblance to any familiar object, and "5" indicated strong resemblance to a familiar object. The mean familiarity score for the 24 novel objects was 1.91 (SD = 0.34). Figure 3 plots the familiarity score for each object against the mean accuracy score at each displaced location in the eye-tracking experiments (the final panel collapses across all displaced conditions). As is clear from these graphs, there is a relation between the judged familiarity of the objects and mean accuracy, but the point we would emphasize is that robust on-line tolerance was still observed with stimuli that were given the lowest familiarity ratings. Indeed, in the two experiments that assessed on-line translation tolerance at 18 degrees, the correlation between familiarity and translation tolerance was non-significant ($r = 0.13$, $p = 0.54$), and performance was approaching 90% for the 5 stimuli that were given lowest familiarity ratings. Clearly, robust on-line translation tolerance extends to objects that were judged to be completely unfamiliar.

Modelling translation tolerance in CNNs

Similar to the behavioral findings, computational studies of on-line translation tolerance have been mixed, with different models supporting a range of

Study	Training/teststimuli	Jittered trainingstimuli	Pretrained on other datasets	GAP	Accuracy at trained location	Largest displacement at test	Performance at largest displacement (chance-level was 10% unless stated)
Kauderer- Abrams (2017)	MNIST (28 × 28 px)	No	No	No	100%	± 10 px	10%
		Yes (±10 px)	No	No	100%	± 10 px	60%
Qi (2017)	MNIST (28 × 28 px)	No	No	No	100%	± 18 px	42%
		Yes (±6 px)	No	No	100%	± 18 px	98%
Furukawa (2017)	SAR satellite images (104 × 104 px)	No	No	No	100%	± 10 px	50% (chance = 20%)
		Yes (±8 px)	No	No	100%	± 10 px	80% (chance = 20%)
Chen et al. (2017)	MNIST (36 × 36 px)	No	No	No	100%	± 30 px	10%
Han et al. (2020)	24 Korean Letters (450 × 450 px)	No	Yes (MNIST ± 3150 px)	No	100%	± 3150 px	95% (chance = 50%)

Table 2. Previous studies that have examined translation tolerance in CNNs when restricting the location of the training image. px = pixels.

outcomes, from near zero to near complete translation tolerance. Table 2 details the range of outcomes with standard CNN models along with some key differences in the simulations, namely, the nature of the test stimuli, whether the models were pretrained to classify other stimuli across a range of locations, whether the test stimuli were trained in a single location or "jittered" over a small range of locations, and whether the model included a GAP layer (see below). The column called Largest Displacement at Test indicates the most distal displacement from the trained location (in pixels) at which the model was tested, and the model's accuracy at that displacement is shown in the column called Performance at Largest Displacement. Strikingly, four out of five CNNs only displaced objects by distances that were smaller than the object's own dimensions (e.g. Chen et al. displaced 36 × 36 images by up to 30 pixels), and most of these showed dramatically reduced performance at that displaced location (e.g. Chen et al.'s model performed at chance-levels following 30 pixel displacements). The only exception to this was the Han et al. (2020) study CNN that supported robust on-line translation tolerance for untrained Korean letters at displacements that were up to seven times the width of the letters. Note, only the Han et al. model was trained to classify a different set of stimuli (digits from the MNIST dataset; LeCun et al., 1998) across multiple locations, suggesting that pretraining may be a critical factor. That is, CNNs may need to learn trained tolerance on one set of stimuli before supporting on-line tolerance for a novel set of stimuli.

In addition, it is worth noting that none of the above CNNs included a GAP mechanism (Lin, Chen, & Yan, 2013) designed to provide larger receptive fields that cover the whole visual field. GAP is a hard-wired mechanism applied to each individual feature map of the final convolutional layer, and averages the values of each feature map into a single value that covers the whole visual field. Given that a GAP layer is commonly added to CNNs in order to make CNNs more robust to spatial translations of the input, this seems a relevant factor to consider as well. In the simulations below, we assessed on-line translation tolerance by training the models on the same set of 3D objects at one retinal location and then testing the model at novel locations while varying three factors: (a) pretraining versus no pretraining, (b) jitter versus no-jitter on test stimuli, and (c) GAP versus no-GAP.

Methods for modeling translation tolerance in CNNs

We systematically investigated on-line tolerance in a popular CNN (VGG16; Simonyan & Zisserman, 2014) by training the network to classify the 24 "Leek" images (Figure 2) as "A" or "B" at restricted locations, and then testing its accuracy at displaced locations equivalent to the psychophysical studies. As illustrated in Figure 2C, the Leek images were 40 × 40 pixels and were presented within 224 × 224 pixel space to allow for relatively large displacements at test (compared with most previous CNN investigations). In all simulations, training continued until the model reached 100% accuracy. We manipulated and crossed the following three factors (giving 8 simulations in total):

(i) **No pretraining versus pretraining.** The VGG16 network was either (a) trained from scratch on the 24 Leek images, or (b) pretrained on ImageNet (Russakovsky et al., 2015; by definition, this meant experiencing exemplars of categories from ImageNet across a range of retinal locations) and then trained to classify the 24 Leek images using the existing visual representations, a procedure known as "transfer learning."

(ii) **No jitter versus jitter.** The 24 Leek stimuli were either trained at (a) the trained location only, or (b) a limited range of locations (Jitter Condition), randomly displaced from the trained location by up to 20 pixels each side (thus introducing some variability as well as mimicking the behavioral studies in which fixations were free to vary by 1.5 degrees before image was masked).

(iii) **Global average pooling (GAP) versus no GAP.** The VGG16 network either (a) uses GAP on the resulting feature map from the final convolutional layer of VGG16, or (b) does not use GAP (akin to

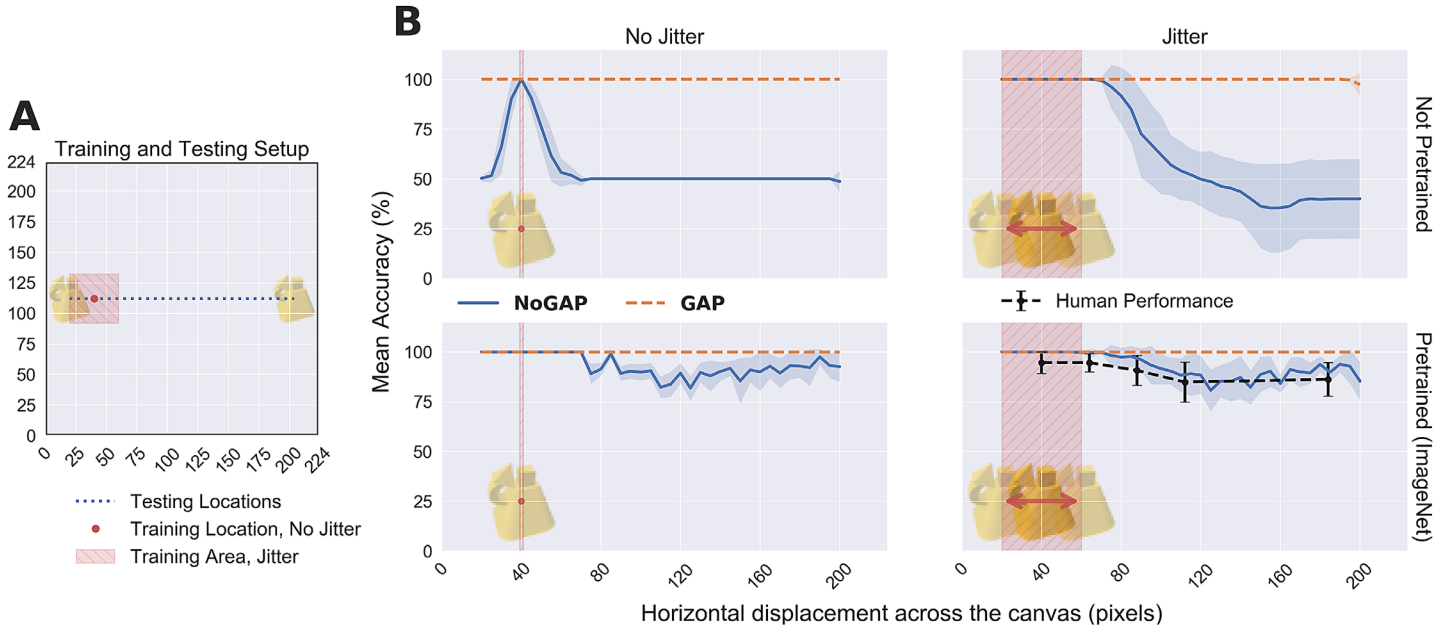

Figure 4. Mean accuracy of CNN when classifying Leek (2016) stimuli over large translations. (**A**) Illustration of the experimental setup, showing the trained and tested locations. The red dot represents the location where the training stimuli were centered for the "No Jitter" condition and the red shaded area represents the locations at which the training stimuli were centered for the "Jitter" conditions. At test, stimuli were displaced by up to 160 pixels, which is 4 times the width of the 40 × 40 pixel stimuli, thus corresponding roughly to the displacements used in our psychophysical experiments. (**B**) Illustration of CNN accuracy at test locations, crossing jitter, pretraining and GAP. Shaded areas represent one standard deviation. Black dashed line in the bottom right panel depicts human performance at analogous test locations.

the previous examinations of on-line translation tolerance in CNNs; see Table 2). At test, the Leek stimuli were displaced up to 160 pixels from the trained location. Therefore, the highest displacement was 4 times the width of the 40 × 40 pixel stimuli, similar to our psychophysical studies (which displaced 5 degrees × 5 degrees images by up to 3.6 times their width). The model's accuracy for each condition was averaged over 20 replications.

Results

Figure 4 summarizes the outcome of simulations when only horizontal displacements were used, consistent with our psychophysical experiments. When the CNN model included a GAP layer perfect on-line translation invariance was observed across all displacements, regardless of the pretraining or jitter (dashed red line in all 4 panels of Figure 4B). By contrast, when a standard CNN model was used (solid lines in Figure 4B), robust translation tolerance was only observed when the model was pretrained on ImageNet (Figure 4B, bottom panels), and following pretraining, there was a small reduction in performance following translations, with performance dropping from approximately 100% to approximately 85%. Jitter only had a minor effect on the pretrained model, but improved performance on the untrained model in the region of the jitter. To facilitate comparison between human and CNN performance, the black dashed line depicted in the bottom right panel of Figure 4B plots mean human performance (collapsing over experiments) after converting each displacement (0 degrees, 3 degrees, 6 degrees, 9 degrees, and 18 degrees) to an equivalent measurement in pixels based on the proportion of the image size.[3] Clearly, the pretrained standard CNNs (without GAP) accounts for the human data most closely, and is consistent with the fact that the humans in our experiment had extensive previous experience seeing familiar objects at multiple retinal locations.

We also repeated the simulations across a greater range of displacements. As illustrated in Figure 5, each Leek image was tested on a 19 × 19 grid in the canvas, centering every stimulus at each point of the grid. The results were averaged across 20 replications, and the untested points in the canvas were estimated through cubic interpolation. The results highlight even more clearly the limited on-line translation tolerance obtained with standard deep convolutional neural networks (DCNNs) without pretraining, the extreme on-line translation tolerance obtained with standard DCNNs that were pretrained on ImageNet, the limited

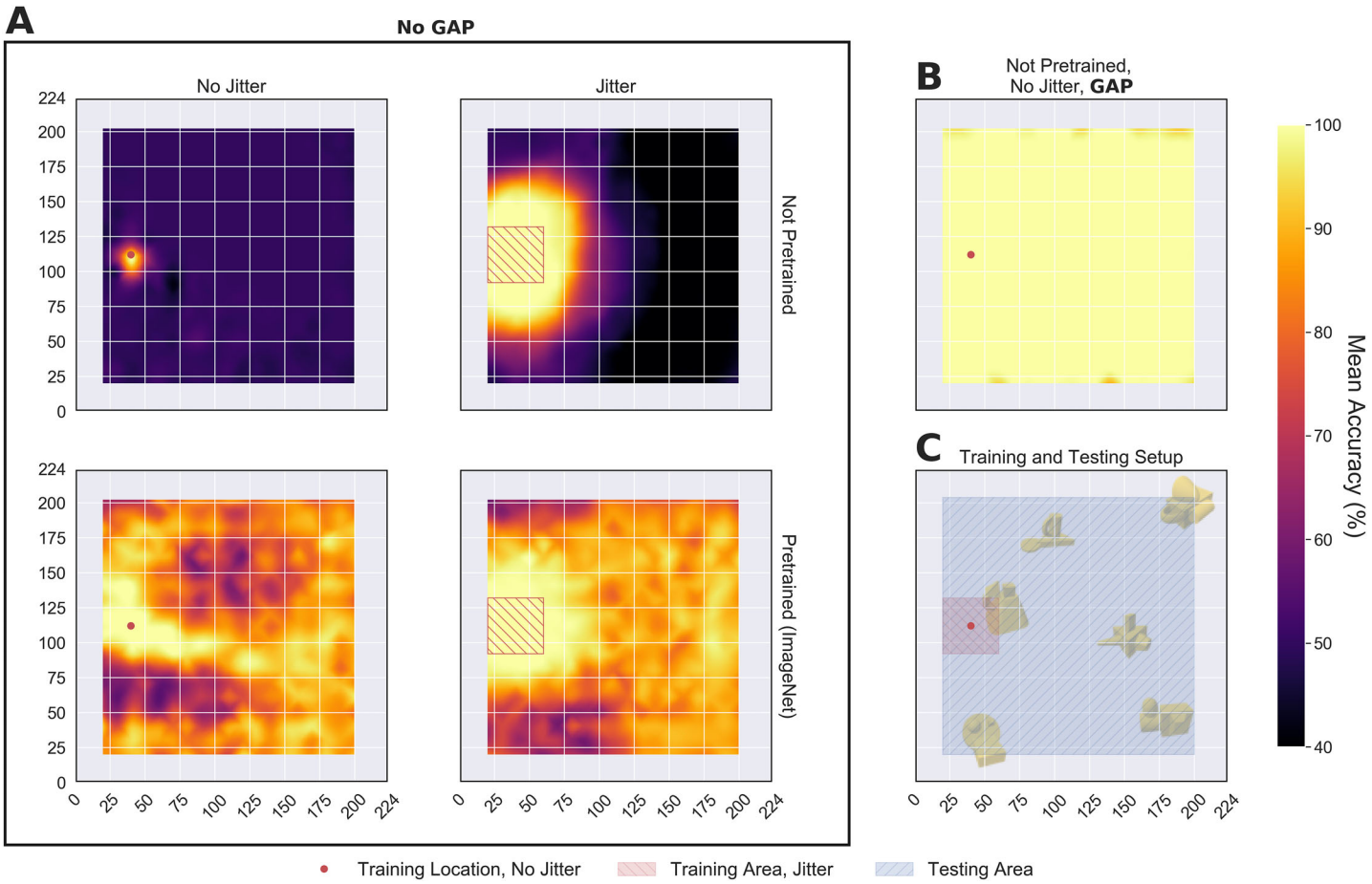

Figure 5. Mean CNN accuracy across 20 runs on the whole canvas, for GAP and no GAP conditions. Test results are expanded across the whole canvas. (**A**) No GAP conditions, with/without jitter and with/without pretraining on ImageNet. When no pretraining was performed, the network does not generalize on translation much further than the trained locations (upper panels). When the network was pretrained, the network could generalize much better across the whole canvas (bottom panels). In these panels, the small areas with lower accuracy might be the results of random features of the pretraining dataset (e.g. photographer bias). (**B**) With GAP, the accuracy was at ceiling everywhere on the canvas. Results for the conditions with pretraining and with jitter presented a similarly high accuracy, and the plots are therefore omitted. (**C**) Diagram of the experimental setup when tested across the whole canvas. The shaded areas represent the locations of the center of the image.

impact of jitter, and the compete online translation invariance obtained with DCNNs that include a GAP layer regardless of pretraining.

Discussion

In a series of behavioral experiments we demonstrate that participants trained to recognise images of novel 3D objects at one retinal position can recognise the same objects at untrained distal retinal-locations with high accuracy (up to 18°). These findings challenge the common claim that on-line translation tolerance is highly limited (Chen et al., 2017; Cox & DiCarlo, 2008; for review, see Kravitz et al., 2008), but are consistent with early priming studies (e.g. Biederman & Cooper, 1991), as well as a number of more recent studies that observed robust translation tolerance up to 13 degrees (e.g. Bowers et al., 2016). Similarly, in a series of simulation studies, we have identified conditions in which CNNs can support this degree of on-line translation tolerance, namely, when the model was pretrained on another set of stimuli presented at multiple locations, or when they included a hard-wired ("innate") mechanism that forces the networks to learn large receptive fields, in this case, a GAP layer (although other mechanisms may also support these large receptive fields). The results help reconcile the mixed findings reported in the literature (as summarized in Table 2), and highlight how pooling and convolutional operations employed in CNNs can account for the robust on-line translation tolerance we observed.

It should be noted that we consistently observed a small decrease in performance across behavioral experiments when novel objects were presented to novel locations at test. This was the case even when we controlled for the eccentricity of trained and novel locations (experiment 2), and when we used a double training procedure so that participants were practiced at identifying stimuli at the critical test locations (experiments 3 and 4b). For this reason, we can only conclude that the visual system supports extreme on-line translation tolerance rather than complete on-line translation invariance. Importantly, we consistently obtained robust translation tolerance when test stimuli were presented for 100 ms, suggesting that bottom-up perceptual processes played an important role in this tolerance. This is consistent with our simulation studies that all used CNNs that operate in a bottom-up manner.

With regard to our simulation studies, we found that a standard CNN (without GAP) needs to be pretrained to classify another set of stimuli (in this case, images from the ImageNet dataset) presented at multiple retinal locations in order to manifest robust on-line tolerance for our novel images of 3D test stimuli. That is, the standard CNNs only exhibited extensive on-line tolerance after acquiring trained-tolerance for a different set of stimuli. It is of course the case that the participants in our behavioral experiments were exposed to many familiar images at multiple retinal locations prior to learning the novel 3D images, so this constraint on on-line tolerance is psychologically plausible. Furthermore, the pretrained CNNs seem to provide a reasonable account of human performance, showing near-perfect accuracy at nearby displacements, and approximately 10% reduction in accuracy at the more distal locations. By contrast, the models with a GAP layer did not need to be pretrained in order to support on-line translation invariance. Although the CNN with a GAP layer appears to provide a better solution to on-line translation tolerance from a machine learning perspective, it does not capture the limitations of human performance.

As far as we are aware, no one has documented this link between trained- and on-line tolerance in models of vision. Indeed, as noted earlier, most models and theories of word and object identification reject robust on-line tolerance and instead assume that trained tolerance explains how humans (and monkeys) identify images across a wide range of retinal locations (e.g., Chen et al., 2017; Cox & DiCarlo, 2008; Dandurand et al., 2013; Di Bono & Zorzi, 2013; Edelman & Intrator, 2003; Elliffe et al., 2002). Although our findings challenge this conclusion, our findings are consistent with the more general point that trained-tolerance on one set of stimuli (i.e. pretraining on other categories) is a prerequisite for the on-line tolerance that we observed. Why pretraining is required for standard DCNNs but not with DCNNs with a GAP layer is an interesting question for future research. One possibility is that the pretraining allows the model to discover invariant low-level features which it then uses to categorize novel stimuli.

In other work, Han et al. (2020) observed robust on-line translation tolerance for a standard pretrained CNN. However, they argued that it provided a poor account of their behavioral findings. These authors observed robust tolerance when novel stimuli were first presented at fixation and translated 7 degrees to the periphery, but tolerance was much reduced when the novel stimuli were first presented in periphery (see Figures 1F, G). The standard, pretrained CNN did not capture this asymmetry. In order to explain their findings, they used an eccentricity-dependent neural network (or ENN), a CNN model that included multiple parallel channels that sampled the inputs at different spatial resolutions, with only the low spatial resolution channel processing images at the larger eccentricities. Although they were able to explain the asymmetry in on-line translation tolerance with this modified CNN architecture, we obtained no evidence of this asymmetry in our psychophysical studies, with performance equally impressive when trained in periphery only (see experiments 4a and 4b) as when trained at fixation only (see experiments 1b and 1c). Why Han et al. observed a different pattern of results is unclear given the many methodological differences between our behavioral studies. Whatever the reason, their model is inconsistent with the current and past results that report near complete on-line translation tolerance for stimuli first presented beyond 7 degrees eccentricity (Bowers et al., 2016; Dill & Edelman, 2001).

In future research, it will be important to explain the mixed behavioral on-line translation tolerance results obtained across studies. For example, one possible explanation for the mixed behavioral findings is that (at least) two representations can be used for object recognition: a representation of shape that is invariant to position and an episodic representation that is bound to the original viewing experience (e.g. Biederman & Cooper, 1991; Biederman et al., 2009). More research is needed to develop this theory, as well as to investigate factors that might mediate the mixed findings. For example, the role of task and stimulus complexity in on-line translation tolerance is yet to be systematically investigated, although there is evidence from other domains of invariance (e.g. pose invariance) that these factors may play a role (e.g. Tjan & Legge, 1998). Additionally, further work needs to assess other forms of invariance in CNNs, including scale, rotation in the picture plane, rotation in depth, etc., and compare with human performance in order to explore the similarities of human vision and CNNs more thoroughly. It is also worth noting that the CNN models used here and in other similar works do not generate reaction times (RTs). In fact, even though modeling RTs through

artificial networks has been substantially explored (Bogacz et al., 2006), there have not been many attempts to produce reaction times out of CNNs: the only example to our knowledge is Holmes et al. (2020), which used CNNs to parameterize the drift in a classic drift diffusion model. As we observed in experiments 1a to 1c, the time of presentation of the stimuli might play a role in human performance, and thus an improved model would take this into account.

Overall, the current simulation studies capture the on-line translation tolerance demonstrated in our psychophysical studies, and also account for the mixed results previously reported with CNNs. Our findings of extreme on-line translation tolerance in humans and CNNs undermine models and theories that posit more limited translation tolerance.

Keywords: translation tolerance, translation invariance, object recognition, convolutional neural networks, global average pooling (GAP)

Acknowledgments

Author contributions: Conceptualization and writing: R.B., V.B., I.V., C.L., and J.B. Investigation: R.B. (psychophysical experiments) and V.B. and I.V. (CNN simulations).

Funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 741134).

Commercial relationships: none.
Corresponding author: Jeffrey S. Bowers.
Email: j.bowers@bristol.ac.uk.
Address: School of Psychological Science, University of Bristol, 12a Priory Road, Bristol BS8 1TU, United Kingdom.

Footnotes

[1]All Bayesian *t*-tests used the default prior option in JASP, that is, a Cauchy distribution with spread set to 0.707 (as recommended in Wagenmakers et al., 2018). Bayes Factor robustness plots were also obtained to ensure that Bayes Factors were stable across different prior specifications (accessible via JASP output files provided at https://osf.io/jahm9/).

[2]Bayes Factors between 1 and 3 are considered weak or inconclusive evidence, BF between 3 and 10 are considered moderate evidence, and Bayes Factor above 10 are considered strong evidence (see Wagenmakers et al., 2018).

[3]For example, in psychophysical experiments, all objects were 5 degrees wide so displacements of 3 degrees, 6 degrees, 9 degrees, and 18 degrees were, respectively, times 0.6, times 1.2, times 1.8, and times 3.6 the size of the image and, thus, the equivalent displacements of the 40 pixel image used in simulations are 24 pixels, 48 pixels, 72 pixels, and 144 pixels.

References

Afraz, S.-R., & Cavanagh, P. (2008). Retinotopy of the face aftereffect. *Vision Research, 48*(1), 42–54.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review, 94*(2), 115–147.

Biederman, I., & Cooper, E. E. (1991). Evidence for complete translational and reflectional invariance in visual object priming. *Perception, 20*(5), 585–593.

Biederman, I., Cooper, E. E., Kourtzi, Z., Sinha, P., & Wagemans, J. (2009). Biederman and Cooper's 1991 paper. *Perception, 38*(6), 809–825.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review, 113*(4), 700.

Bowers, J. S., Vankov, I. I., & Ludwig, C. J. H. (2016). The visual system supports online translation invariance for object identification. *Psychonomic Bulletin & Review, 23*(2), 432–438.

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 Series, Science of Intelligence. Retrieved from: https://cbmm.mit.edu/publications/eccentricity-dependent-deep-neural-networks-modeling-invariance-human-vision.

Cooper, E. E., Biederman, I., & Hummel, J. E. (1992). Metric invariance in object recognition: a review and further evidence. *Canadian Journal of Psychology/Revue Canadienne de Psychologie, 46*(2), 191.

Cox, D. D., & DiCarlo, J. J. (2008). Does learned shape selectivity in inferior temporal cortex automatically generalize across retinal position? *The Journal of Neuroscience, 28*(40), 10045–10055.

Dandurand, F., Hannagan, T., & Grainger, J. (2013). Computational models of location-invariant orthographic processing. *Connection Science, 25*(1), 1–26.

DiCarlo, J. J., & Maunsell, J. H. R. (2003). Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. *Journal of Neurophysiology, 89*(6), 3264–3278.

Dill, M., & Fahle, M. (1998). Limited translation invariance of human visual pattern recognition. *Perception & Psychophysics, 60*(1), 65–81.

Dill, M, & Edelman, S. (2001). Imperfect invariance to object translation in the discrimination of complex shapes. *Perception, 30*(6), 707–724.

Di Bono, M. G., & Zorzi, M. (2013). Deep generative learning of location-invariant visual word recognition. *Frontiers in Psychology, 4*, 635.

Edelman, S., & Intrator, N. (2003). Towards structural systematicity in distributed, statically bound visual representations. *Cognitive Science, 27*(1), 73–109.

Elliffe, M. C., Rolls, E. T., & Stringer, S. M. (2002). Invariant recognition of feature combinations in the visual system. *Biological Cybernetics, 86*(1), 59–71.

Ellis, R., Allport, D. A., Humphreys, G. W., & Collis, J. (1989). Varieties of object constancy. *The Quarterly Journal of Experimental Psychology Section A, 41*(4), 775–796.

Fiser, J., & Biederman, I. (2001). Invariance of long-term visual priming to scale, reflection, translation, and hemisphere. *Vision Research, 41*(2), 221–234.

Furukawa, H. (2017). Deep learning for target classification from SAR imagery: Data augmentation and translation invariance. ArXiv Preprint ArXiv. Retrieved from https://arxiv.org/abs/1708.07920.

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.

Holmes, W. R., O'Daniels, P., & Trueblood, J. S. (2020). A joint deep neural network and evidence accumulation modeling approach to human decision-making with naturalistic images. *Computational Brain & Behavior, 3*(1), 1–12.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review, 99*(3), 480–517.

JASP Team. (2019). *JASP (version 0.10. 2)[computer software]*. Amsterdam, The Netherlands: University of Amsterdam.

Kauderer-Abrams, E. (2017). Quantifying translation-invariance in convolutional neural networks. ArXiv Preprint ArXiv. Retrieved from https://arxiv.org/abs/1801.01450.

Kravitz, D. J., Vinson, L. D., & Baker, C. I. (2008). How position dependent is visual object recognition? *Trends in Cognitive Sciences, 12*(3), 114–122.

Kubilius, J., Kar, K., Schmidt, K., & DiCarlo, J. (2018). Can deep neural networks rival human ability to generalize in core object recognition? In Cognitive Computational Neuroscience (Ed.), *2018 Conference on Cognitive Computational Neuroscience.* Brentwood, Tennessee, USA: Cognitive Computational Neuroscience. Retrieved from https://doi.org/10.32470/CCN.2018.1234-0.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324.

Leek, E. C., Roberts, M., Oliver, Z. J., Cristino, F., & Pegna, A. J. (2016). Early differential sensitivity of evoked-potentials to local and global shape during the perception of three-dimensional objects. *Neuropsychologia, 89*, 495–509.

Lin, M., Chen, Q., & Yan, S. (2013). Network in network. ArXiv Preprint ArXiv. Retrieved from https://arxiv.org/abs/1312.4400.

Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology, 5*(5), 552–563.

Martelli, M., Majaj, N. J., & Pelli, D. G. (2005). Are faces processed like words? A diagnostic test for recognition by parts. *Journal of Vision, 5* (1), 6.

Op De Beeck, H., & Vogels, R. (2000) Spatial sensitivity of macaque inferior temporal neurons. *The Journal of Comparative Neurology, 426*, 505–518.

O'Regan, J. K., & Nazir, T. A. (1990). Some results on translation invariance in the human visual system. *Spatial Vision, 5*(2), 81–100.

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. ArXiv Preprint ArXiv. Retrieved from https://arxiv.org/abs/1511.08458.

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*, 22–27.

Peirce, J., & MacAskill, M. (2018). *Building Experiments in PsychoPy*. Thousand Oaks, CA: Sage Publications.

Qi, W. (2018). A quantifiable testing of global translational invariance in Convolutional and Capsule Networks: Conference Submissions. Retrieved from https://openreview.net/pdf?id=SJlgOjAqYQ.

Riesenhuber, M, & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*(11), 1019–1025.

Riesenhuber, M, & Poggio, T. (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology, 12*(2), 162–168.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., & Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211–252.

Serre, T. (2019). Deep learning: the good, the bad, and the ugly. *Annual Review of Vision Science, 5*, 399–426.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. ArXiv Preprint ArXiv. Retrieved from https://arxiv.org/abs/1409.1556.

Stankiewicz, B. J., & Hummel, J. E. (2002). Automatic priming for translation- and scale-invariant representations of object shape. *Visual Cognition, 9*(6), 719–739.

Stankiewicz, B. J., Hummel, J. E., & Cooper, E. E. (1998). The role of attention in priming for left–right reflections of object images: Evidence for a dual representation of object shape. *Journal of Experimental Psychology: Human Perception and Performance, 24*(3), 732.

Strasburger, H., Rentschler, I., & Jüttner, M. (2011). Peripheral vision and pattern recognition: a review. *Journal of Vision, 11*(5), 13.

Tjan, B. S., & Legge, G. E. (1998). The viewpoint complexity of an object-recognition task. *Vision Research, 38*(15-16), 2335–2350.

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., & Verhagen, J., ... & Meerhoff, F. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25*(1), 58–76.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. New York, NY: Springer.

Xiao, L.-Q., Zhang, J.-Y., Wang, R., Klein, S. A., Levi, D. M., & Yu, C. (2008). Complete transfer of perceptual learning across retinal locations enabled by double training. *Current Biology, 18*(24), 1922–1926.

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 of the United States of America, 111*(23), 8619–8624.