Common object representations for visual production and recognition

Judith E. Fan[1,2], Daniel L. K. Yamins[3,4] and Nicholas B. Turk-Browne[1,2]

[1]Department of Psychology, Princeton University

[2]Princeton Neuroscience Institute, Princeton University

[3]McGovern Institute for Brain Research, Massachusetts Institute of Technology

[4]Department of Psychology, Stanford University

Abstract

Production and comprehension have long been viewed as inseparable components of language. The study of vision, by contrast, has centered almost exclusively on comprehension. Here we investigate drawing — the most basic form of visual production. How do we convey concepts in visual form, and how does refining this skill, in turn, affect recognition? We developed a crowdsourcing platform for collecting large amounts of drawing and recognition data, and applied a deep neural network model of visual cortex to explore the hypothesis that drawing recruits the same abstract object representations that subserve visual recognition. Consistent with this hypothesis, we discovered that drawings contain the features most important for recognizing objects in photographs, and that learning to make more recognizable drawings of objects generalizes to enhanced recognition of those objects. These findings could explain why drawing is so effective for communicating visual concepts, they suggest novel approaches for evaluating and refining conceptual knowledge, and they highlight the potential of deep networks for understanding human learning.

Keywords: communication; drawing; learning; perception and action; computer vision

Since the earliest etchings onto cave walls were made 40,000 years ago in modern-day Spain (Pike et al., 2012) and Indonesia (Aubert et al., 2014), people have devised many ways to render their thoughts in visual form, employing media ranging from stone and clay to paper and digital displays. The most basic and direct among these visualization techniques is drawing, in which a person produces marks that form an image. Drawing constitutes just one of several methods for externalizing mental representations in graphical form, an ability we term *visual production*. Drawn images predate symbolic writing systems (Clottes, 2008), are pervasive in many cultures (Gombrich, 1989), and are produced prolifically by children (Kellogg, 1969).

Why is drawing such a core aspect of human behavior and culture? For one, drawings are efficient carriers of meaning — just a few strokes can express the identity of a face (Bergmann, Dale, & Lupyan, 2013), a suggested route (Agrawala & Stolte, 2001), complex logical relations (Bauer & Johnson-Laird, 1993), or an intention to act (Galantucci, 2005). Moreover, despite large differences in the appearance of drawings of objects versus physical objects, drawings are just as recognizable (Biederman & Ju, 1988; Biederman & Gerhardstein, 1993; Gibson, 1971). Here we provide a computational framework for understanding the effectiveness of drawings in conveying visual concepts. These investigations are guided by the overarching hypothesis that the act of drawing an object recruits the same mental representation used for recognizing the object.

A first prediction of this hypothesis is that drawings — although impoverished in many ways — retain precisely those features that enable recognition of real-world objects. To test this prediction, we employed a biologically inspired deep neural network model of the ventral visual pathway (Yamins et al., 2014; Hong, Yamins, Majaj, & DiCarlo, 2016) to characterize the features expressed in drawings quantitatively. We compared these feature representations in different layers of the model to those that support the identification of objects in photos. This led to the discovery of a striking isomorphism in the representations of object categories in drawings and photos.

A second prediction is that learning how to draw might refine the representations shared by both drawing and recognition. To test this prediction, we trained people to draw a set of objects and examined, across several experiments, how their drawings of these objects improved and what effect this had on recognition of the objects. To quantify drawing performance, we assessed how well the deep neural network model could recognize the object being drawn. We found that it performed better in classifying drawings after training and that these improved drawings exhibited less feature-level overlap with each other, suggesting that practice drawing these objects had differentiated their underlying representations. This was further reflected in a psychophysical recognition task as enhanced categorical perception of the objects that people had learned to draw.

These findings provide a first direct demonstration, to our knowledge, that visual production can alter object representations. Although drawing has long been used by behavioral scientists to investigate the contents of conceptual representations in children (Minsky & Papert, 1972) and in clinical populations (Folstein, Folstein, & McHugh, 1975; Bozeat et al., 2003), prior studies have relied on qualitative assessments of drawings, which limit their explanatory power (Kosslyn, Heldmeyer, & Locklear, 1977; Karmiloff-Smith, 1990; Cohen & Bennett, 1997; Bozeat et al., 2003), or on quantitative measures of low-level properties (e.g., pixel distance and area), which fail to capture high-level object identity information (Tchalenko, 2009; Fay, Garrod, Roberts, & Swoboda, 2010; Perdreau & Cavanagh, 2014). The approach taken here of using deep convolutional neural network models to characterize the high-level properties of drawings enhances the scientific potential of visual production as a window into human cognition. While there is precedent for this kind of anlaysis in the machine learning literature (Yu, Yang, Song, Xiang, & Hospedales, 2015), predicting and explaining *human* learning and behavior with these models is innovative and could find broad applicability across many fields, including cognitive development, education, communication, and human-computer interaction.

Results

Generalized object representations

Recognition of visual objects is achieved by a set of hierarchically organized brain regions known as the ventral visual stream (Goodale & Milner, 1992; Rolls, 2000; Malach, Levy, & Hasson, 2002). Simple visual features encoded in lower areas (e.g., orientation, spatial frequency in V1) are successively combined and transformed into more complex features across levels of the hierarchy (Gross, Rocha-Miranda, & Bender, 1972; Hung, Kreiman, Poggio, & DiCarlo, 2005), allowing for read out of abstract object properties from higher areas (e.g., category, identity in inferior temporal [IT] cortex). Recently, these computations have been modeled using deep convolutional neural networks. Such models can approach human-level performance in challenging object recognition tasks and learn features that predict neural population responses in multiple sites along the ventral stream, including V4 and IT (Krizhevsky, Sutskever, & Hinton, 2012; Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014). As such, they present an attractive candidate for investigating object recognition when it requires invariance across image domains, such as the recognition of drawings.

We tested the hypothesis that training a deep convolutional neural network to recognize photographs of objects across variable views (Yamins et al., 2014; Hong et al., 2016; Fig. 1A) would provide a sufficiently robust basis set of high-level features to enable the model to also recognize drawings. We further predicted that the model's representations of drawings and photographs should become progressively more similar across successive layers, and peak at the highest layer (approximating IT), consistent with the notion that drawings possess the same abstract features used to recognize natural objects. However, because of large differences between drawings and photographs at the pixel level, we expected their representation in early layers of the model to be much more distinct.

To evaluate these predictions, we employed a deep convolutional neural network (Yamins et al., 2014; Hong et al., 2016) that had been optimized exclusively to recognize photographs of objects and had never seen a line drawing before. We then tested the model on a large number of drawings (Eitz, Hays, & Alexa, 2012) and photographs (Deng et al., 2009) of objects belonging to 105 real-world categories, none of which was included in the training set. Each image elicited a pattern of feature activations at every layer in the model, each pattern being equivalent to a vector in a feature space with the same number of dimensions as units in that layer. Separately for drawings and photographs, we averaged the feature vectors within each object class for a given layer, then computed a layer-specific matrix of the Pearson correlation distances between these average vectors across classes (Kriegeskorte et al., 2008). Each of these 105x105 representational distance matrices (RDMs) provides a compact description of the layout of objects in the high-dimensional feature space inherent to each layer of the model (Fig. 1B).

The matrices for drawings and photographs computed based on the top model layer are strikingly similar by visual inspection and as quantified by correlation (r = 0.547, $p < 0.001$). By contrast, the matrices from the bottom layer exhibited low similarity (r = 0.144, $p < 0.001$), which, although statistically reliable, was substantially lower than that of the top layer (independent correlations test, $p < 0.001$). Indeed, similarity increased over successive layers in the model (Spearman correlation of similarity with layer number, r = 0.943 $p < 0.001$; Fig. 1C). To explicitly test the model's accuracy in recognizing drawings, we trained a 105-way support vector machine (SVM) linear classifier on the feature vectors from a given model layer for 80% of the drawings in each class and tested it on the remaining drawings. Consistent with the representational distance analyses, 105-way classification was highly accurate for the top layer (64.8% vs. chance=0.95%, s.e.m. = 1.1% across 5 cross-validation splits), and substantially more so than in the bottom layer (18.6%, cross-validated, s.e.m. = 0.6%).

The model was trained to identify photographs based on human-provided labels, so we interpret this successful recognition of drawings as mirroring how humans would recognize the drawings. To validate this assumption, we recruited an independent cohort of human participants (N=327) to provide labels for drawings from a subset of these categories. As expected, human and computer recognition performance (d') was highly

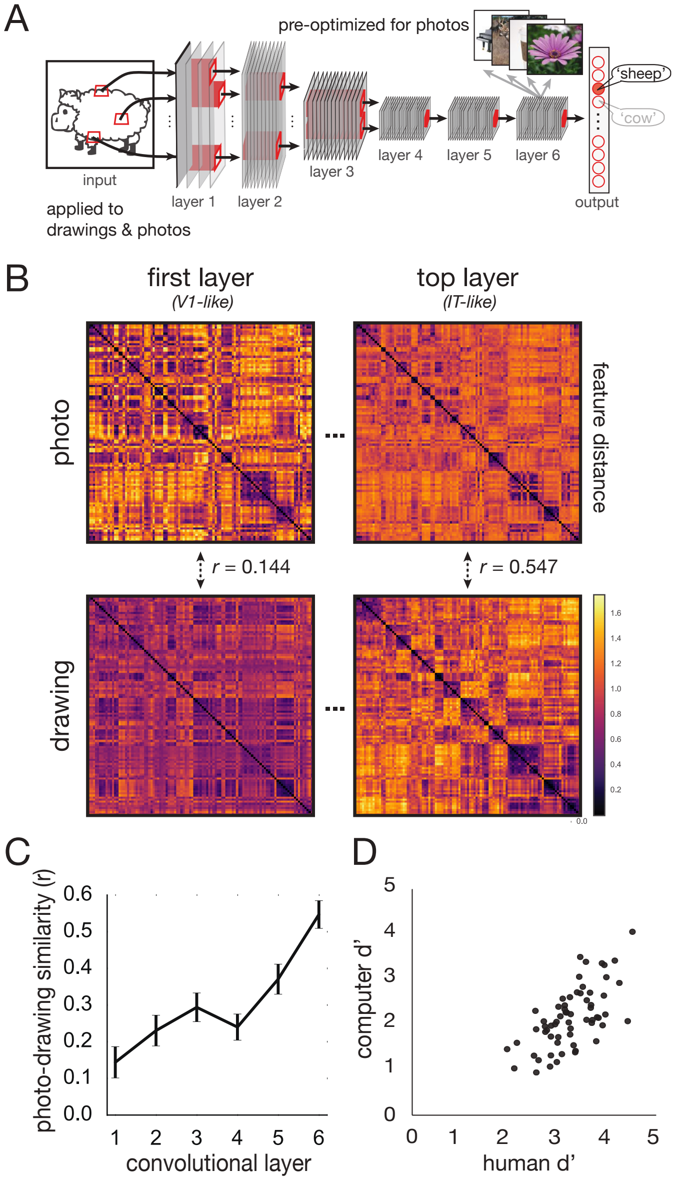

Figure 1: A: Features were extracted from drawings and photos using a neurally predictive, deep convolutional neural network model previously optimized for performance on object recognition tasks involving natural photographs but not drawings. Each feature reflects the activation of a single unit in a given layer of the neural network. B: Representational distance matrices (RDMs) of model features for each image domain, displaying the overall layout of objects in high-dimensional feature space. Each entry shows correlation distance (1-*r*) between feature vectors for a pair of objects. Darker values reflect relatively proximal pairs of objects. Reflecting the presence of higher-order structure (i.e., clustering of objects with similar features), each top-layer matrix shows clear block-diagonality. C: Cross-domain similarity between image domains increases as a function of model layer. Error bars represent 1 s.e.m., estimated by jackknife resampling of objects. D: Human participants (N=327) provided labels for drawings from a subset of these categories (64 of 105). We found that human and computer recognition performance (d') was highly consistent across objects.

consistent across objects (r = 0.649, $p < 0.001$; Fig. 1D). Thus, the model's pattern of correct identifications and confusion errors was similar to that of humans performing the same task.

Together, these results show that building a computational model that achieves the level of visual abstraction required to recognize real-world objects under high image variation yields converging feature representations for object drawings and photographs. In other words, the high-level features that support natural object recognition are captured in drawings, allowing a model embodying these features to easily generalize to artificial line drawings. This is consistent with our broader hypothesis that the abstract representation of an object formed during recognition may provide the basis for producing a recognizable image of the object by drawing. A further implication of these results is that drawings may be so effective at conveying visual concepts in part because they take advantage of computational mechanisms already in place (i.e., along the ventral stream) to extract abstract information, such as object identity, from natural visual input.

Visual production training

If drawing an object involves accessing the representation used to recognize it, then learning how to draw the object better may refine this representation and improve recognition. Here we test for improved recognition in the deep neural network model; in the next section we test for improved recognition in humans. We hypothesized that training people to draw objects would enhance the model's ability to recognize their drawings of these objects, and that this occurs because they learn to emphasize those features of an object that distinguish it from other objects. This makes the specific prediction that the model's top-layer representations of the trained objects should differentiate from each other. Such differentiation has traditionally been induced using recognition tasks (Goldstone, 1998), but here we examine it as a consequence of training in visual production.

A natural starting point for examining learning is to identify objects for which untrained participants have trouble producing recognizable drawings — that is, objects whose drawings are frequently confused by the model with drawings of other objects. To identify confusable groups of objects, we exploited clustering of objects in the model's top-layer representations of drawings (Eitz et al., 2012). From these clusters, we defined eight visual "categories" that each contained eight objects. Each participant was randomly assigned two of these categories (Fig. 2A). During training, participants drew four randomly selected objects in one category (Trained) multiple times. Before and after training (Fig. 2B), participants produced one drawing of each of those objects, of each of the other four objects in that category (Near), and of all of the objects in the second category (Far). These conditions allow us to assess the specificity of training effects: Trained objects provide a measure of object-specific learning, Near objects provide a measure of category-specific learning, and Far objects provide a baseline measure of generic task-level or motor improvement. We hypothesized that the Trained objects would become more recognizable to the model after training, relative to both their recognizability before training and to the recognizability of the Near and Far objects.

Training was conducted via an online game we developed ("Guess My Sketch"), in which participants (N=593) were repeatedly cued to convey particular object concepts to a digital avatar (representing the model) via drawings (Fig. 2C). Approximately half of the participants were cued with words and the other half with images. When the participant's drawing was submitted on each trial, the avatar listed its top three guesses as to the identity of the cued object, thus providing participants with immediate feedback about the quality of their drawing. These guesses were generated by submitting the drawing bitmap to the model running on a server in real-time and passing the top-layer responses though a 64-way classifier pre-trained on photos of the objects from all categories. The classifier returned a list of 64 margin values, corresponding to the level of confidence that the test image belonged to each object class. The rank of the cued object in this list provided a measure of the goodness-of-fit of the submitted drawing to the cued object's representation in the model and thus served as our primary index of drawing quality (lower rank means better performance).

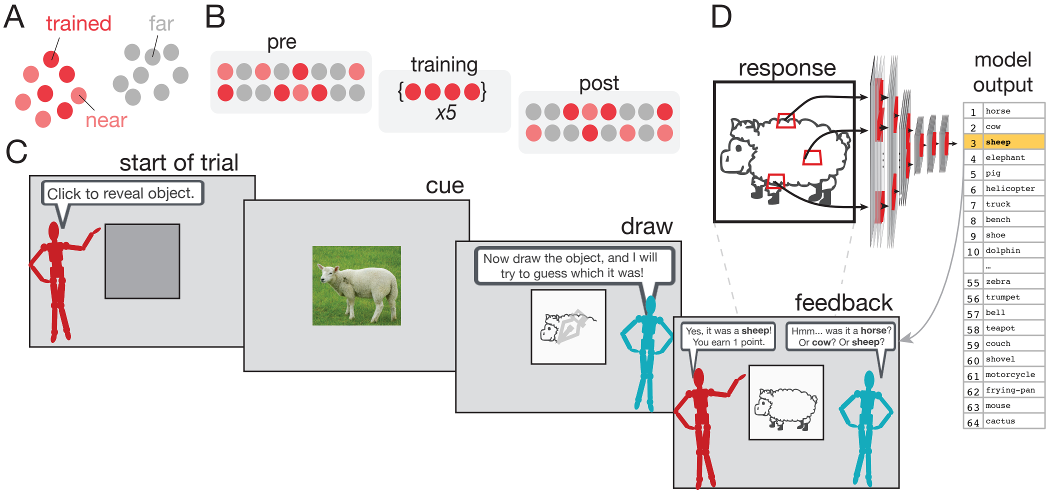

Figure 2: Each participant was randomly assigned two of eight possible object categories. B: During training, participants (N=593) practiced drawing four randomly selected objects in one category (Trained) multiple times. Before and after training, participants drew the other four objects in the same category (Near), and the objects in the second category (Far). C: On each trial, participants were cued with a trial-unique photo (N=311) or word (N=282) that referred to a target object for them to draw. The neural network model guessed the identity of the drawn object in real time, providing participants immediate feedback about the quality of their drawing. The rank of the cued object in the list of all 64 guesses returned by the neural network model, ordered by confidence, was used to track changes in performance.

Because objects were randomly assigned to condition across participants, we expected no differences in mean rank for Trained, Near, and Far objects in the pretest. Indeed, a 3 (condition: Trained, Near, Far) x 2 (cue type: word, image) repeated-measures ANOVA revealed no main effect of condition ($F_{2,587}$ = 1.16, p = 0.315). There was a main effect of cue type ($F_{2,587}$ = 14.8, $p < 0.001$), with image-cue performance (M=8.69, SD=6.55) exceeding word-cue performance (M=10.24, SD=7.58). However, cue type did not interact with condition ($F_{2,587}$=0.147, p = 0.863).

During the training phase, participants drew the four Trained objects five times each, in a randomly interleaved order. To assess changes in performance over training, we computed the mean rank across objects for each repetition and assessed its relationship to repetition number. Across participants, the trend was reliably negative (mean Spearman's r = -0.075; t = 3.37, p = 0.0007), demonstrating that drawings improved with practice (Fig. 3A).

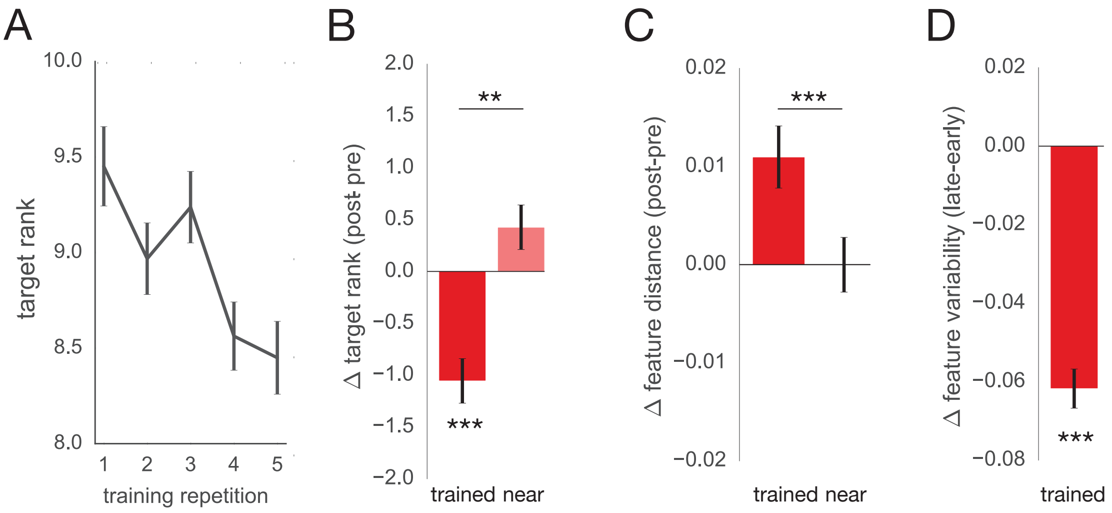

Figure 3: A: Drawing performance (mean target rank) as a function of training repetition. B: Mean change in drawing performance for Trained and Near objects, relative to control Far objects. Decrease in target rank reflects improvement. C: Change in mean feature distance among objects in Trained and Near conditions, relative to Far. D: Mean change in root-mean-squared feature distances among early drawings (first three) and late drawings (final three) of Trained objects. *$p < 0.05$. **$p < 0.01$. ***$p < 0.001$. Error bars represent within-participant s.e.m. Pairwise comparisons are based on two-sided significance tests.

To assess learning across conditions, we compared differences between pretest and posttest performance (Near and Far objects only appeared during these phases). For each object in all conditions, we calculated the change in rank ($\Delta_{rank} = rank_{post} - rank_{pre}$), then averaged these Δ_{rank} values across objects in each condition. We then performed the same type of ANOVA across participants as for the pretest analysis, revealing a significant difference in rank change between conditions ($F_{2,1182}$ = 7.67, $p < 0.001$). There was no main effect of cue type ($F_{1,591}$=1.66, p = 0.198), nor an interaction between condition and cue type ($F_{2,1182}$ = 0.162, p = 0.851), so we collapse across cue type in subsequent analyses. Drawings of Trained objects were better recognized by the model following training ($\Delta_{rank} < 0$: t_{592} = 4.04, $p < 0.001$, two-sided in this and all subsequent t-tests);

no such improvement was found for Near objects (t_{592} = 0.511, p = 0.609) or Far objects (t_{592} = 1.22, p = 0.223). The improvement for Trained exceeded that of Near (t_{592} = 3.44, $p < 0.001$) and Far (t_{592} = 2.91, p = 0.004), which themselves did not differ (t_{592} = 1.15, p = 0.252). Taken together, these results indicate that production training primarily resulted in object-specific benefits (Fig. 3B). In particular, improved classification performance for Trained objects was driven primarily by an increase in the hit rate (+5.2%, percentage of trials, pooling across participants; $p < 0.001$, bootstrap resampling of participants), with marginally significant decreases in the rate that Trained drawings were misclassified as being Near (-1.3%, p = 0.080) or Far (-1.2%, p = 0.082) objects.

The improved rank score for Trained objects shows that their high-level feature representations became more linearly discriminable. We investigated two potential sources of this differentiation (not mutually exclusive): increased feature distances among Trained objects (“separation”) or decreased feature variance of individual Trained objects (“sharpening”).

To test for separation, we first extracted the model’s top-layer feature representation of all drawings from the pretest and posttest and computed a matrix of the correlation distances between these feature vectors (Fig. 3C). Then, for each Trained object, we compared its distance before training with other Trained objects before training (pre/pre) vs. after training (pre/post). The same difference was calculated for Near and Far objects as controls. Increased distance for pre/post vs. pre/pre in the Trained condition relative to the Near and Far baselines would indicate that drawing induced separation of object representations. Consistent with this possibility, a one-way repeated-measures ANOVA revealed a main effect of condition ($F_{2,1773}$ = 3.12, p = 0.044). Planned comparisons confirmed that Trained objects separated more than Near objects (t_{592} = 3.48, p = 0.0005) and Far objects (t_{592} = 3.46, p = 0.0005), which did not differ from each other (t_{592} = 0.005, p = 0.996).

To test for sharpening, we tracked changes in the distance between feature vectors from the top layer across successive drawings of the same object during training. For each Trained object, we constructed a distance matrix relating drawings across all repetitions (Fig. 3D); this analysis was not possible for Near or Far objects because they were only drawn at the start and end of the study. We quantified change in feature variability in two ways: by comparing root-mean-squared feature distances among early drawings (first three) and late drawings (final three), and by measuring the trend in feature distances across pairs of successive drawings. We found that late drawings were reliably more similar to one another than early drawings (mean Δ =-0.061, s.e.m. = 0.005, $p < 0.001$, bootstrap resampling of participants), and that this reflected a gradual decrease in the amount by which successive drawings differed across repetitions (Spearman’s r =-0.211, s.e.m. = 0.022, $p < 0.001$, bootstrap resampling of participants).

To evaluate the respective contributions of separation and sharpening to the improvement in drawing performance (as quantified by the model), we regressed both of these measures on the change in rank for Trained vs. Far objects. Across participants, we found that sharpening (β=7.62, t_{590} = 2.53, p = 0.012) but not separation (β=0.02, t_{590} = 0.004, p = 0.997) predicted model performance. This suggests that decreased variability was most directly responsible for the increased discriminability of Trained objects.

The results so far have been interpreted as a consequence of repeatedly drawing the Trained objects. However, as these objects were being drawn, participants also received additional perceptual experience with them. This experience could have induced perceptual learning, providing an alternative explanation for the benefit for Trained over Near and Far objects, which were only encountered in the pretest and posttest.

To evaluate this alternative, we conducted a second training study. For each participant in the original cohort, we recruited a new participant to repeat the same sequence of trials. However, instead of drawing during the training phase, they were instead presented with the finished drawing produced by their matched participant from the original cohort (Fig. 4A). In some ways, this provides even more perceptual experience, as they always viewed the completed, most recognizable drawing, rather than incomplete, ambiguous versions. To keep participants engaged, they were instructed to “guess what the computer thinks the drawing looks like” by typing in a label, and they received bonus points when their guess matched the model’s top guess, which occurred on 36.7% of trials.

We found that viewing finished drawings yielded only modest changes in drawing performance for Trained objects (t_{592} = 1.72, p = 0.086), and no reliable improvement for Trained objects relative to Far (t_{592} = 0.580, p = 0.562) or Near objects (t_{592} = 1.21, p = 0.228). Again, neither Near (t_{592} = 0.160, p = 0.873) nor Far (t_{592} = 1.27, p = 0.205) objects improved relative to their pretest baseline, nor differed significantly from one another (t_{592} = 0.714, p = 0.475). Together, these results suggest that mere exposure to drawings is insufficient for improving the ability to produce recognizable drawings (Fig. 4B).

Although viewing completed drawings did not improve drawing performance, this only captures part of the perceptual experience of drawing. In particular, completed drawings are the result of composing a series of individual strokes into parts, and parts into whole drawings, over time. We hypothesized that observing these stroke-level dynamics may be more beneficial for learning how to draw, because they convey information about the procedure for composing a drawing that may be subsequently used to support successful drawing. To evaluate this possibility, another cohort of naive participants was recruited to each repeat the same sequence of trials of a participant in the original cohort, except that training now involved viewing a stroke-by-stroke replay of each drawing being produced (Fig. 4C). Again, their task was to guess the object being drawn, and they matched the model's top guess on 35.9% of trials.

We found that observing dynamic reconstructions of drawings produced reliable pre-post improvement for Trained objects ($t_{592} = 2.880$, $p = 0.004$), which significantly exceeded that of Far ($t_{592} = 2.09$, $p = 0.037$) and Near ($t_{592} = 2.55$, $p = 0.011$) objects. These findings suggest that observing the process of drawing construction improves participants' subsequent ability to make recognizable drawings of those objects they had previously observed being drawn (Fig. 4D). We again found no reliable pre-post changes in performance for Near ($t_{592} = 0.448$, $p = 0.655$) or Far ($t_{592} = 0.606$, $p = 0.545$) objects, and these conditions did not differ from one another ($t_{592} = 0.751$, $p = 0.453$).

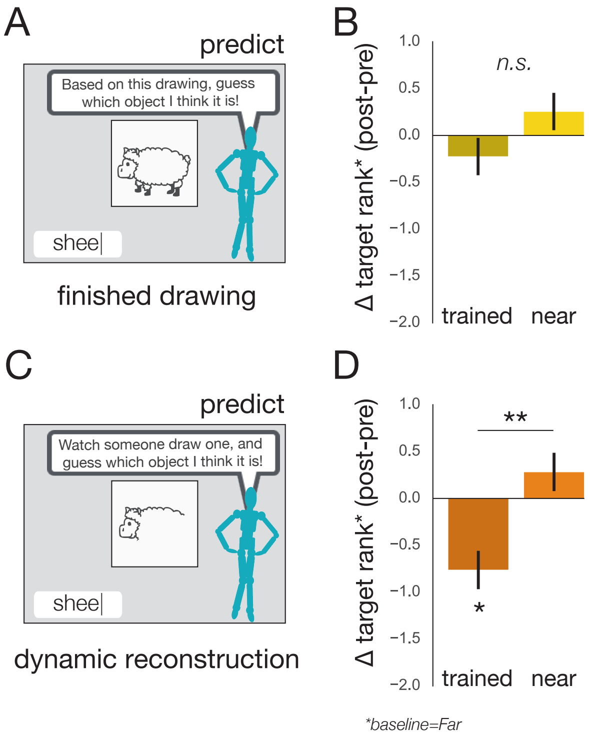

Figure 4: Two additional cohorts of participants (N=593 per cohort), each member of which was matched with a participant in the original cohort, were recruited to repeat the same sequence of trials. Rather than drawing during the training phase, these participants attempted to predict the top guess returned by the neural network model about a drawing produced by their matched participant. A: One group viewed only finished drawings. B: Change in drawing performance for Trained and Near objects, relative to control Far objects. Decrease in target rank reflects improvement. C: The other group observed stroke-by-stroke reconstructions of each drawing being produced. D: Change in drawing performance for Trained and Near objects, relative to control Far objects. *$p < 0.05$. **$p < 0.01$. Error bars represent within-participant s.e.m.

Consequences for object recognition

Above we showed that training participants how to draw objects resulted in drawings that were more recognizable to a deep neural network model of the ventral visual stream. The model was pre-trained and its parameters fixed, so these changes in the rank of the cued object can be interpreted as evidence that the object representations of participants were refined by the training. However, although suggestive, this is only indirect evidence for the claim that these refined representations are the ones in the ventral visual stream that subserve human object recognition abilities. Participants may have improved their drawings without any change in their internal visual representation of these objects (i.e., in the ventral stream), for example, as a result of object-specific motor learning of stroke sequences.

To evaluate more directly how learning to draw impacts human recognition we conducted a transfer study (N=72) in which the drawing training phase was bookended by a recognition pretest and posttest. Motivated by the earlier finding that the feature representation of Trained objects in the top, IT-like layer of the model differentiated with training, we hypothesized that drawing would increase the perceptual discriminability between Trained objects. We tested this hypothesis using a categorical perception paradigm (Goldstone, 1998; Livingston, Andrews, & Harnad, 1998), predicting that training would cause morphs between Trained objects to be perceived as more categorically distinct — that is, morphs in the middle of the range should become more consistently recognized as the majority object, resulting in a steeper slope of the psychometric function relating the morphing proportion to object labels. The rationale for this prediction is that differentiation should reduce features shared between object representations such that intermediate morphs are represented more in terms of the distinguishing features of the majority object, supporting more consistent identification of that object.

We adapted our training task in several ways to enhance our ability to measure the hypothesized perceptual changes. Estimating the parameters of psychometric curves requires many trials, so we used a reduced stimulus set of two categories with four objects each (furniture: table, bench, bed, chair; cars: sedan, limo, SUV, smartcar; Fig. 5A). To enable morphing of these objects, we commissioned an artist to design 3D models of each object with the same number of vertices per category, and then generated intermediate objects via interpolation between these endpoint objects. To increase the probability of inducing transfer effects, we also increased the number of training trials per object from 5 to 16. Finally, we removed feedback from the training phase to more cleanly isolate the consequences of production as such.

Each participant was randomly assigned one of the two categories and was trained to draw two of the objects in that category (Trained). The other two objects (Control) served as a baseline for changes in recognition. On each training trial, participants were cued with one of the Trained endpoint objects (trial-unique viewpoint) and then made a drawing of it. Before and after training, participants performed a recognition task in which they discriminated morphs of the two Trained objects or morphs of the two Control objects. On each trial of this task, participants were briefly presented with a morph and made a forced-choice judgment about which of the two objects they saw (Fig. 5B). Responses were fit with a logistic function, separately for Trained and Control pairs, and for pretest and posttest. Our key prediction concerned the slope parameter from the fitted logistic: enhanced perceptual discriminability should lead to a greater increase in slope for Trained vs. Control pairs (Fig. 5C).

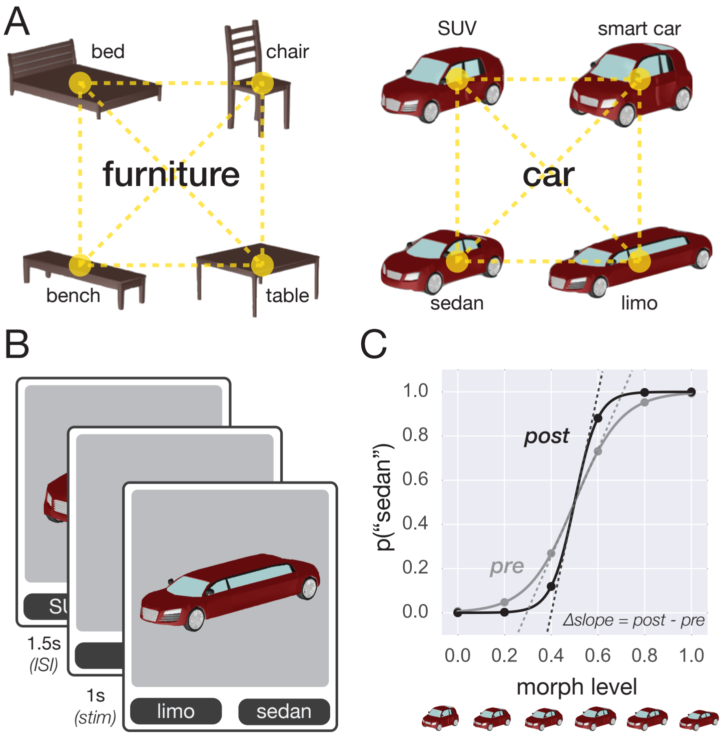

Figure 5: A: Stimuli in the categorical perception experiment belonged to two object categories (furniture, cars) containing four objects each (table, bench, bed, chair; sedan, limo, SUV, smartcar). Morphs were generated by linearly interpolating between endpoint objects. B: Before and after training, participants performed an object recognition task. On each trial, participants were briefly presented (duration = 1000ms) with a morph between either the two Trained endpoints or the two Control endpoints, and made a 2AFC judgment about which of the two objects they saw. C: Psychometric data were fit with a logistic function, whose slope parameter was predicted to increase as a consequence of training, if drawing impacts recognition.

Slope estimates did not differ between Trained and Control pairs during the pretest ($p = 0.473$, boostrapped resampling), which was expected since there was no difference between conditions prior to training. After training, the slope for the Trained pair reliably increased ($p = 0.004$; Fig. 6A), and more than for the Control pair ($p = 0.005$), whose slope did not change ($p = 0.533$). The threshold parameter did not change significantly for either condition (Trained: $p = 0.092$; Control: $p = 0.308$). These results show that visual production training can generalize to a recognition task, lending key support to our hypothesis that production and recognition engage a common high-level representation for objects.

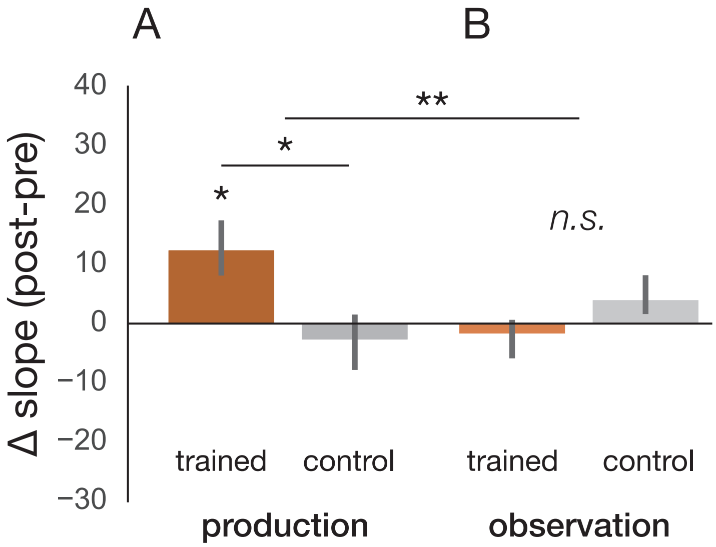

Figure 6: Mean change in slope parameter of logistic function fit to psychometric data from pretest and posttest for Trained and Control object pairs. A: Production participants (N=72) drew the two Trained objects 16 times during training. B: Observation participants (N=72) viewed stroke-by-stroke reconstructions of each drawing made by their matched Production participant. *$p < 0.05$. **$p < 0.01$. Error bars represent 1 s.e.m.

The question raised earlier about the role of perceptual experience during drawing is especially salient in this study, which employed a perceptual measure. That is, enhanced perceptual discrimination for the Trained pair may reflect perceptual learning due to greater visual exposure to these objects while they were being drawn repeatedly. To address this, we conducted another experiment modeled on the earlier dynamic replay experiment, in which a new sample of participants (N=72) viewed stroke-by-stroke reconstructions of drawings during training, but did not produce any drawings themselves. Although such observation improved drawing performance in the earlier experiment, we hypothesized that transfer to a recognition task would require more significant representational changes, as would be induced by drawing, and thus that this group might not show improved perceptual discrimination.

Indeed, there was no reliable change in slope for either the Trained pair ($p = 0.475$; Fig. 6B) or Control pair ($p = 0.332$), and no difference between these conditions ($p = 0.169$). Moreover, there was an interaction between training group and condition ($p = 0.002$), with a larger increase in slope for Trained vs. Control in the participants who were trained to draw than in those who observed somebody else drawing. The threshold parameter did not change for either condition (Trained: $p = 0.838$; Control: $p = 0.147$). These results suggest that the generalization of drawing training to perceptual discrimination was driven by aspects of visual production beyond observation of the consequences of action.

Discussion

The present study investigated the relationship between the ability to recognize objects — a biological endowment shared with other species – and to produce images of objects by drawing – a relatively recent development from human prehistory. We examined the hypothesis that these two behaviors are at least partly served by a common representational substrate for objects.

We discovered that a deep neural network model trained only on photos succeeded in recognizing drawings, suggesting that this kind of abstraction can arise from the same neural architecture evolved to make sense of natural visual inputs (Sayim, 2011). These findings provide a computational basis for understanding how line drawings can drive object recognition so effectively (Biederman & Ju, 1988; Walther, Chai, Caddigan, Beck, & Fei-Fei, 2011). We also found that learning how to draw increased the discriminability of trained object representations, as quantified by improved recognition performance and reduced feature overlap in the model and by enhanced categorical perception of trained objects in human participants. These findings are reminiscent of the way other generative behaviors such as memory retrieval (Slamecka & Graf, 1978; Crutcher & Healy, 1989; Karpicke & Roediger, 2008) and self-explanation (Chi, De Leeuw, Chiu, & LaVancher, 1994; Williams & Lombrozo, 2013) can powerfully guide learning.

The learning mechanisms responsible for such changes are not yet known, but a promising avenue forward is to build on extant theories of how differentiation between mental representations occurs. Two broad classes of candidate mechanisms may be particularly worthwhile to test: (1) strengthening of diagnostic features of objects through increased weighting of relevant dimensions (Goldstone, 1998), and (2) weakening of features that overlap between objects through competitive dynamics (Norman, Newman, Detre, & Polyn, 2006). These mechanisms are not mutually exclusive, as certain learning rules, such as non-monotonic plasticity (Lewis-Peacock & Norman, 2014), can produce both strengthening and weakening depending on the amount of activation. Nevertheless, they do make different predictions about learning outcomes under certain conditions. For instance, in the context of prolonged competition between two similar objects (e.g., alternating drawing of sheep vs. goat), the first mechanism could stabilize and refine object representations in a generalized manner, whereas the second mechanism would exaggerate differences specifically along the axis separating the competing objects in representational space.

The claim that visual production recruits and refines the same high-level object representation used during visual recognition may appear to be in tension with the distinction between vision-for-action and vision-for-recognition, which are functionally segregated into dorsal and ventral streams, respectively (Goodale & Milner, 1992). However, our findings can be reconciled with this view by considering the type of action investigated here. Namely, our drawing task involves producing a recognizable image of an object held in mind rather than reaching toward or manipulating a physical object in the world. Couched this way, the act of drawing coincides with a core function of the ventral stream – the computation of abstract, geometric properties of objects that are diagnostic of their identity (DiCarlo, Zoccolan, & Rust, 2012). We do not claim that ventral stream representations are *sufficient* for visual production, as rendering a physical image still requires translating these features into a motor program to execute the appropriate sequence of actions.

Indeed, future studies could investigate how processes engaged uniquely during visual production but not recognition affect object representations. For example, the role of motor execution could be examined by having participants trace over previously-made drawings, and the role of mental construction could be examined by having participants simulate drawing objects without producing a physical image. The rich sensory feedback generated during production may play an important role in learning, as the visual traces of current and prior movements provide an explicit basis for performance monitoring and error-based updating (Wolpert, Diedrichsen, & Flanagan, 2011; Taylor, Hieber, & Ivry, 2013). The functional value of these external traces is important to understand because they are unique to visual production and not shared with other generative cognitive processes known to influence learning, such as selective attention (Chun & Turk-Browne, 2007; Uncapher & Rugg, 2009; Fan & Turk-Browne, 2013) and memory retrieval (Slamecka & Graf, 1978; Crutcher & Healy, 1989; Karpicke & Roediger, 2008).

In the present study, the benefits of training for production and recognition were specific to the objects participants had practiced drawing. However, a hallmark of learning is generalization; thus, an important goal for future research is to understand the training conditions under which participants improve (or worsen) their ability to draw objects they did not practice, a form of skill learning also known as structural learning or learning-to-learn (Braun, Aertsen, Wolpert, & Mehring, 2009; Lake, Salakhutdinov, & Tenenbaum, 2015). Various factors merit detailed investigation, including the amount of training, variability in the objects and categories practiced (Schmidt & Bjork, 1992; Wrisberg & Liu, 1991), the type of performance feedback returned (Taylor et al., 2013; Nikooyan & Ahmed, 2015), and the effects of progressive training sequences from simpler to more complex visual forms, as advocated in classic instruction manuals (Ruskin, 1881).

Humans draw for many reasons: including to depict, to record, to plan, to explain, and to create (Tversky, 2010). Just as investigations of both verbal comprehension and production are indispensable to theories about linguistic communication, a more complete understanding of visual communication will require examining how visual recognition and production interact to support behavioral goals. Ultimately, inquiries into the psychological basis of visual production may shed new light upon the origins of symbolic writing systems for communication, and the nature of our ability to apprehend abstract meanings from visual artifacts.

Methods

Model-based feature analysis of drawings and photos

Imageset. We obtained 8,400 drawings of 105 common, real-world objects from an existing corpus (Eitz et al., 2012). From the Imagenet database (Deng et al., 2009), we acquired 22,843 photographs of the same 105 objects, depicting diverse exemplars from each object category embedded in natural backgrounds.

Computational model. We extracted their features using a deep convolutional neural network model that had been developed using hierarchical modular optimization, a procedure for efficiently searching among mixtures of convolutional neural networks for candidate hierarchical model architectures that achieve high performance on basic-level object recognition (Yamins et al., 2014). This training procedure was performed on an independent image dataset containing millions of photographs from hundreds of object categories other than the 105 categories in our study (Deng et al., 2009). In addition to approaching human-level performance in recognizing these objects, the higher layers of the model quantitatively predict neural population responses in high-level visual cortex (e.g., V4 and IT; Yamins et al., 2014; Hong et al., 2016). As such, this model was an attractive candidate for investigating object recognition invariant to image domain.

Two generations of the hierarchical convolutional neural network model were employed here. The first-generation model (Yamins et al., 2014) was used for the initial experiments, including to extract features from drawings and photographs for the representational distance analyses (Fig. 1) and to provide object-label feedback to participants in the visual production experiment (Fig. 2). The second-generation model (Hong et al., 2016), which became available partway through the study, was used for all subsequent feature analyses, owing to its superior performance, resulting from adding error backpropagation to the training of filter weights.

Representational distance analysis. Separately for drawings and photographs, we averaged the feature vectors within each object class for a given layer and then computed a layer-specific matrix of the Pearson correlation distances between these average vectors across classes (Kriegeskorte et al., 2008). Formally, this entailed computing: $RDM(R)_{ij} = 1 - \frac{cov(\vec{r}_i,\vec{r}_j)}{\sqrt{var(\vec{r}_i)\cdot var(\vec{r}_j)}}$, where $\vec{r}_i$ and $\vec{r}_j$ are the mean feature vectors for the ith and jth object classes, respectively. Each of these 105x105 representational dissimilarity matrices (RDMs) provides a compact description of the layout of objects in the high-dimensional feature space inherent to each layer of the model. Following Kriegeskorte et al. (2008), we measured the similarity between object representations in different layers by computing the Spearman rank correlations between the RDMs for those corresponding layers.

Estimates of standard error for the Spearman correlation between RDMs (i.e., between domains or between layers) were generated by jackknife resampling of the 105 object classes. This entails iterating through each of the 105 subsamples that exclude a single class, computing the correlation on each iteration, then aggregating these values. Specifically, the jackknife estimate of the standard error can be computed as: $s.e._{(jackknife)} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\bar{x}_i - \bar{x}_{(.)})^2}$, where $\bar{x}_i$ is the correlation based on leaving out the ith object class and $\bar{x}_{(.)} = \frac{1}{n}\sum_i^n \bar{x}_i$, the mean correlation across all subsamples (of size 104). This estimate of standard error allows us to construct 95% confidence intervals and compute two-sided p-values for specific comparisons (Tukey, 1958; Efron, 1979).

Classification of drawings. Model features were also used to train linear SVM classifiers (http://scikit-learn.org/) with L2 regularization to evaluate the degree to which category information was linearly accessible in each model layer. Linear classifiers determine a linear weighting of the feature activations which best predicts classification labels on a sample set of training images. Predictions are then made for images held out from the training set, and accuracy is assessed on these held-out images. The robustness of classifier accuracy scores was determined using stratified 5-fold cross validation on 80% train/20% test class-balanced splits.

Visual production training experiment

Stimuli. In order to identify groups of objects that are drawn similarly prior to training, we applied a clustering algorithm (affinity propagation with damping = 0.9; Frey & Dueck, 2007) to the features extracted from the 105-object drawing corpus described above (Eitz et al., 2012). This yielded 16 clusters containing between 3 and 20 objects each. Among clusters containing at least 8 objects, we defined 8 visual categories containing 8 objects each (Table 1). Each participant was randomly assigned two of these categories, and only the 16 objects from these two categories appeared as drawing targets during their session.

Category	Objects	Mutual Confusion Rate
1	airplane, blimp, crocodile, fish, helicopter, ship, trumpet,violin	24.4%
2	bed, bench, chair, couch, harp, ladder, laptop, table	59.0%
3	bell, frying pan, hat, pear, shoe, socks, tablelamp, teapot	29.3%
4	cat, cow, elephant, horse, kangaroo, pig, rabbit, sheep	55.7%
5	banana, dolphin, duck, mosquito, mouse, seagull, shark, swan	41.0%
6	floor lamp, fork, guitar, hammer, microphone, shovel, snake, spoon	37.1%
7	cactus, crab, giraffe, lobster, palm tree, pineapple, tiger, zebra	39.2%
8	SUV, bicycle, bus, motorbike, race car, train, truck, van	77.3%

Table 1: Objects belonged to eight visual categories, each containing eight items. These categories were derived by applying a clustering procedure to the high-level feature representation of drawings from the Eitz et al. (2012) corpus. Mutual confusion rate reflects the percentage of human labeling errors that involved another object belonging to the same category as the target object (uniform = 11%).

Experimental procedure. A total of 593 unique participants, who were recruited via Amazon Mechanical Turk (AMT), completed the experiment. Owing to the novelty of the paradigm employed in this study, we could not rely upon preexisting studies to estimate effect sizes. Instead, we used our experimental design as a guide to develop a target sample size that included a few participants (i.e., at least 5) for each of the 56 possible assignments of condition (Trained, Near, and Far) to pairs of categories (out of eight total), and for each of the image-cue and verbal-cue task conditions (see below). Participants were paid a base amount of $1.50 and up to a $3.00 bonus for high task performance. In this and all subsequent studies, participants provided informed consent in accordance with the Princeton IRB. Allocation of participants to groups was conducted anonymously via AMT, and thus the investigators were effectively blinded to the assignment of participants to group and condition during data collection.

Drawings were generated as part of a game ("Guess My Sketch") in which participants earned points by communicating visual concepts to a blue avatar (Fig. 2C). Participants initiated each trial by clicking a central gray square (500 x 500 pixels). Then, a red avatar cued participants to draw an object with either a photo (N=311) or word (N=282). In the image-cue condition, each trial used a unique photograph of the target object on a natural background and participants were instructed to "make a drawing in which someone else is likely to recognize the object depicted " but were informed that the drawing did not have to exactly depict what was in the photo. In the verbal-cue condition, the label of the target object appeared below the square. After cue offset the blue avatar appeared, prompting the participant to begin drawing.

Drawing responses were collected on a digital canvas (500 x 500 pixels) embedded in a web browser window using Raphael Sketchpad (https://ianli.com/sketchpad). Participants drew in black ink (pen width = 5 pixels) using the mouse cursor, and were not able to delete previous strokes. There were no restrictions on how long participants could take to make their drawings, and on average they spent 30.7 s per drawing (95% confidence interval = 8 - 86s). After clicking a submit button, the blue avatar listed its top three guesses as to the identity of the drawn object, thus providing participants with immediate feedback about the quality of their drawing. Participants earned points if any of these guesses were correct, proportional to its position in the top three.

These guesses were produced from a 64-way support vector machine (SVM) linear classifier. The classifier was pre-trained on model responses to photographs of the objects used in this study. On each drawing trial, the features of the submitted drawing were extracted from the model's top layer in real time and passed through the classifier. The output was a list of 64 margin values, corresponding to the level of confidence that the drawing belonged to each object class. The three objects with the most positive margin values (highest confidence) were returned to the participant as guesses. In the verbal-cue condition, when none of the top three guesses were correct, the rank of the target object in this ordered margin list was also returned to the participant (e.g., “Too bad...‘giraffe’ would have been my 9th guess.”). This target rank provides a non-parametric measure of the goodness-of-fit of the submitted drawing to the representation of the target object in the model, and so it was used as the primary dependent measure of drawing quality.

Model validation: category assignments. To validate the model’s representations of the drawings from this experiment, we extracted their features from the top layer. Separately for the image-cue and verbal-cue conditions, we computed the average feature vectors for all drawings of the same object and computed correlation matrices of these average vectors. In both cue conditions, the objects within each of the 8 categories were more similar to each other (image-cue: r = 0.633, verbal-cue: r = 0.623) than to objects in other categories (image-cue: r = -0.083, verbal-cue: r = -0.072; ps<0.001 based on object-level resampling). The matrices from the two cue conditions were also highly similar to each other (Spearman’s r = 0.897), showing that the model successfully captured object identity in both task settings. Moreover, both matrices were highly similar to that computed based on top-layer features of drawings of these 64 objects from the Eitz et al. (2012) drawing corpus (image-cue/original: r = 0.789; verbal-cue/original: r = 0.818).

Model validation: measuring human drawing recognition. To validate object-label feedback from the classifier, an independent cohort of human participants (N=327) provided three labels for each drawing from the image-cue condition, in order of confidence, from the set of 64 object labels. To compute d’ and compare human and model recognition performance, we analyzed only the top label provided by humans and the model. We found that human and model recognition (d’) were highly consistent across objects (Spearman’s r = 0.649, Fig. 1D). We additionally computed the mutual confusion rate, defined as the percentage of (first-guess) labeling errors that involved another object belonging to the same category as the target object. If these errors were spread uniformly over all distractors, the expected mutual confusion rate would be: 7/63 = 0.11. All categories exhibited mutual confusion rates reliably above uniform responding (across objects within category: t_7s > 2.76, ps < 0.028), which further validates category assignments.

Control experiment: viewing finished drawings. 593 participants were recruited via AMT and paid a base amount of $2.00 and up to a $3.00 bonus for high task performance. Each participant was matched with one of the original participants and received the same sequence of trials. On each training trial, they were presented with the finished drawing produced by the matched participant. Participants had the goal of predicting the model’s top guess for the drawing (this deviated from the target label on approximately 60% trials). Participants typed their response into a text field, and only the labels of the 64 objects in the set were accepted. They had to wait 4000 ms before being able to submit their response, to encourage them to pay attention to the criteria used by the model to classify drawings. Participants drew all objects once each before and after training, allowing us to measure the consequences of viewing finished drawings on drawing performance.

Control experiment: observing stroke dynamics. 593 participants were recruited via AMT and paid a base amount of \$2.00 and up to a \$3.00 bonus for high task performance. Again, each participant was matched with one of the original participants and received the same sequence of trials. On each trial, they observed a stroke-by-stroke reconstruction of the drawing produced by their matched participant, with a stroked added every 500 ms. They performed the same prediction task with 4000-ms waiting period from the start of the animation, ensuring that at least eight strokes appeared (or all of the strokes if eight or fewer). Participants drew all objects once each before and after training, allowing us to measure the consequences of observing stroke dynamics on drawing performance.

Statistics. Before performing statistical tests, we visualized data and examined assumptions. Quantile-quantile plots revealed a reasonable approximation to normality, an assumption of the paired t-test. Mauchly's test of sphericity indicated that the assumption of sphericity for the repeated-measures ANOVAs had not been violated. All p-values reported are two-sided. We also found that employing non-parametric analysis techniques (i.e., bootstrap resampling) gave similar results, suggesting that the choice of test (and assumptions therein) does not impact our conclusions. We did not have a prespecified way of handling outliers in this study, so we report analyses with all data included.

Categorical perception experiment

Stimuli. The four objects in each of the two categories were selected to allow the construction of a quartet of 3D mesh models (Autodesk Maya) sharing the same vertices, thereby enabling quantitatively precise morphing and full control over category-orthogonal image parameters (e.g., pose, size, background). This resulted in six axes ('morphlines') connecting all pairs of objects within each category and 12 total axes for both categories. For each axis, we derived a perceptually uniform space from which to sample morphs, in order to increase sensitivity for measuring the slope of the psychometric function. As a first step, we rendered a series of 12 morphs for each axis, linearly interpolated between the endpoint objects. Each morph was rendered from a 10° viewing angle (i.e., slightly above) at a fixed distance on a gray background in 40 viewpoints (i.e., each rotated by an additional 9° about the vertical axis). We recruited a separate cohort of 40 participants via AMT to provide 288 identification judgments each for random subsets of these morphs, yielding 80 baseline judgments per morph (two per viewpoint). For each morph (e.g., sedan/limo), we computed the proportion of trials that the morph was identified as being one of the endpoint objects (e.g., sedan) and fit these data with a logistic function to derive a population psychometric curve. We used this curve to estimate the morphing levels that produced 0% (or the minimum), 20%, 40%, 60%, 80%, and 100% (or the maximum) identifications as one of the endpoint objects. The resulting six morphs per axis evenly spanned the subjective transition between endpoints and were included in the training study.

Category	Objects
Cars	sedan, limo, SUV, smartcar
Furniture	table, bench, bed, chair

Table 2: Endpoint objects included in transfer experiments.

Experimental procedure. Based on initial piloting, we developed a target sample size of 72 participants, across whom all condition and object assignments would be fully counterbalanced. Of the original group of 72 participants recruited from AMT, 25 were excluded because their data could not be fit with a logistic function in at least one condition either before or after drawing training. This occurred either because of non-monotonicity in their psychometric curves (i.e., inconsistent responding) or hypersteepness of the slope parameter (i.e., approaching infinity). We recruited additional participants to fill these sessions to ensure that the design was counterbalanced. In total, 97 participants completed the task and received $5.00.

At the beginning of each training session, participants were familiarized with each of the eight endpoint objects that might appear. On each familiarization trial, an animation of one of the objects continuously rotating was played. Objects appeared on a gray background at the same viewing angle, distance, and viewpoints used in the experiment. The name of the object was displayed in large font above the animation ("This is a SEDAN."). The participant viewed the object completing at least one full rotation (6 s) before proceeding to the next object.

After familiarization, participants completed three phases of the experiment: pretest discrimination, drawing training, and posttest discrimination. During training, participants practiced drawing two objects (Trained) from one category; the other two objects (Control) from that category served as a baseline. The Trained objects were drawn 16 times each during the training phase (32 total trials), in a randomized order and cued with trial-unique viewpoints. No feedback was provided but the task was otherwise identical to the earlier drawing training study.

Before and after training, participants were tested on perceptual discrimination of the Trained object pair, as well as the unpracticed Control object pair. All 12 morphs (six morphs per pair) were shown 12 times each during both the pretest and posttest, always from a trial-unique viewpoint. On each trial, participants were briefly presented (1000 ms) with the morph and made a forced-choice judgment about which of the two objects they saw by clicking one of two labeled buttons that appeared below the image. The assignment of labels to buttons was randomized across trials. Participants who did not achieve greater than 80% accuracy on the unambiguous 0%/100% endpoint objects in the pretest phase did not proceed to the training phase.

For each participant and for both the pretest and posttest, we constructed psychometric curves for the Trained and Control object pairs, relating different morphing levels to the proportion of trials (out of 12) in which a given object was chosen. These curves were fit with a logistic function $p(x) = \frac{1}{1+e^{-k(x-x_0)}}$ using the Levenberg-Marquardt algorithm as implemented in Scipy (https://www.scipy.org/), where $p(x)$ is the proportion of trials on which the first object was chosen, k represents the slope, and x_0 the midpoint of the sigmoid. If drawing enhances the discriminability of Trained objects, then morphs in the middle of the range to become more consistently recognized as the majority object, and k for the Trained pair should increase from pretest to posttest more than for the Control pair.

Control experiment: observing stroke dynamics. For each of the 72 sessions in the main experiment, we recruited a naive participant to repeat the same sequence of trials, except that during the training phase, they observed a stroke-by-stroke reconstruction of the drawing produced by their matched participant, with a stroked added every 500 ms. Of this initial new cohort of 72 participants, 16 were excluded because their data could not be fit with a logistic function in at least one condition either before or after drawing training. As in the main experiment, we recruited additional participants to fill these sessions to ensure that the design was counterbalanced. In total, 88 participants completed the task and received $5.00. Before and after training, all participants were tested on perceptual discrimination of the Trained and Control object pairs.

Statistics. Before performing statistical tests, we visualized data and examined assumptions. Quantile-quantile plots revealed that data did not follow a normal distribution, so classical inference tests (e.g., ANOVA, *t*-test) that rely upon assumptions of normality were not appropriate. Instead, we employed bootstrap resampling (Efron & Tibshirani, 1986) to construct 95% confidence intervals and compute p-values for key parameters and comparisons of interest (i.e., change in slope from pretest to posttest). This entailed resampling 72 participants' worth of data with replacement, then computing the mean, on each of 10,000 iterations, for each experiment. The two-sided p-value was defined as the proportion of these iterations on which this mean fell below zero, multiplied by two.

Code availability

The code for the analyses presented in this article will be made publicly available in a figshare repository upon acceptance of this manuscript, and shared with editors and reviewers upon request.

Data availability

The data presented in this article will be made publicly available in a figshare repository upon acceptance of this manuscript, and shared with editors and reviewers upon request.

Materials and correspondence

All correspondence regarding this manuscript and requests for materials may be addressed to Judith Fan (jefan@princeton.edu), Peretsman-Scully Hall, Princeton University, Princeton, NJ 08544. Starting February 1, 2017, please note that J.E.F.'s primary affiliation will be the Department of Psychology at Stanford University (jefan@stanford.edu; Jordan Hall, 450 Serra Mall, Stanford, CA 94305).

Competing financial interests

The authors declare no competing interests.

Acknowledgments

This work was supported by NSF GRFP DGE-0646086, NIH R01 EY021755, R01 MH069456, and the David A. Gardner '69 Magic Project at Princeton University. Thanks to Ken Norman, Jordan Suchow, Jordan Taylor, and the Turk-Browne lab for helpful comments. Thanks also to NVIDIA for a grant of GPU hardware, and to Lauren Aleza Kaye for designing the 3D morph stimuli.

Author contributions

J.E.F. and D.L.K.Y. formulated initial idea, performed computational modeling, and analyzed data. J.E.F. performed the experiments. J.E.F., D.L.K.Y., and N.B.T.-B. designed experiments, planned analyses, interpreted results, and wrote the paper.

References

Agrawala, M., & Stolte, C. (2001). Rendering effective route maps: improving usability through generalization. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques* (pp. 241–249).

Aubert, M., Brumm, A., Ramli, M., Sutikna, T., Saptomo, E. W., Hakim, B., … Dosseto, A. (2014, September). Pleistocene cave art from Sulawesi, Indonesia. *Nature*, *514*(7521), 223–227.

Bauer, M. I., & Johnson-Laird, P. N. (1993). How diagrams can improve reasoning. *Psychological Science*, *4*(6), 372–378.

Bergmann, T., Dale, R., & Lupyan, G. (2013). The impact of communicative constraints on the emergence of a graphical communication system. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1887–1992).

Biederman, I., & Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(6), 1162.

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive Psychology*, *20*(1), 38–64.

Bozeat, S., Lambon Ralph, M. A., Graham, K. S., Patterson, K., Wilkin, H., Rowland, J., … Hodges, J. R. (2003, January). A duck with four legs: Investigating the structure of conceptual knowledge using picture drawing in semantic dementia. *Cognitive Neuropsychology*, *20*(1), 27–47.

Braun, D. A., Aertsen, A., Wolpert, D. M., & Mehring, C. (2009, February). Motor Task Variation Induces Structural Learning. *Current Biology*, *19*(4), 352–357.

Chi, M., De Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science: A Multidisciplinary Journal*, *18*(3), 439–477.

Chun, M. M., & Turk-Browne, N. B. (2007, April). Interactions between attention and memory. *Current Opinion in Neurobiology*, *17*(2), 177–184.

Clottes, J. (2008). *Cave art*. Phaidon London.

Cohen, D. J., & Bennett, S. (1997). Why can't most people draw what they see? *Journal of Experimental Psychology: Human Perception and Performance*, *23*(3), 609.

Crutcher, R. J., & Healy, A. F. (1989). Cognitive operations and the generation effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 669-675.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009.* (pp. 248–255).

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

Efron, B. (1979). 1977 Rietz Lecture - Bootstrap Methods - Another Look at the Jackknife. *Annals of Statistics*, *7*(1), 1–26.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 54–75.

Eitz, M., Hays, J., & Alexa, M. (2012). How Do Humans Sketch Objects? *ACM Transactions on Graphics (TOG)*, *31*(4), 44.

Fan, J. E., & Turk-Browne, N. B. (2013). Internal attention to features in visual short-term memory guides object learning. *Cognition*, *129*, 2.

Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive Science*, *34*(3), 351–386.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975, November). ”Mini-mental state”. A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189–198.

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, *315*(5814), 972–976.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, *29*(5), 737–767.

Gibson, J. J. (1971, January). The Information Available in Pictures. *Leonardo*, *4*(1), 27–35.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, *49*(1), 585–612.

Goodale, M., & Milner, A. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25.

Gross, C. G., Rocha-Miranda, C. E. d., & Bender, D. B. (1972). Visual properties of neurons in inferotemporal cortex of the macaque. *Journal of Neurophysiology*, *35*(1), 96–111.

Hong, H., Yamins, D. L. K., Majaj, N. J., & DiCarlo, J. J. (2016, February). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, *19*(4), 613–622.

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, *310*(5749), 863–866.

Karmiloff-Smith, A. (1990). Constraints on representational change: Evidence from children’s drawing. *Cognition*, *34*(1), 57–83.

Karpicke, J., & Roediger, H. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966-968.

Kellogg, R. (1969). *Analyzing children’s art*. National Press Books Palo Alto, CA.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Comput Biol*, *10*(11), e1003915.

Kosslyn, S. M., Heldmeyer, K. H., & Locklear, E. P. (1977). Children's drawings as data about internal representations. *Journal of Experimental Child Psychology*, *23*(2), 191–211.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., … Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097–1105).

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Lewis-Peacock, J. A., & Norman, K. A. (2014). Competition between items in working memory leads to forgetting. *Nature Communications*, *5*.

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(3), 732.

Malach, R., Levy, I., & Hasson, U. (2002). The topography of high-order human object areas. *Trends in Cognitive Sciences*, *6*(4), 176–184.

Minsky, M., & Papert, S. (1972). Artificial intelligence progress report.

Nikooyan, A. A., & Ahmed, A. A. (2015, January). Reward feedback accelerates motor learning. *Journal of Neurophysiology*, *113*(2), 633–646.

Norman, K. A., Newman, E., Detre, G., & Polyn, S. (2006). How inhibitory oscillations can train neural networks and punish competitors. *Neural Computation*, *18*(7), 1577–1610.

Perdreau, F., & Cavanagh, P. (2014). Drawing skill is related to the efficiency of encoding object structure. *i-Perception*.

Pike, A. W. G., Hoffmann, D. L., Garcia-Diez, M., Pettitt, P. B., Alcolea, J., De Balbin, R., … Zilhao, J. (2012, June). U-Series Dating of Paleolithic Art in 11 Caves in Spain. *Science …*, *336*(6087), 1409–1413.

Rolls, E. T. (2000). Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron*.

Ruskin, J. (1881). *The elements of drawing: in three letters to beginners*. Wiley.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*(4), 207–217.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 592-604.

Taylor, J., Hieber, L. L., & Ivry, R. (2013). Feedback-dependent generalization. *Journal of Neurophysiology*.

Tchalenko, J. (2009). Segmentation and accuracy in copying and drawing: Experts and beginners. *Vision Research*.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics*.

Tversky, B. (2010, August). Visualizing Thought. *Topics in Cognitive Science*, *3*(3), 499–535.

Uncapher, M., & Rugg, M. D. (2009, June). Selecting for Memory? The Influence of Selective Attention on the Mnemonic Binding of Contextual Information. *Journal of Neuroscience*, *29*(25), 8270–8279.

Walther, D. B., Chai, B., Caddigan, E., Beck, D. M., & Fei-Fei, L. (2011). Simple line drawings suffice for functional mri decoding of natural scene categories. *Proceedings of the National Academy of Sciences*, *108*(23), 9661–9666.

Williams, J. J., & Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. *Cognitive Psychology*, *66*(1), 55–84.

Wolpert, D. M., Diedrichsen, J., & Flanagan, J. R. (2011). Principles of sensorimotor learning. *Nature Reviews Neuroscience*, *12*(12), 739–751.

Wrisberg, C. A., & Liu, Z. (1991). The effect of contextual variety on the practice, retention, and transfer of an applied motor skill. *Research Quarterly for Exercise and Sport*, *62*(4), 406–412.

Yamins, D. L., 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*, 201403112.

Yu, Q., Yang, Y., Song, Y.-Z., Xiang, T., & Hospedales, T. (2015). Sketch-a-net that beats humans. *arXiv preprint arXiv:1501.07873*.