Using drawings and deep neural networks to characterize the building blocks of human visual similarity

Kushin Mukherjee and Timothy T. Rogers

Department of Psychology & Wisconsin Institute for Discovery
University of Wisconsin-Madison

Early in life and without special training, human beings discern resemblance between abstract visual stimuli such as sketch drawings and the real-world objects they represent. We used this capacity for visual abstraction as a tool for evaluating deep neural networks (DNNs) as models of human visual perception. Contrasting 5 contemporary DNNs, we evaluated how well each explains human similarity judgements among line drawings of recognizable and novel objects. For object sketches, human judgements were dominated by semantic category information; DNN representations contributed little additional information. In contrast, such features explained significant unique variance perceived similarity of abstract drawings. In both cases, a vision transformer trained to blend representations of images and their natural-language descriptions showed the greatest ability to explain human perceptual similarity—an observation consistent with contemporary views of semantic representation and processing in the human mind and brain. Together the results suggest that the building blocks of visual similarity may arise within systems that learn to use visual information, not for specific classification, but in service of generating semantic representations of objects.

Drawing | Deep Learning | Perception | Similarity Judgements | Multi-modal models

Correspondence: *kmukherjee2@wisc.edu*

Introduction

A central question for theories of visual perception and cognition concerns the nature of the features the visual system deploys to represent its inputs and of the processes it uses to assemble these into a recognized object in the world. Much work in this area has understandably focused on explaining visual perception/recognition of naturalistic inputs, such as color photographs of objects or scenes. Yet human vision is also remarkable in its capacity to perceive, recognize, and make inferences about even highly abstract stimuli that depart radically from the veridical visual structure of the real world, from cave drawings to illustrations in children's books to expressionist paintings to figures in scientific papers.

Specifically, the ability to discern resemblance between drawings and real-world objects develops early and without special training in infancy: children as young as 5 months discern the similarity between a photograph and line drawing depicting the same face (1, 2), and drawing recognition is generally robust in childhood (3, 4). It also appears special to human cognition: adult chimpanzees can generalize learned responses across photographic depictions of object classes, but do not extend this generalization to line drawings or other abstract depictions of the same objects (5); pigeons, despite their famed capacity for visual recognition, show the same pattern (6). Drawings thus offer a useful opportunity for testing different proposals about the building-blocks of human visual cognition: whatever features and processes the visual system develops to support perception and recognition of objects in the real world must also extend to explain perception and recognition of abstract object depictions in drawings and other visual media.

The current paper uses people's ability to perceive similarities between simple line drawings as a tool for evaluating a class of vision models that has garnered sustained interest across the related disciplines of machine vision, visual neuroscience, and visual cognition, namely deep neural networks (DNNs). Such models have been applied to several problems including image captioning (7), answering questions about a given image using natural language (8, 9), generating sketches (10), and even solving entire families of visual tasks (11). Cognitive science and visual neuroscience, however, have focused primarily on deep image classifiers: models trained via gradient descent to assign objects shown in millions of photographs into one of 1000 possible mutually-exclusive categories (12–14). From the perspective of human visual cognition, such models are interesting because they generalize well to images depicting new examples of the trained classes (15) and thus offer a potential mechanism for understanding key phenomena such as recognition invarance across category exemplar, viewpoint, spatial location/orientation, lighting conditions, etc., and how these abilities may be acquired via learning from the visual structure of the environment. From the perspective of neuroscience, the models are interesting partly because the internal representations they acquire resemble, in certain ways, the patterns of neural activity evoked by visual stimuli in the ventral processing streams of both humans and nonhuman primates (12, 13, 16–18).

Perhaps surprisingly, some deep image classifiers, despite being trained exclusively on photographs, nevertheless acquire internal representations that capture a degree of similarity between sketches and photographs depicting the same class of objects (19, 20). In learning to categorize photorealistic images, such models thus appear to acquire feature representations and mechanisms for combining them that extend, at least to some extent, to abstract depictions of objects like those appearing in line drawings. Taken together, these observations suggest that deep image classifiers may provide a useful tool for connecting computational, cognitive, and neuroscientific accounts of visual object processing.

Yet there are also many reasons for questioning the utility of DNN image classifiers as scientific models of human visual cognition:

The features DNNs acquire are opaque. It is notoriously difficult to understand precisely what information in the input neural networks models exploit across different layers in exhibiting the behaviors that they do. While some researchers have proposed heuristics for tackling this question (21, 22) and others have investigated inductive biases in such models (23, 24), it remains unclear exactly what kinds of visual features DNNs acquire. Besides DNNs, machine vision also offers many more transparent techniques for characterizing the 'low-level' visual information expressed in an image or drawing, and little work has assessed whether DNN-derived features capture important aspects of human perception beyond those already expressed by these other easier-to-comprehend methods (25).

There are many different DNN architectures and training methods. Contemporary interest in DNNs as models of human perception began with convolutional networks (26), which represented a step change in classification accuracy while also possessing some resemblances to the object-processing visual stream in the human brain–for instance, an organization in which both feature complexity and receptive field size increase from earlier to later processing stages. Today, however, newer architectures that bear little clear relation to ventral visual stream often perform better on benchmark tasks (e.g. transformer models (27)); recently-introduced heuristics for training models (e.g. contrastive methods such as CLIP (28)) appear to have a larger effect on their behavior than does the architecture *per se*; and models with qualitatively distinct architectures appear to capture macro-scale neural patterns in ventral visual stream about equally well (29), despite behaving according to quite different principles. It is unclear whether these variants differ in their utility for understanding human visual perception.

Human vision supports more than just object classification. Whereas DNNs classifiers can categorize natural images accurately, human vision yields up much richer information about its inputs (30), including other semantic information about the objects beyond its subordinate or basic category label; the parts that go together to compose it; its orientation in space; its size; how one might interact with it, etc. Such information may importantly constrain the visual similarities that people discern amongst stimuli, in ways that various current DNN image classifiers may or may not capture (31).

It is not known whether DNN representations capture the visual structure that humans perceive. While considerable research has evaluated the ability of DNNs to generalize their classification behavior, and have assessed similarity between model and neural structure, comparatively less work has assessed whether/how representations that arise in such models explain the similarities that people perceive in images (30). Where such studies have been conducted, they have focused on representation of photographic stimuli like those that constitute the model's training environment (12, 32, 33) and it is not clear whether similar results would obtain for perception of more abstract and out-of-distribution stimuli such as sketches.

These considerations raise three key questions about the degree to which DNNs provide useful scientific models of human visual object perception, which are the focus of this paper:

1. Are the internal representations/features acquired by DNNs sufficient, either alone or in combination with other common expressions of visual structure, to explain the similarities that people detect amongst abstract depictions of objects (such as line drawings)?

2. Do the internal representations/features acquired by DNNs merely recapitulate other better-understood kinds of visual features, or do they capture aspects of perceptual similarity beyond such features?

3. Do different model architectures and/or training procedures offer different answers to these questions?

To answer these questions, we adopt an approach similar to that taken by Jozwik and colleagues (34), who sought to explain the contributions of categorical and visual features, in addition to DNN features, towards explaining human-perceived similarities amongst photographs of objects. Their work evaluated two convolutional DNN architectures, AlexNet and VGG-16, across different layers. To assess human-perceived structure they had participants list visual features such as parts, colors, or shapes, and also provide category labels, such as 'elephant', 'animal', or 'natural', for their photographs. They then tested whether these human-generated features reliably predicted judgements of similarity amongst their photographs. They found that deeper layers of the DNNs outperformed visual features, but that categorical features outperformed both.

Our work builds on these results, and those of Fan and colleagues (35), by considering which features best explain and predict the similarities that humans perceived amongst line drawings. This focus extends prior work in two nontrivial ways. The first is simply that there exist a variety of computational techniques for measuring similarities between sketch images that do not rely solely on human-generated propositional descriptions of structure. Each such technique quantifies a kind of similarity between pairs of sketches, which might then provide a basis for guiding human perceptual decisions. The use of drawings allows us to investigate these metrics alongside features extracted from DNNs and human-generated labels when understanding the factors governing perceptual similarity.

Second, as noted above, drawings represent a test case for out-of-sample generalization that is important for many aspects of human visual communication. It may be that, by virtue of learning from very large sets of naturalistic images, DNNs acquire a kind of domain-general basis set for expressing visual information that then naturally capture, without specific training, perceived similarities amongst sketches and other abstract depictions of objects. If so, mechanisms embodied in DNNs are *sufficient* to explain the human ability to cope with abstract visual depictions. Alternatively, it may be that the features acquired by DNNs are insufficient to explain the structure that people discern amongst drawings without special training/tuning, or that some architectures fare better than others, or that other features beyond those expressed in DNNs provide a better or more transparent account of the central features.

In the experiments that follow, we began by estimating the similarities that people discern amongst various line drawings using a triadic comparison or *triplets* task in which participants must decide which of two sketch images is most similar to a third reference image. From many such judgments, the component sketches can be embedded within a low-dimensional space such that the Euclidean distance between pairs of points relates to the probability that the two items will be selected as "more similar" relative to some arbitrary third image (36). The resulting embeddings thus encode a low-dimensional perceptual similarity space. To determine which features govern the organization of this space, We then used regression techniques to predict the coordinates of the various drawings in the perceptual space from representational spaces derived from 5 different DNNs, from other measures of similarity, or from both together. Comparison of model fit and regression coefficients across these analyses then shed light on the three core questions raised above.

Study 1

Experiment 1 applied the general approach to understand factors governing similarities perceived amongst drawings of common real-world objects produced online by non-expert participants. While line drawings lack much of the detailed information present in photographs of objects, they nevertheless share structural isomorphisms with their real-world counterparts such as part-structure and global shape (37), and people may additionally infer from such features semantic information such as the category to which the depicted item belongs. Perceptual judgments of similarity may additionally be influenced by lower-level characteristics of the image such as the "jaggedness" of contours, the density of lines, overall size, or the orientation of the shape on the page — properties that can be quantitatively estimated via various machine-vision techniques. Experiment 1 measured the perceived similarities amongst 128 sketches depicting items from 4 different categories, then assessed how well DNN-based features and other more transparent feature sets can explain the resulting structures, either alone or in combination. Figure 1 provides a high-level overview of the workflow.

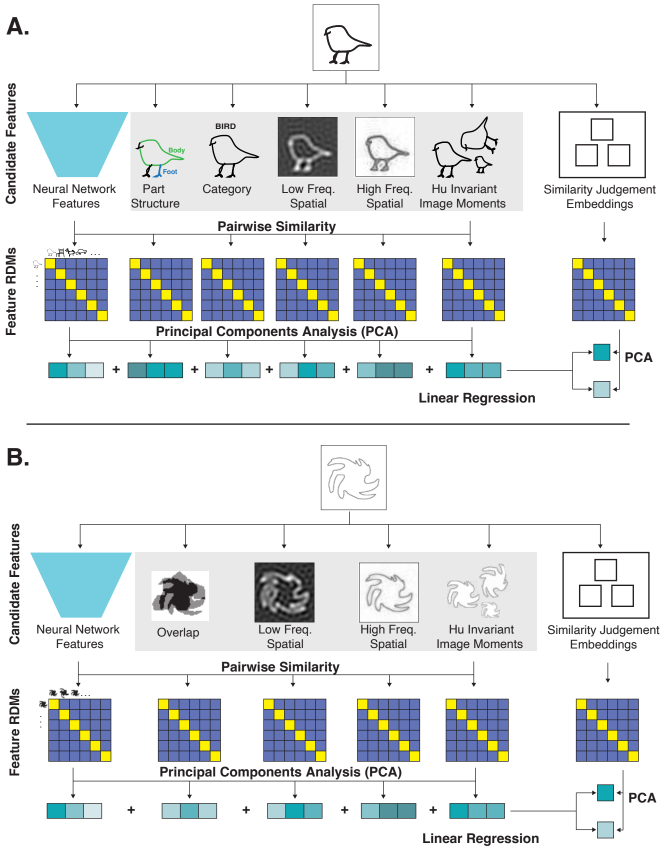

Fig. 1. Procedure for fitting linear regression to predict human judgement embeddings from candidate features. (A) In experiment 1, features were constructed using part-structure, category, low spatial frequency, high spatial frequency, and shape information. Additionally latent feature activations were extracted from 5 different neural network architectures. The features enclosed in the gray box were used for all models fit, with the neural network features varying depending on which of the 5 models were being tested. Representational dissimilarity matrices were computed from all these features and each matrix was represented using the first few principal components. These principal components computed from all the candidate features were used together in independent models to predict the first and second component of human similarity judgement embeddings. (B) In experiment 2, the process was largely the same except that part-structure and category features were no longer applicable for abstract shapes. Additionally, the degree of overlap in enclosed area between the shapes was included as a candidate feature.

Behavioral methods. *Participants.* 85 participants were recruited via Amazon Mechanical Turk (mTurk) using CloudResearch (36 Female, 47 Male, 2 other; Mean age = 38.69). Participants provided consent in accordance with the University of Wisconsin-Madison IRB and received compensation for their participation.

Stimuli. We used a subset of drawings collected by Fan and colleagues (38) for our similarity judgement study. These drawings were made in Pictionary-style *reference game* where a sketcher and a guesser were simultaneously shown the same set of 4 images. The sketcher was tasked with drawing one of the 4 images and the guesser had to guess which of the 4 images the sketcher was tasked to draw. Each image belonged to one of 4 categories — birds, dogs, cars, or chairs, and each category had 8 unique exemplars. Additionally, in some trials, the target image belonged to the same basic-level category as the 3 distractors leading to more detailed drawings by the sketcher, while on other trials all 4 images belonged to different categories leading the sketcher to make simpler drawings. We sampled 2 drawings from each condition (2) x category (4) x exemplar (8) cell resulting in a final set of 128 drawings.

Additionally, in a separate experiment, each stroke in each drawing was annotated by human-raters with a part label thus providing fine-grained information regarding the semantic part structure people observed within a given drawing (39). This information was operationalized as *part-based* vector representations for each drawing. The total number of unique parts was first computed for the entire dataset of drawings and the amount of ink and number of unique strokes for each part were then computed. These two sources of information were concatenated to create a 48 dimensional representation for each sketch, where the first 24 dimensions corresponded to the number of strokes allocated to each of the 24 unique parts and the next 24 dimensions corresponded to the amount of ink used to draw those parts.

Triplet-judgment procedure. To measure human-perceived similarity between drawings, we had participants complete a triplet similarity judgement task (36) implemented using the SALMON online tool for collecting triplet queries and fitting embeddings (https://github.com/stsievert/salmon). On each trial, participants viewed 3 drawings: a *target* positioned at the top of the screen two *options* positioned below it. They were instructed to select which of the 2 option drawings was *most similar* to the target drawing using either their mouse or the left and right arrow keys on their keyboard. If they perceived the two options to be equally similar, they were asked to pick one randomly.

We did not specify *how* participants should assess similarity when doing this task, allowing for a variety of potential strategies. Each participant completed 200 trials, including 180 sampled randomly with uniform probability from the set of all possible triplets and 20 consisting of a fixed set of 'validation' triplets that every participant saw. The validation triplet trials were randomly interleaved within the random triplet trials (Figure 2) and were used to estimate mean inter-subject agreement for the task. Based on prior work using this paradigm, participants with a mean response time less than 1500ms were excluded from any further analyses.

Computing candidate image representations. For all sketch images, we estimated low-dimensional embeddings that capture similarity structure apparent in (1) human perceptual judgments from the triplet task, (2) internal activation vectors from the deepest fully-connected layers of the five DNN models, and (3) vectors derived from alternative methods for expressing similarity structure in sketches. We refer to the vector spaces from neural networks and other techniques as *candidate image representations*, as each captures structure amongst images that may aid in predicting the perceptual similarities expressed by the triplet-based embeddings. Here we briefly describe the methods used for each candidate representation.

Similarity judgement-based embeddings. From the full set of triplet judgments, an ordinal embedding algorithm was applied to situate all 128 sketches within a low-dimensional space such that Euclidean distances amongst points minimize the crowd-kernel loss on the triplet data. (40). The optimal dimensionality was chosen by fitting embeddings in an increasing number of dimensions, evaluating each on their ability to predict human judgments in held-out validation triplet trials, and choosing the lowest-dimensional solution showing hold-out performance equal to inter-participant agreement on these trials. The results was a 2D embedding shown in Figure 3A that predicted human decisions for held-out items with accuracy of 72.70%, comparable to inter-participant agreement of 73.10% (one-sample t-test, p = .62) for the same triplets.

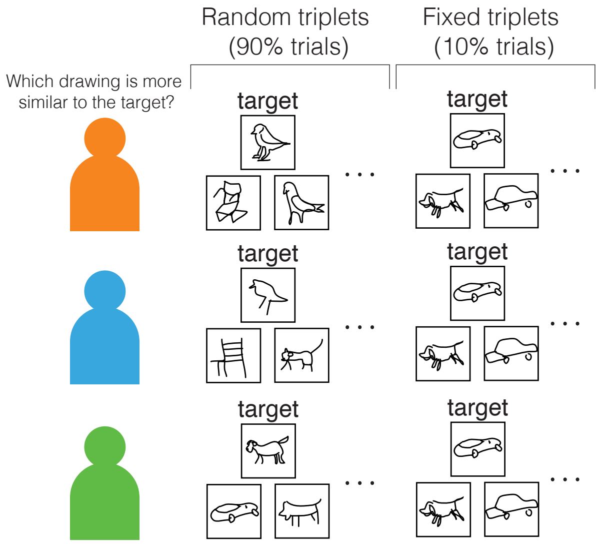
Random triplets (90% trials)
Fixed triplets (10% trials)
Which drawing is more similar to the target?
target
target
target
target
target
target
. . .

Fig. 2. Structure of the triplet similarity judgement task. Each participant (here represented with different colors) completed 200 trials, indicating which of two options was most similar to a target drawing in each trail. 180 trials sampled triplets randomly from the set of all possible triplets. The remaining 20 were 'fixed' triplets judged by all participants. Fixed and random triplets were interleaved with a different random ordering across participants.

Neural network feature activations. Neural network features were extracted using the THINGSVision Python Toolbox (41) and focusing on 5 different DNNs including (1) AlexNet, a convolutional neural network (26) that was one of the first to achieve near human-level performance at image categorization; (2) VGG-19 (42), a deeper convolutional neural network with 19 layers; (3) ResNet-18, an 18-layer convolutional image classifier that additionally employs 'residual' connections to ensure that each layer learns new structure relative to the preceding layer; (4) The Vision Transformer (ViT) (27), a (non-convolutional) Transformer-based neural network (43) trained for image classification; and (5) CLIP-ViT, a multimodal variant of the vision transformer trained on a large dataset of image-caption pairs using a contrastive loss that maximizes the similarity between valid pairs and minimizes the similarity between invalid pairs.

Models (1)-(3) utilize the well-established convolution

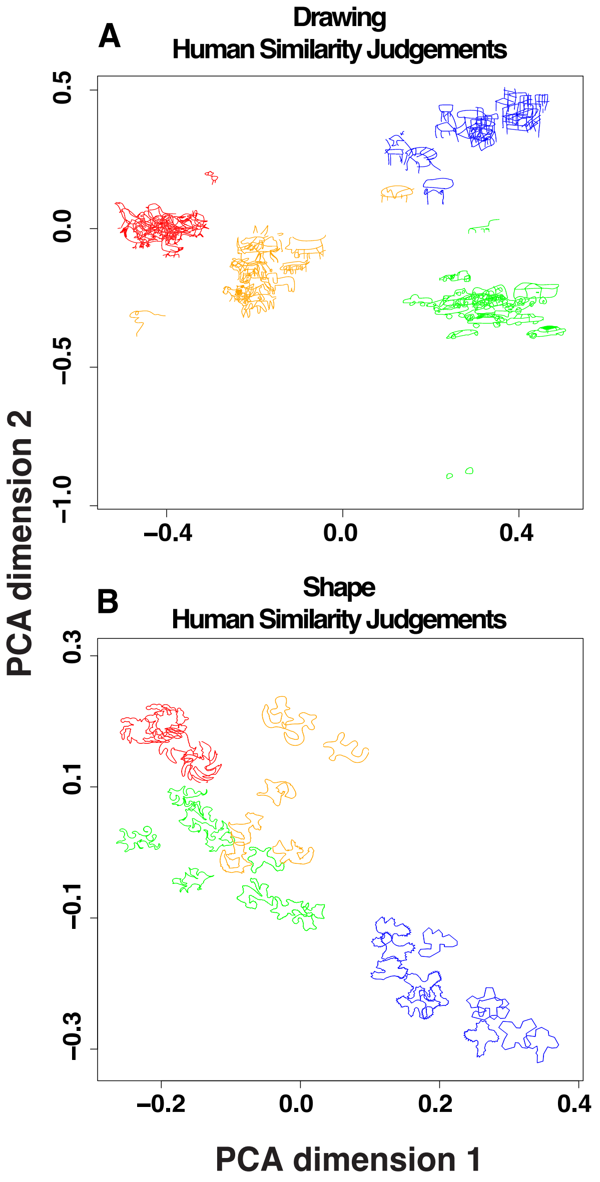

Fig. 3. Visualization of the locations of (top) line drawings and (bottom) abstract shapes along the first and second principal components computed from human similarity judgements. In the top panel, drawings of living objects are separated from non-living objects along dimension 1.

operation, where a shared set of weights is broadcast to different parts of the input tensor, enforcing an inductive bias toward spatial invariance. In convolutional models, early units with narrow receptive fields acquire simple visual feature 'filters,' which give way with greater depth to units that encode more complex features across broader receptive fields. These properties mirror some aspects of the human ventral visual stream, with some researchers suggesting they provide useful tools for understanding the primate visual system (13, 18). The three variants we studied differ in two respects. First, (2) and (3) possess many more convolutional layers (ie are deeper) than (1), an architectural difference that can lead to better overall performance and a greater level of abstraction. Second, (3) possesses 'residual' connections that allow information from earlier layers to 'skip ahead' to deeper layers so that learning in the intervening layer is driven primarily by error gradients unexplained by the preceding layer. While the effects of these architectural differences on multi-way image categorization has been well documented in the computer sciences, prior work has not considered whether they likewise affect a model's ability to capture human-like perceptual structure amongst abstract, out-of-distribution images like sketches.

Models (4) and (5) discard convolutional structure and instead utilize a *transformer* architecture (43) borrowed from the world of natural language processing. Transformers replace convolutional operations with an attention mechanism that represents each image patch as a weighted blend of representations of other patches, iteratively performs this operation until a classification of the input image's category has to be made (27). Weights governing these representations, including weights on the relevant similarity metric, are all learned via gradient descent on error. Unlike convolutional models, units in transformer models do not locally encode a spatially-bounded part of the image—instead all units can potentially encode information from all regions of the image at once. This difference allows transformers to develop remarkably flexible and context-sensitive internal representations, while performing exceedingly well on a variety of benchmark tasks in machine learning, but with little clear connection to the organization of visual processing streams in the brain. While some have addressed the relevance of the differences between convolutional and transformer vision models in modeling human vision (44), few have tested these models on abstract stimuli that nevertheless convey semantic information such as line drawings. The critical difference between (4) and (5) is not in architecture but in training objective. While (4) is trained to minimize categorization error, model (5) is trained to maximize the similarity between a visual representation of the image and 'semantic' natural-language representation of a text-description of the image while also minimizing the similarity to all other possible text-descriptions—an approach known as 'contrastive image-language pretraining' or CLIP.

To extract model internal representations, each drawing was first transformed to a standard 224x224 size. Since the drawings are grayscale and most models expect a 3D tensor, the same 224x224 image of grayscale values was copied and stacked 3 times as is standard practice. Each image tensor was applied to the model input layer and we recorded the activation vectors arising in the final hidden layer for the classification models and from the image-encoding layer for the CLIP-based model. Given the broad differences in architecture and optimization techniques, we expected to observe quantitative and qualitative differences in the structure encoded by vectors from different models. The key question was whether these structures also vary in how well they capture human perceptual representations.

Other candidate representations. Finally, for each image we also computed candidate representations using 5 alternative techniques taken from cognitive psychology and machine vision literatures. Each expresses a different kind of structure that might reasonably govern human perceptual decisions for these stimuli. They include:

Category vectors: People rapidly and automatically discern the basic-level semantic category to which sketches of common objects belong, a tendency that may influence the degree to which the sketches are perceived/judged as similar. Since each drawing in our dataset belonged to one of 4 basic-level categories (dog, bird, car or chair), we captured this information by simply representing each drawing as a four-element one-hot vector indicating to which category it belonged. If observers heavily weight the recognized category of a drawing in determining similarity over other visual properties of the image such as shape or 'style', this feature should reliably predict human similarity judgements. Note that, even though four of the five DNNs we consider were trained on image classification, it is not clear whether the representations they acquire will capture such structure, for two reasons. First, the output labels employed in this work denote classes more specific than the basic-level categories that govern non-expert visual classification in people–for instance, the classifier must assign different labels to different breeds of dog, rather than a single common label to all varieties of dog. Second, the classification models were trained only on photographs, and it is not clear whether the image features they acquire will extend to capturing basic-level category information about sketches.

Part vectors: Beyond basic-level categories, people also discern the part structure within objects (37, 45). Indeed, classic structural descriptive theories have posited that visual representations are built from the constituent parts that make up an object (46). Furthermore, people are capable of ascribing meaningful labels to the constituent parts (34). To capture the part-based knowledge that people possess, using the part annotation information in each drawing, we constructed part-based feature vectors as described in Mukherjee et al. (2019)(39). Each drawing was represented using a 48-dimensional vector containing information about (1) the number of strokes and (2) the amount of ink allocated to each of the 24 unique part labels represented in the dataset.

Hu invariant image moments: People may judge two sketches to be similar if they possess an similar overall shape, even if that shape varies in its orientation, its size and location on the page, or the viewing angle(47–49). Machine vision offers a variety of techniques for quantifying shape similarity among black-and-white line images in a size-, location-, and orientation-invariant way. Since our stimuli were 2D sketches, we adopted a technique for estimating shape-similarity in an affine-invariant (i.e., rotation-, translation-, and scale-invariant) manner. Specifically, we computed *Hu image moments* for each drawing (50) using the openCV library. Hu moments, specifically, are a set of 7 numbers that combine simpler *image moments*, which in turn represent weighted intensities of the pixel values in an image based on where on the canvas the pixel is located.

High and low spatial frequencies: Observers might be sensitive to both the overall global shape of the drawings or the local details within each drawing when assessing their similarity. To capture these qualities we computed the fast Fourier transform of each drawing and and created low and high-pass filter variants of the drawing by either setting the high or low frequencies of the drawing in the frequency-domain to 0 and reversing the transformation. This resulted in images that preferentially highlighted either global shape (low-pass) or local details (high-pass). We then flattened these image tensors and treated them as vectors. If people reliably use global shape or local details to make similarity decisions, then distances between these vector spaces should be predictive of their decisions.

Dimension reduction. Using the different representational bases outlined above, we computed representational dissimilarity matrices (RDM) by computing the pairwise distances between each of the 128 drawings. We used Euclidean distances for the similarity judgement embeddings as this is the metric that is optimized by the ordinal embedding algorithm. The remaining RDMs, save for one, encoded cosine dissimilarities between pairs of items in each vector space. The exception was the RDM for Hu image moments, which were computed using the following standard distance function D -

$$D(X,Y) = \sum_{i=0}^{6} \left| \frac{1}{H_i^X} - \frac{1}{H_i^Y} \right|$$

where X and Y are the 2 images being compared and H_i refers to the ith log-transformed Hu moment for that image.

Finally, in addition to the RDMs themselves, we computed low-dimensional embeddings of the resulting distances using singular value decomposition. Specifically, from the RDMs computed for each vector space, we extracted the first three singular vectors weighted by their respective singular values as a three-dimensional image representation approximating the distances expressed in the original high-dimensional space. These low-dimension approximations were then used in regression analyses to determine which candidate vector spaces best explain human perceived similarity. For DNN-based representations, the 3D embeddings captured 75% of the variance in the original RDM on average; we used the same dimension for reductions of other vector spaces to ensure that no single representation was over-represented in the downstream analyses.

Results of study 1. *How well do DNN-based embeddings explain human perceptual similarity?* To answer this question we first used linear regression to fit models predicting the coordinates of images along two orthogonal dimensions in the human-perception based embeddings from coordinates in each DNN-based embedding. To get the target values for regression, the 2D embedding shown in Figure 3A was subjected to a singular-value decomposition, extracting two singular vectors and weighting each by the respective singular value. This had the effect of rotating the embedding to ensure that first component aligned with the direction of greatest variation and that the second component

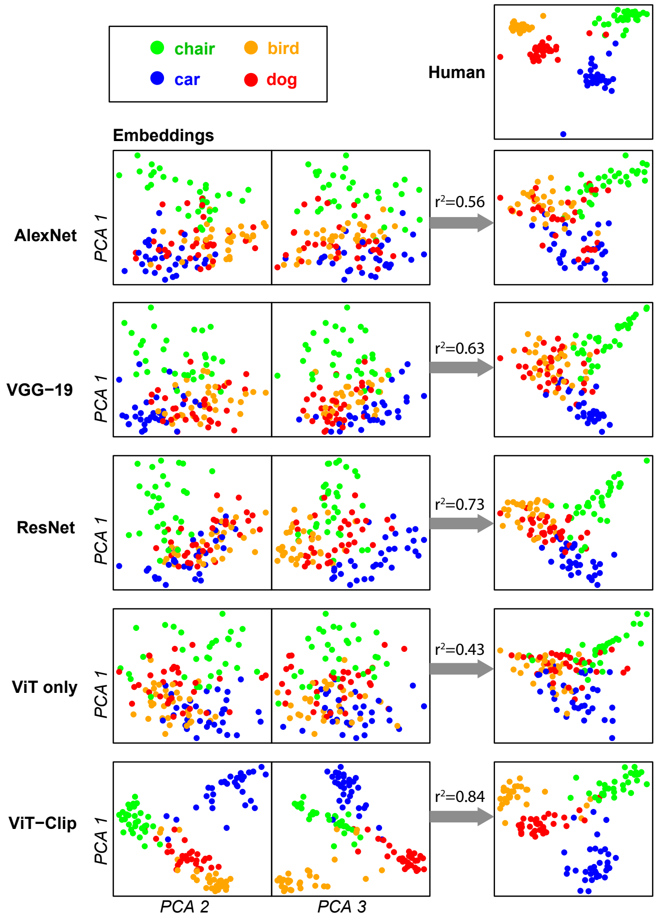

Fig. 4. Sketch embeddings in the regression analyses. The top right panel shows the human-based embeddings rotated to ensure the two components that constitute the dependent measure in the regressions are orthogonal to one another. Within each remaining row, the left plots show the 3D embeddings generated from each DNN, and the right plot shows the predicted coordinates of the sketches within the human-based space after fitting regression models.

was orthogonal to the first. We then fit separate regression models to predict each sketch's location along each of these two orthogonal dimensions from their coordinates in each 3D DNN-based embedding, including all interactions amongst the three components. The results are shown graphically in Figure 4.

The top right panel shows the human-based embeddings as rotated by the SVD technique, with colors indicating the semantic category to which each item belongs using the same scheme shown in Figure 3. The remaining rows show the 3D embedding generated from the corresponding DNN (left) and the predicted coordinates of each image in the human perceptual space after fitting the regression. The arrows indicate the proportion of variance in pairwise distances from the true human embeddings explained by the predicted embeddings. All regression fits were statistically highly reliable ($p < 0.001$ for all contrasts against null hypothesis), indicating that all architectures capture structure that is non-arbitrarily related to the similarities that people perceive. To understand how much variation in the pairwise distances from the original human-based space is explained by predicted coordinates from the regressions for each model, we took the square of the Procrustes correlation between predicted and true spaces. These are the values shown as r^2 in Figure 4. The different models varied somewhat in this metric, but the the CLIP-trained transformer model captured the most variance ($r^2 = 0.84$), reliably better than the next-best ResNet model ($p < 0.001$). By observation the reason seems clear: human-based judgments strongly cluster sketches by semantic category, and such categories are more clearly expressed in the CLIP-based model embeddings than any other model. Interestingly, the transformer architecture trained to classify images–ie, without CLIP–did not cleanly separate semantic classes, and showed the worst accuracy predicting human-based embedding coordinates.

Feature	R^2	p**-value**
category	.91	**<.001**
parts	.80	**<.001**
low freq. spatial	.22	**<.001**
high freq. spatial	.17	**<.001**
Hu moments	.16	**<.001**

Table 1. The amount of variance in human perceived similarity in drawings explained by each non-DNN candidate feature. For each feature, 2 independent regression models were fit to predict the first and second principal coordinate of the human similarity embeddings. R^2 values were computed by first computing a Procrustes correlation between the true and predicted coordinates and computing its squared value.

Predicting human similarities from other features. We next considered how well the other candidate representations fared at predicting coordinates in the human-based space, applying the same procedure but with the 3D embedding coordinates (and their interactions) from each candidate space as the predictors. Squared Procrustes correlations between predicted and true coordinates are shown for each regression in Table 1. All candidates spaces, taken individually, accounted for significant variance in the human perceptual space, but the amount of variance differed radically. The category-based vectors on their own accounted for a remarkable 91% of the variance in human-based distances–more than the best-performing DNN. Part-based vectors explained 80%, about as much as the CLIP-based transformers. The other metrics each individually explained a relatively smaller amount of variance. Like the DNN analysis, these results suggest that human judgments are dominated by information about semantic category.

Which methods account for unique variance in human-perceived similarities? Since all candidate representations independently explain some variance in human perceived similarities, a further question is whether a given candidate representation accounts for reliable variation after other representations are taken into account. To answer this question, we again fit regression models predicting human-based coordinates, but including as predictors the 3D embedding coordinates from one of the DNNs and from each of the other embeddings. We fit one such regression for each DNN type, each then including 18 different predictors (the 3 DNN components and 3 each from category, part, Hu-moment, low-frequency, and high-frequency embeddings).

Due to the large number of independent variables, we fit models using only simple effects. For each predictor, we evaluated whether its inclusion improved model accuracy more than expected under the null.

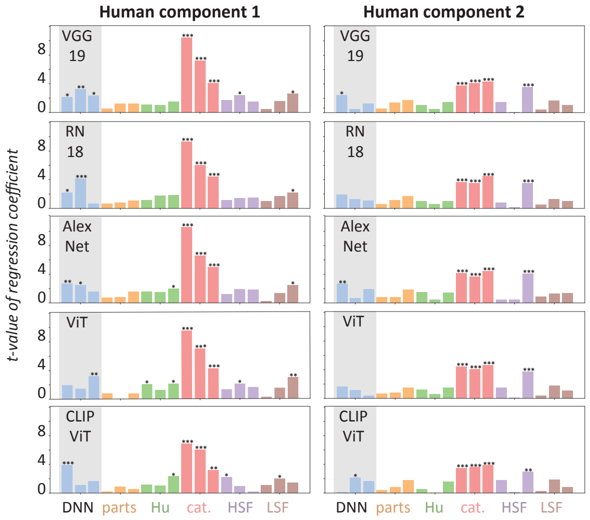

Fig. 5. Regression coefficients from Experiment 1. Rows shows t-values on regression coefficients predicting components 1 (left) or 2 (right) of the human embeddings from a combination of handcrafted features and neural network features extracted from five different architectures. Asterisks indicate coefficients that reliably improve model fit with p < 0.01*, p<0.01** or p<0.001***. DNN=Deep neural network; parts = part-based vectors; Hu = Hu moments; HSF = high spatial frequency; LSF = low spatial frequency.

Figure 5 shows t-values on regression coefficients from these analyses, with asterisks indicating which coefficients reliably reduced prediction error over and above inclusion of other predictors. For both components of the human-based embeddings, coefficients on the category-based embedding space are largest, but other spaces also received coefficients that were reliably non-zero, including embeddings from all five DNN-based representations. Thus, at least considering simple effects, DNN representations do appear to capture some elements of structure relevant to human similarity judgments over and above structure captured by category and by other, simpler metrics. How much additional structure? We compared the fits of models fit only using the non-DNN-based embeddings to those using all such features plus the DNN-based embeddings, for each architecture. Embeddings from all architectures explained significant variance over and above the other features on at least one dimension ($p < 0.05$ for all contrasts), but in all cases the amount of additional variance explained was at most 1%. Thus, while these models do appear to capture some unique aspects of human-perceived similarities, such influences appear to be relatively small.

Are these results an artifact of dimension reduction? The predictors in the preceding regressions were low-dimensional embeddings computed from very high dimensional representations. Is it possible that the various candidate representations would better explain human judgments without such reduction? To answer this question, we evaluated how well similarities encoded in the original RDMs, from both DNNs and other metrics, could predict human decisions in the triplet-judgment task. Recall that each human participant judged a set of 20 'validation' triplets, which in turn were used to measure the mean inter-subject agreement in the task and to find the best embedding dimension for characterizing human-perceived similarities. To evaluate how well the original vector spaces explain human perceptual decisions for sketches, we predicted responses on the validation triplets from each candidate space by simply looking to see, within the corresponding RDM, which of the two option sketches was least dissimilar to the target sketch in the full high-dimensional space. For each candidate representation, the predicted responses were then compared to corresponding human decisions and counted as "correct" when these matched and incorrect otherwise.

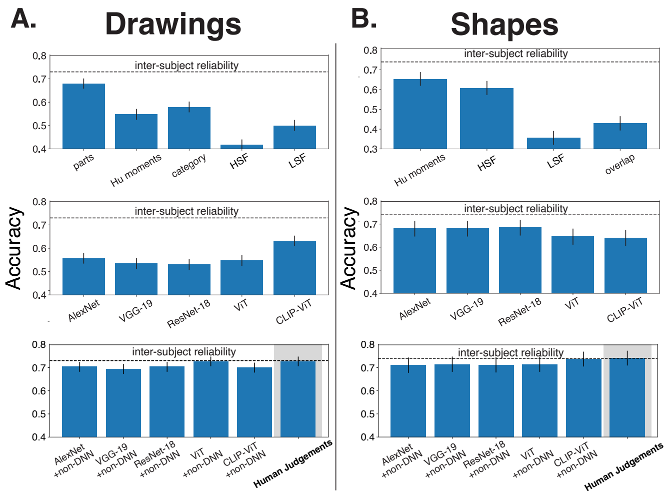

Fig. 6. Accuracy of predicted human similarity decisions for drawings (A) and shapes (B). Row 1 shows the predictive ability of psychologically motivated candidate features. Row 2 shows the predictive ability of neural network features. Row 3 shows the predictive ability of estimated human similarity judgement embeddings from DNN and non-DNN candidate features and true human similarity judgement embeddings (gray background), all models' performance was statistically comparable to inter-subject reliability. Error bars indicate 95% confidence intervals.

Figure 6A top and middle show the results. The dotted horizontal line indicates the mean inter-subject agreement, which represents an upper limit on how well any predicting model can do. While all candidate neural network representations predict human responses better than chance, no representation on its shows predictive accuracy equal to the inter-subject agreement. In other words, none of the high-dimensional representations, taken individually, fully explains the similarities that humans perceive amongst these sketches. Amongst DNNs, the CLIP-trained transformer showed better predictions than other models, consistent with the earlier regression results. Amongst non-DNN features, the part-based vectors showed highest predictive accuracy, better than the category-based vectors. Note that, while part- and category-based vectors capture similar structure, the category-based RDMs are derived from one-hot vectors, and so do not express any within-category structure, nor any broader structure across categories.

To understand how these results relate to the earlier regression analyses, we conducted a similar analysis on the *predicted coordinates* of the sketches generated from the regression models fit using embedding coordinates from one DNN plus each other candidate representation (all treated as simple effects as described above). Each regression model generated predicted coordinates for every sketch in the 2D human-based space and from these we computed the corresponding predicted Euclidean distances between all image pairs. The resulting RDM was then used to predict human decisions on the validation triplet set. The results are shown in Figure 6A (bottom) for predictions using each DNN embedding together with embeddings from other candidate representations. All models predicted human decisions at a level of accuracy similar to the inter-subject agreement. Thus low-dimension approximations of structure encoded by each DNN, when combined with comparable approximations from other spaces, are sufficient to explain human-perceived similarities amongst these stimuli.

Discussion of Study 1. Study 1 suggests that human similarity judgments for sketches of real objects are largely governed by semantic category membership: regression models built on category-based embeddings explained 91% of the variance in human-derived similarity spaces. The internal representations in DNNs capture this perceived structure to the extent they cluster images by semantic category. As shown in Figure 4, each model expresses at least some category structure, but the transformer architecture trained with CLIP shows the clearest clustering by category and accordingly yielded the best predictions of human-perceived structure amongst the different neural networks. While DNNs and other candidate representations each capture some unique variance in human perceptual structure, the amount of variance captured is much smaller than that explained by semantic category membership. These conclusions do not hinge on the low-dimensional compression of the core representations, since predictions of human decisions on the triplet task from full-dimension DNN spaces (a) were better for CLIP than other models and (b) did not fully explain human decisions. Instead, regression models that combined low-dimensional DNN embeddings with low-dimensional information from other metrics all predicted such decisions as well as possible given the level of inter-subject agreement.

These observations accord with the prior results of several studies (34, 51, 52), whose studies of perceived similarities amongst photographs of objects likewise found that such structure is dominated by semantic category membership. The current work shows a similar pattern even for abstract, out-of-distribution stimuli like sketches, and including a range of alternative representational structures beyond propositional features listed by people. Perhaps more interestingly, the results show that the CLIP training procedure, which constrains learning by ensuring that images and natural language descriptions receive similar representations when they denote similar context, leads to much clearer emergence of semantic category structure, even for abstract sketch images.

The contrasting behavior of the vision transformers with/without CLIP training is interesting because it suggests that the good performance of the CLIP-trained model does not arise from the transformer architecture *per se*. Indeed, the transformer trained on classification–the same task used with the convolutional models–showed worse ability to explain human-perceived similarities. Since CLIP training encourages the model to represent images and their natural-language descriptions as similar, it may be that this constraint leads to improved ability to capture semantic similarity structure in sketch images. This possibility is only tentative, however, since there are many other differences between the two models, most notably the corpora on which they were trained.

A remaining question concerns the degree to which our behavioral results reflect, not the representation of perceptual structure within visual processing systems, but the human tendency to rely instead on rich semantic knowledge about the items depicted. It may be that semantic category membership dominates the similarity space simply because, once participants recognize a sketch as a member of a familiar class, they retrieve names and a range of other familiar properties common to the category, and base their similarity judgments on these inferred semantic characteristics rather than on visual similarity alone. If so, the preceding results may not shed much light on the degree to which DNNs and other metrics capture visual structure independent of this semantic information. Experiment 2 tests this possibility.

Experiment 2

Experiment 2 followed the same design as Experiment 1, but instead using line drawings depicting abstract, unrecognizable shapes. If human similarity judgments for object sketches are largely driven by semantic features retrieved when the stimulus is recognized, we might expect quite different results for such stimuli. The items we chose were a set of 64 line drawings devised by Schmidt and Fleming (53), which show bounded but visually complex shapes that are not recognizable as real-world objects (see Figure 3B). The shapes were designed to fall into both broader and finer-grained groups on the basis of their visual similarity alone, and so provide a useful contrast case for the results in Experiment 1.

Methods for study 2. *Participants.* 40 participants were recruited via Amazon Mechanical Turk (mTurk) using CloudResearch (14 Female, 26 Male; Mean age = 36.25). Participants provided consent in accordance with the University of Wisconsin-Madison IRB and received compensation for their participation.

Stimuli. The dataset consisted of 64 unique shapes, each derived from one of 4 base shapes (53). Within a family of base shapes, each exemplar varied in low-level perceptual properties such as whether the contours were smooth, angular, or corrugated. Thus, the dataset had systematic perceptual regularities in addition to within-family variation. To standardize the images, each shape was extracted, made into a grayscale contour, and positioned in the center of a 525x525 pixel canvas.

Procedure for triplet judgement task. The task was identical to that described in Study 1, but using the shape stimuli in place of sketches. Participants with a mean response time of over 1500ms were again excluded from further analyses. The same algorithm was used to situate the 64 items in a 2D Euclidean space to minimize the crowd-kernel loss on the triplet judgment dataset. The resultant embeddings, shown in Figure 3B, predicted human judgments on a held-out validation set with 73.76% accuracy.

Candidate representations. Study 2 used the same techniques as Study 1 to derive RDMs and corresponding 3D embeddings for the 64 items from each DNN and from the additional candidate representational similarity spaces, with two exceptions. First, since the stimuli do not correspond to familiar categories of items and do not possess familiar, identifiable parts, we did not include category- or part-based vectors. Second, since each image is a bounded figure typically perceived as an object situated against a background, we included one additional measure of visual similarity, namely shape overlap. For this metric, we filled the area within the contour for each shape with a value of 1 and the area outside the contour with a value of 0, then computed overlap as:

$$O(X,Y) = \frac{\sum(X\&Y)}{\sum(X\,|\,Y)}$$

...where X and Y are flattened binary bitmaps of the 2 images being compared. Thus the candidate representations in this dataset included RDMs and associated 3D embeddings for the 5 DNNs and for Hu moments, low-frequency reconstructions, high-frequency reconstructions, and shape overlap. The central questions was whether and how these different spaces could explain human-perceived similarities amongst these unfamiliar, non-meaningful shape drawings.

Results of Study 2. To assess how well the various DNN representations explain human-perceived similarities, we again conducted regression analyses predicting coordinates in the human similarity space from the 3D embedding coordinates derived from each model, including all interaction terms. The results are shown in Figure 7. The human-derived embeddings (top right) clearly capture the 'family' groupings intended by the designers (dot colors), an organization reflected to varying degrees across the embeddings from different models. Regressions predicting human-based coordinates from the embeddings all account for significant variange ($p < 0.001$ for all contrasts to null), with the CLIP-trained transformer again accounting for the model (79%) and the VGG-19 model coming a near second (76%). As with study 1, the transformer architecture trained without the CLIP loss was the worst-performing model, accounting for 64% of variation in human-perceived similarities. Regressions predicting human-based coordinates from the alternative spaces all accounted for significant variance ($p < 0.001$ vs. the null) but did not fare as well as the DNN embeddings, with Hu moments accounting for the most variance (63%), followed by low-spatial-frequency embeddings (48%), shape overlap (41%), and high-spatial-frequency (34%).

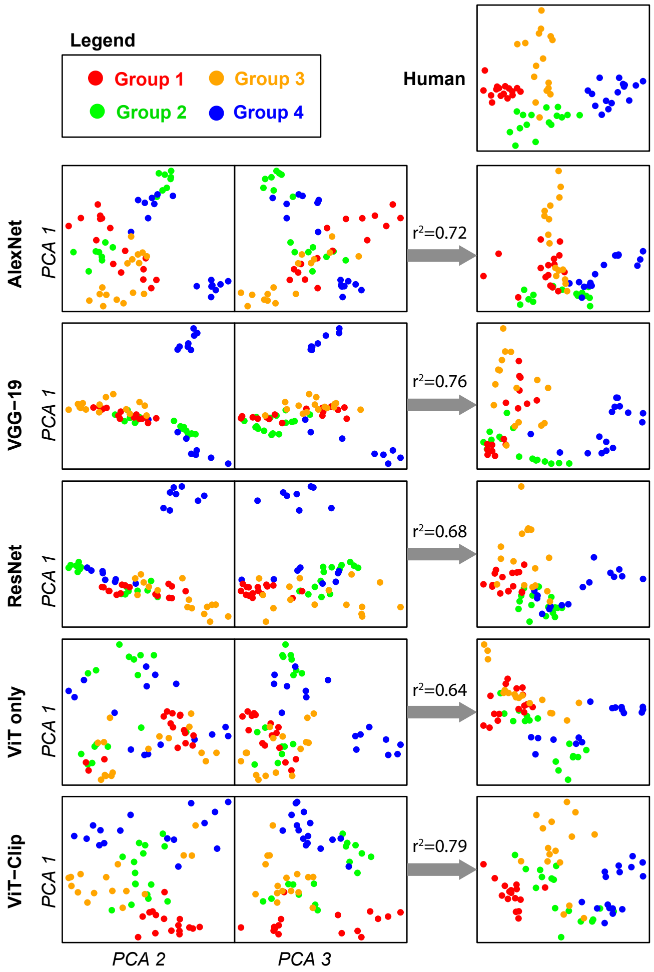

Fig. 7. Shape embeddings in the regression analyses. The top right panel shows the human-based embeddings rotated to ensure the two components that constitute the dependent measure in the regressions are orthogonal to one another. Within each remaining row, the left plots show the 3D embeddings generated from each DNN, and the right plot shows the predicted coordinates of the sketches within the human-based space after fitting regression models.

Table 2 shows the corresponding fit values for regressions using each alternative embedding space as the predictor. While each alternative again accounted for significant variance in the target space ($p < 0.001$ vs. the null), no alternative space accounted for as much variance as the better-performing DNNs. Hu moments on their own explained 63% of the variance in the human-derived space, about the same as the worst-performing DNN.

To determine whether the various representation spaces capture unique aspects of human-perceived structure, we again combined coordinates from each DNN embedding with those from alternative candidate representations, investigating only simple effects. These results are shown in Figure 8. While all metrics account for significant

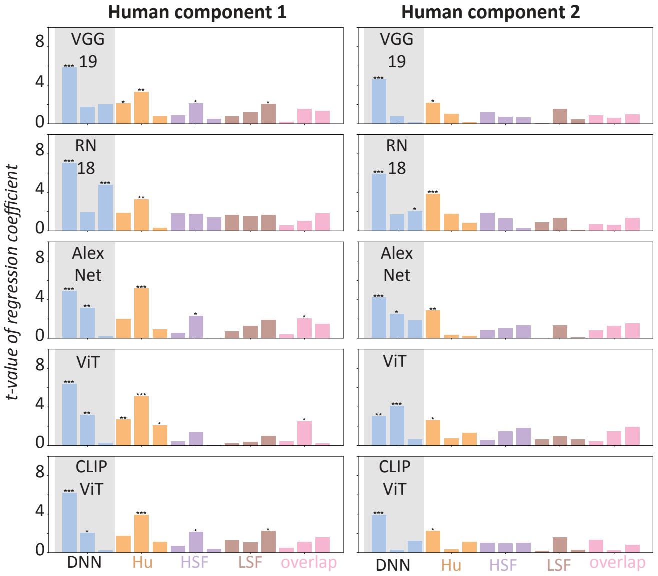

Fig. 8. Regression coefficients from Experiment 2. Rows shows t-values on regression coefficients predicting components 1 (left) or 2 (right) of the human embeddings from a combination of handcrafted features and neural network features extracted from five different architectures. Asterisks indicate coefficients that reliably improve model fit with p < 0.01*, p<0.01** or p<0.001***. DNN=Deep neural network; Hu = Hu moments; HSF = high spatial frequency; LSF = low spatial frequency; Overlap=degree of shape overlap.

unique variance on at least one target dimension, DNN embeddings attracted the largest coefficients in the regression model, followed by the shape-similarity measure captured by Hu moments. Table 3 shows the change in r^2 observed when contrasting models fit with/without the DNN-based embeddings. All five explained significant additional variance beyond Hu moments and other spaces. The amount of unique variance explained by each was an order of magnitude larger than observed in Study 1, ranging from 11% to 29% across the two dimensions. In this analysis, ResNet-18 and the CLIP-trained transformer each accounted for the most additional variance.

Feature	R^2	p**-value**
Hu moments	.63	**<.001**
high freq. spatial	.34	**<.001**
low freq. spatial	.48	**<.001**
overlap	.41	**<.001**

Table 2. The amount of variance in human perceived similarity in abstract shapes explained by each non DNN candidate feature. For each feature, 2 independent regression models were fit to predict the first and second principal coordinate of the human similarity embeddings. R^2 values once again correspond to the squared Procrustes correlation.

Feature	ΔR^2	F**-statistic**	p**-value**
human judgements component 1			
VGG-19	.11	17.84	**<.001**
ResNet-18	.16	35.55	**<.001**
AlexNet	.11	18.25	**<.001**
ViT	.12	21.29	**<.001**
CLIP-ViT	.17	45.21	**<.001**
human judgements component 2			
VGG-19	.16	10.14	**<.001**
ResNet-18	.29	19.42	**<.001**
AlexNet	.22	13.71	**<.001**
ViT	.19	10.51	**<.001**
CLIP-ViT	.28	20.77	**<.001**

Table 3. The amount of unique variance explained by DNN features in ensemble models with all other candidate features. Unlike in the case of drawings, DNN features explain a larger part of the variance.

Finally, to assess whether these results reflect the dimension reduction step, we again used the original RDMs for each vector space to predict human judgments on the validation items from the triplet task. As with Study 1, we evaluated the predictions from each representational space considered independently, and also from the 2D space generated by predictions of the regressions models that combine DNN and other feature embeddings. The results are shown in Figure 6B. Relative to the results with sketches, the DNN feature spaces alone show higher accuracy predicting human judgments for these non-meaningful stimuli, though they do not reach the ceiling level defined by inter-subject agreement. Interestingly, without data reduction and parameter fitting via regression, the CLIP-trained transformer performs worst among the DNNs, suggesting that the very high dimension native space may encode much information irrelevant to human perception.

Predictions from Hu moments perform as well as the worst-performing DNN embeddings, suggesting that human judgments are, unsurprisingly, largely driven by overall similarity in shape for these stimuli. Embeddings computed from high-frequency spatial information also do relatively well. Note that regressions based on embeddings of the high-spatial-frequency vectors explained the least variance in the human-based embeddings. The contrasting pattern suggests that these vectors contain information relevant to human judgments that is lost by the compression to three dimensions. For instance, for these stimuli such judgments may be partly informed by patterns in high spatial frequencies such as the rounded, jagged, or square contours that form each shape.

Discussion of Study 2. Study 1 suggested that, for drawings of recognizable objects, semantic information dominates human similarity judgments, and DNN representations capture little additional structure. Study 2 suggests that, when object category (and other semantic information) is not available to inform similarity judgments about drawings, DNNs can capture nontrivial aspects of human-perceived similarity not expressed by the other metrics we considered. While human perceptual judgments for these items seem strongly informed by shape similarity, all DNN representations accounted for significant additional variation beyond Hu moments, the overlap metric, and spaces derived from high and low spatial frequency information. Moreover, regression analyses placed the largest coefficients on DNN-based predictors, which reliably improved predictive accuracy over and above all other feature types.

The best-performing DNN-based embeddings were again those computed from the CLIP-trained transformer model, while the worst-performing were again those computed from the classification-trained transformer. This pattern echoes the results of Study 1, with interesting implications. As noted earlier, CLIP encourages networks to assign similar internal representations to images and their natural-language descriptions. When the sketch images depict real, recognizable objects, it seems reasonable to suppose that such training promotes the discovery of semantic-category-like internal representations for these items, since such structure will be expressed in the natural-language descriptions of images. In Study 2, the stimuli do not correspond recognizable items; no such items have appeared in the model training environment; and no natural-language descriptions exist to aid in organizing their structure. Nevertheless the CLIP-trained transformer performed markedly better than the classification-trained transformer, and the shape-similarity-based families built by design into the stimuli are clearly captured by the CLIP-based internal representations. This suggests that CLIP training may aid in more than just capturing semantic similarities amongst familiar visual stimuli–perhaps such learning allows the system to find a representational basis that more accurately captures human perceptual similarity even for completely novel shape stimuli. That is, perhaps the features that support perception of visual similarity for novel objects are precisely those that best promote representation of semantic structure from vision for familiar objects.

General Discussion

From early in life and without special training, human beings, perhaps alone among animals, can recognize abstract depictions of objects in the world. Theories of human vision are challenged to explain such abilities: what computational or information-processing mechanisms do human minds possess that support such abstraction? This paper considered whether contemporary deep neural network models, independently or together with other representational spaces, provide an answer to this question. While prior work has investigated how deep neural network features might contribute towards explaining patterns of human similarity judgements (34, 51), these studies were conducted in the domain of real-world photographs and efforts that have looked at the performance of deep neural networks on simple silhouettes (31, 54) or simple drawings (55) haven't contrasted DNNs to simpler feature spaces, or tested a suite of models with sufficient variance in architecture and training methods. For both sketches of real objects and line drawings depicting unrecognizable shapes, we used human behavior in a triplet-judgement task to map a low-dimensional space capturing perceived similarities amongst stimuli. We then assessed whether internal representations extracted from various DNNs can explain the resulting structure. Broadly, our results suggest that the utility of DNN-based representations for understanding human similarity judgments hinges on whether the stimuli depict recognizable objects.

For sketches of real items, we found that human similarity judgments were overwhelmingly driven by the depicted item's basic-level semantic category. Vector-space representations based only on basic-level category explained 91% of the variance in inter-item distances from the human embedding space. While features extracted from each DNN architecture did account for statistically significant additional variance beyond category and other candidate feature spaces, the amount of additional variance was 1% or less. Moreover, the DNN-based representations that independently explained the most variance in human-perceived similarity were those that most cleanly separated stimuli by semantic category. Taken together these observations suggest that structure encoded by DNNs does not add greatly to an understanding of human similarity judgments for drawings of real objects, since such judgments mainly express semantic category structure.

For drawings of unrecognizable shapes, however, DNNs capture important information not expressed by the other metrics we considered. Unsurprisingly, human judgments are partly driven by overall similarity in shape, a property captured by Hu moments. Yet after regressing out this structure and other purely visual measures (including shape overlap and similarity in low- and high-spatial-frequency information), DNN-based representations still explained an additional 11-29% of variance amongst inter-item distances in the human-derived similarity space. Considered independently, the best-performing DNN accounted for 79% of the variance in such distances, substantially more than the best-performing non-DNN-based representations (Hu moments, accounting independently for 63% of variance). That is, in contrast to results with sketches of real objects, no alternative representation fared better at predicting human-perceived similarity than did the best-performing DNN (the CLIP-trained transformer). Thus, when no semantic information about the stimulus is available to guide judgments, DNN-based representations generally appear to better capture human-perceived structure than do other measures.

With these observations in mind, we can revisit the three questions raised in the introduction and the answers our results suggest.

1. Are the internal representations/features acquired by DNNs sufficient, either alone or in combination with other common expressions of visual structure, to explain the similarities that people detect amongst abstract depictions of objects (such as line drawings)? For neither dataset did DNN-based representations alone capture all of the information needed to model human similarity judgments. When low-dimensional embeddings of DNN-based structure were used to predict human-based embeddings, the best-performing networks captured a remarkable amount of variance for both sketches (84%) and shapes (79%). Without compression and regression, raw distances in DNN representational spaces did not fully predict human decisions on the triplet task. Only when low-dimension DNN embeddings were combined with other non-DNN-based features in a regression model was it possible to predict human decisions on triplet judgments at ceiling level for both datasets.

2. Do the internal representations/features acquired by DNNs merely recapitulate other better-understood kinds of visual features, or do they capture aspects of perceptual similarity beyond such features? For both datasets, DNN-based representations accounted for significant additional variance when predicting coordinates in the human-derived similarity space. The amount of additional variance explained, however, was quite small for sketches and substantially larger for novel shapes. For sketches, simply knowing the category to which an item belongs carries a great deal of information about the similarity decisions people will make, and it is not clear that DNN-based representations capture much useful information beyond category, especially given their opacity. In contrast, for novel shapes, no alternative representational basis explained as much variation in human decisions as did the best-performing DNN, and all DNNs explained non-trivial additional variance in the human-derived distances. Thus, when semantics is removed from the table, DNN-based features express aspects of human perceptual structure difficult to capture in simpler techniques.

3. Do different model architectures and/or training procedures offer different answers to these questions? Our results suggest that the training task may matter more than the model architecture. For both sketches and shapes, the best performing model was the CLIP-trained transformer, while the worst-performing model was the classification-trained transformer. Convolutional models, all trained only on classification, fell somewhere between these poles. The contrast is instructive as it suggests that good performance is not attributable to the transformer architecture alone. Instead the CLIP training procedure, which promotes affinity in representation between images and their verbal descriptions, promotes representations of sketches that better capture semantic category structure (and so better explain human similarity decisions) *and* representations of novel shapes that better express human-perceived similarities amongst these.

Broader implications. The CLIP training procedure promotes acquisition of common latent representations that jointly support visual and natural language representations. In this respect it resonates with a well-known perspective on semantic representation in the mind and brain, namely the *hub-and-spokes* approach(56–59). The hub and spokes model proposes that different receptive and expressive information channels in the brain–vision, language, action, hearing, etc–communicate with one another via a shared representational "hub", which serves to mediate interactions amongst the various modality-specific "spokes." In so doing, it acquires distributed representations that are shaped by patterns of high-order co-variation across modalities and over time(60, 61), which in turn express conceptual or semantic similarity relations. CLIP-trained transformers capture this idea for vision and language by enforcing a learning constraint so that images and language with similar semantic content receive similar internal representations. That is, the distributed image representations acquired in a CLIP-trained transformer reflect information from both vision and language, and in this sense are analogous to the representations arising in the proposed cross-modal semantic hub.

The concordance is interesting for two reasons. First, the hub-and-spokes model has proven useful for understanding a range of phenomena in the cognitive neuroscience of semantic memory, including patterns of semantic dysfunction from brain injury (56, 57, 62), the large-scale connectivity of the cortical semantic network (61, 63, 64), functional imaging of neuro-semantic processing (64–66), and results of transcranial magnetic stimulation (67). The observation that a loosely parallel technique for training DNNs likewise improves both semantic representation and agreement with human similarity judgements provides an avenue for connecting contemporary machine-learning models to a scientific framework useful for understanding this important aspect of human cognition.

Second, the CLIP training procedure yielded better agreement with human-derived similarities even for novel object shapes that do not denote recognizable entities. That is, encouraging agreement between vision and language representations of real stimuli promoted acquisition of visual features that better capture human perception generally. This suggests that the visual features governing human perceptual similarity may be precisely those that best aid, not image classification, but distributed representations of semantic/conceptual structure. The optimal visual basis for generating distributed semantic representations may differ significantly from the basis optimal for specific item classification—in which case, DNNs trained only on classification may provide a poor approximation of the computations carried out in human vision.

In this work we focused on line drawings, both because they serve as a class of stimuli beyond the standard repertoire of deep image-classifier training datasets and because it is possible to compute low-level image features and annotate part-structure more easily in them relative to real-world photographs. While our simple approaches suffice for characterizing visual and perceived semantic structure in sketches and simple shapes, recent advances in the automatic computation of robust shape dimensions from generative adversarial networks trained to generate realistic silhouettes of objects (68) provide a promising avenue to extend our approaches to the domain of naturalistic images. Coupled with novel methods for image-computable part-structure (69), future work can not only apply our methods to a broader range of stimuli but also evaluate the performance of DNNs trained to specifically understand finer-grained semantic information, such as parts and scene-segmentations, in both photographs (70, 71) and drawings (72).

Acknowledgements

We thank Joe Austerweil, Emily Ward, and Karen Schloss for early comments on this project. We thank Gary Lupyan for feedback and referring us to the shapes dataset used in experiment 2. We thank Robert Nowak for valuable feedback on shape-matching metrics. Finally, we thank the MTurk workers who created the drawings in our dataset, annotated strokes with the part-labels, and provided similarity judgements.

Neither of the experiments reported in this manuscript were preregistered.
All code and materials will be accessible at:
https://github.com/kushinm/sketch_triplets

References

1. Judy S DeLoache, Mark S Strauss, and Jane Maynard. Picture perception in infancy. *Infant behavior and development*, 2:77–89, 1979.
2. Megumi Kobayashi, Ryusuke Kakigi, So Kanazawa, and Masami K Yamaguchi. Infants' recognition of their mothers' faces in facial drawings. *Developmental Psychobiology*, 62(8): 1011–1020, 2020.
3. Julian Hochberg and Virginia Brooks. Pictorial recognition as an unlearned ability: A study of one child's performance. *the american Journal of Psychology*, 75(4):624–628, 1962.
4. Maureen V Cox. *Children's drawings of the human figure.* Psychology Press, 2013.
5. Masayuki Tanaka. Recognition of pictorial representations by chimpanzees (pan troglodytes). *Animal cognition*, 10(2):169–179, 2007.
6. Patrick A Cabe. Transfer of discrimination from solid objects to pictures by pigeons: A test of theoretical models of pictorial perception. *Perception & Psychophysics*, 19(6):545–550, 1976.
7. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
8. Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.
9. Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017.
10. Yael Vinker, Ehsan Pajouheshgar, Jessica Y Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. Clipasso: Semantically-aware object sketching. *arXiv preprint arXiv:2202.05822*, 2022.
11. Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018.
12. Nikolaus Kriegeskorte. Deep neural networks: a new framework for modelling biological vision and brain information processing. *biorxiv*, page 029876, 2015.
13. Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.
14. Aran Nayebi, Daniel Bear, Jonas Kubilius, Kohitij Kar, Surya Ganguli, David Sussillo, James J DiCarlo, and Daniel LK Yamins. Task-driven convolutional recurrent models of the visual system. *arXiv preprint arXiv:1807.00053*, 2018.
15. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
16. Nicholas J Sexton and Bradley C Love. Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science advances*, 8(28):eabm2219, 2022.
17. Katherine R Storrs, Tim C Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus Kriegeskorte. Diverse deep neural networks all predict human it well, after training and fitting. *BioRxiv*, 2020.
18. Charles F Cadieu, Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*, 10(12): e1003963, 2014.
19. Judith E Fan, Daniel LK Yamins, and Nicholas B Turk-Browne. Common object representations for visual production and recognition. *Cognitive science*, 42(8):2670–2698, 2018.
20. Justin Yang and Judith E Fan. Visual communication of object concepts at different levels of abstraction. *arXiv preprint arXiv:2106.02775*, 2021.
21. Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
22. Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.
23. Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.
24. Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33:19000–19015, 2020.
25. Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):119, 2016.
26. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.
27. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
28. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.
29. Colin Conwell, Jacob S Prince, George A Alvarez, and Talia Konkle. What can 5.17 billion regression fits tell us about artificial models of the human visual system? In *SVRHM 2021 Workshop@ NeurIPS*, 2021.
30. Jeffrey S Bowers, Gaurav Malhotra, Marin Dujmović, Milton Llera Montero, Christian Tsvetkov, Valerio Biscione, Guillermo Puebla, Federico G Adolfi, John Hummel, Rachel Flood Heaton, et al. Deep problems with neural network models of human vision. 2022.
31. Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, 14(12): e1006613, 2018.
32. Joshua C Peterson, Joshua T Abbott, and Thomas L Griffiths. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, 42(8):2648–2669, 2018.
33. Brenden M Lake, Wojciech Zaremba, Rob Fergus, and Todd M Gureckis. Deep neural networks predict category typicality ratings for images. In *CogSci*, 2015.
34. Kamila M Jozwik, Nikolaus Kriegeskorte, Katherine R Storrs, and Marieke Mur. Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in psychology*, 8:1726, 2017.
35. Judith Fan, Daniel Yamins, and Nicholas Turk-Browne. Common object representations for visual production and recognition. *Cognitive Science*, 2018.
36. Kevin G Jamieson, Lalit Jain, Chris Fernandez, Nicholas J Glattard, and Robert D Nowak. Next: A system for real-world development, evaluation, and application of active learning. In *NIPS*, pages 2656–2664. Citeseer, 2015.
37. Barbara Tversky. Parts, partonomies, and taxonomies. *Developmental Psychology*, 25(6): 983, 1989.
38. Judith E Fan, Robert D Hawkins, Mike Wu, and Noah D Goodman. Pragmatic inference and visual abstraction enable contextual flexibility during visual communication. *Computational Brain & Behavior*, 3(1):86–101, 2020.
39. Kushin Mukherjee, Robert D Hawkins, and Judith E Fan. Communicating semantic part information in drawings. 2019.
40. Omer Tamuz, Ce Liu, Serge Belongie, Ohad Shamir, and Adam Tauman Kalai. Adaptively learning the crowd kernel. *arXiv preprint arXiv:1105.1033*, 2011.
41. Lukas Muttenthaler and Martin N. Hebart. Thingsvision: A python toolbox for streamlining the extraction of activations from deep neural networks. *Frontiers in Neuroinformatics*, 15: 45, 2021. ISSN 1662-5196. doi: 10.3389/fninf.2021.679838.
42. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
43. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
44. Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L Griffiths. Are convolutional neural networks or transformers more like human vision? *arXiv preprint arXiv:2105.07197*, 2021.
45. David Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9(3):353–383, 1977.
46. Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological Review*, 94(2):115, 1987.
47. MC Booth and Edmund T Rolls. View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral cortex (New York, NY: 1991)*, 8 (6):510–523, 1998.
48. Hamid Karimi-Rouzbahani, Nasour Bagheri, and Reza Ebrahimpour. Invariant object recognition is a personalized selection of invariant features in humans, not simply explained by hierarchical feed-forward vision models. *Scientific reports*, 7(1):1–24, 2017.
49. Hamid Karimi-Rouzbahani, Nasour Bagheri, and Reza Ebrahimpour. Hard-wired feed-forward visual mechanisms of the brain compensate for affine variations in object recognition. *Neuroscience*, 349:48–63, 2017.
50. Zhihu Huang and Jinsong Leng. Analysis of hu's moment invariants on image scaling and rotation. In *2010 2nd international conference on computer engineering and technology*, volume 7, pages V7–476. IEEE, 2010.
51. Joshua C Peterson, Joshua T Abbott, and Thomas L Griffiths. Adapting deep network features to capture psychological representations. *arXiv preprint arXiv:1608.02164*, 2016.
52. Marieke Mur, Mirjam Meys, Jerzy Bodurka, Rainer Goebel, Peter A Bandettini, and Nikolaus Kriegeskorte. Human object-similarity judgments reflect and transcend the primate-it object

representation. *Frontiers in psychology*, 4:128, 2013.

53. Filipp Schmidt and Roland W Fleming. Visual perception of complex shape-transforming processes. *Cognitive psychology*, 90:48–70, 2016.
54. Jonas Kubilius, Stefania Bracci, and Hans P Op de Beeck. Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, 12(4): e1004896, 2016.
55. Johannes JD Singer, Katja Seeliger, Tim C Kietzmann, and Martin N Hebart. From photos to sketches-how humans and deep neural networks process objects across different levels of visual abstraction. *Journal of vision*, 22(2):4–4, 2022.
56. James L McClelland and Timothy T Rogers. The parallel distributed processing approach to semantic cognition. *Nature reviews neuroscience*, 4(4):310–322, 2003.
57. Timothy T Rogers, James L McClelland, et al. *Semantic cognition: A parallel distributed processing approach*. MIT press, 2004.
58. Timothy T Rogers, Matthew A Lambon Ralph, Peter Garrard, Sasha Bozeat, James L McClelland, John R Hodges, and Karalyn Patterson. Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychological review*, 111 (1):205, 2004.
59. Karalyn Patterson, Peter J Nestor, and Timothy T Rogers. Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature reviews neuroscience*, 8(12):976–987, 2007.
60. Matthew A Lambon Ralph, Elizabeth Jefferies, Karalyn Patterson, and Timothy T Rogers. The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, 18(1):42–55, 2017.
61. Rebecca L Jackson, Timothy T Rogers, and Matthew A Lambon Ralph. Reverse-engineering the cortical architecture for controlled semantic cognition. *Nature human behaviour*, 5(6):774–786, 2021.
62. Matthew A Lambon Ralph, Christine Lowe, and Timothy T Rogers. Neural basis of category-specific semantic deficits for living things: evidence from semantic dementia, hsve and a neural network model. *Brain*, 130(4):1127–1137, 2007.
63. Richard J Binney, Geoffrey JM Parker, and Matthew A Lambon Ralph. Convergent connectivity and graded specialization in the rostral human temporal lobe as revealed by diffusion-weighted imaging probabilistic tractography. *Journal of cognitive neuroscience*, 24 (10):1998–2014, 2012.
64. Lang Chen, Matthew A Lambon Ralph, and Timothy T Rogers. A unified model of human semantic knowledge and its disorders. *Nature human behaviour*, 1(3):0039, 2017.
65. Timothy T Rogers, Julia Hocking, UTA Noppeney, Andrea Mechelli, Maria Luisa Gorno-Tempini, Karalyn Patterson, and Cathy J Price. Anterior temporal cortex and semantic memory: reconciling findings from neuropsychology and functional imaging. *Cognitive, Affective, & Behavioral Neuroscience*, 6(3):201–213, 2006.
66. Timothy T Rogers, Christopher R Cox, Qihong Lu, Akihiro Shimotake, Takayuki Kikuchi, Takeharu Kunieda, Susumu Miyamoto, Ryosuke Takahashi, Akio Ikeda, Riki Matsumoto, et al. Evidence for a deep, distributed and dynamic code for animacy in human ventral anterior temporal cortex. *Elife*, 10:e66276, 2021.
67. Gorana Pobric, Elizabeth Jefferies, and Matthew A Lambon Ralph. Category-specific versus category-general semantic impairment induced by transcranial magnetic stimulation. *Current biology*, 20(10):964–968, 2010.
68. Yaniv Morgenstern, Frieder Hartmann, Filipp Schmidt, Henning Tiedemann, Eugen Prokott, Guido Maiello, and Roland W Fleming. An image-computable model of human visual shape similarity. *PLoS computational biology*, 17(6):e1008981, 2021.
69. Henning Tiedemann, Filipp Schmidt, and Roland W Fleming. Superordinate categorization based on the perceptual organization of parts. *Brain Sciences*, 12(5):667, 2022.
70. Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
71. Feng Li, Hao Zhang, Shilong Liu, Lei Zhang, Lionel M Ni, Heung-Yeung Shum, et al. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. *arXiv preprint arXiv:2206.02777*, 2022.
72. Lei Li, Hongbo Fu, and Chiew-Lan Tai. Fast sketch segmentation and labeling with deep learning. *IEEE computer graphics and applications*, 2018.