Memory & Cognition
https://doi.org/10.3758/s13421-024-01580-1

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

Kushin Mukherjee[1] · **Timothy T. Rogers**[1]

Accepted: 22 April 2024

Abstract

Early in life and without special training, human beings discern resemblance between abstract visual stimuli, such as 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 five contemporary DNNs, we evaluated how well each explains human similarity judgments among line drawings of recognizable and novel objects. For object sketches, human judgments 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.

Keywords Drawing · Deep Learning · Perception · Similarity Judgements · Multi-modal models

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 perceived shape or a recognized object. 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 (Hoffmann et al., 2018) to illustrations in children's books (Ganea et al., 2008) to forms in abstract paintings (Schmidt et al., 1989; Vinker et al., 2022) to figures in scientific papers (Franconeri et al., 2021).

✉ Kushin Mukherjee
kmukherjee2@wisc.edu

[1] Department of Psychology & Wisconsin Institute for Discovery, University of Wisconsin-Madison, Madison, WI, USA

The ability to discern resemblance between drawings and the shapes or objects they depict 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 (DeLoache et al., 1979; Kobayashi et al., 2020), and drawing recognition is generally robust in childhood (Cox, 2013; Hochberg & Brooks, 1962). 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 (Tanaka, 2007); pigeons, despite their famed capacity for visual recognition, show the same pattern (Cabe, 1976). 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, as well as the ability to perceive similarity of form even for novel or unrecognizable figures.

The current paper uses people's ability to perceive similarities between simple line drawings of objects and abstract

Published online: 30 May 2024

shapes 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 (Lin et al., 2014), answering questions about a given image using natural language (Goyal et al., 2017; Jang et al., 2017), generating sketches (Vinker et al., 2022), and even solving entire families of visual tasks (Zamir et al., 2018). 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 1,000 possible mutually exclusive categories (Kriegeskorte, 2015; Nayebi et al., 2018; Yamins et al., 2014; although refer to Konkle & Alvarez, 2020; Orhan et al., 2020; Zhuang et al., 2021) for some notable exceptions). From the perspective of human visual cognition, such models are interesting because they generalize well to images depicting new examples of the trained classes (Deng et al., 2009) and thus offer a potential mechanism for understanding key phenomena such as recognition invariance 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 (Cadieu et al., 2014; Kriegeskorte, 2015; Sexton & Love, 2022; Storrs et al., 2020; Yamins et al., 2014).

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 (Fan et al., 2018; Yang & Fan, 2021). 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 (Selvaraju et al., 2017; Shrikumar et al., 2017) and others have investigated inductive biases in such models (Geirhos et al., 2018; Hermann et al., 2020), 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 (Sangkloy et al., 2016).

There are many different DNN architectures and training methods Contemporary interest in DNNs as models of human perception began with convolutional networks (Krizhevsky et al., 2017), 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; Dosovitskiy et al., 2020); recently-introduced heuristics for training models (e.g. contrastive methods such as CLIP; Radford et al., 2021) appear to have a larger effect on their behavior than does the architecture *per se*; and models with qualitatively distinct architectures appear to capture macroscale neural patterns in ventral visual stream about equally well (Conwell et al., 2021), 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 (Bowers et al., 2022), including its decomposition into component parts; its orientation in space; its size; its distance from the observer; and, for recognizable objects, additional semantic information beyond the subordinate or basic category label. 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 (Baker et al., 2018).

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 (Bowers et al., 2022). Where such studies have been conducted, they have often focused on representation of photographic stimuli like those that constitute the model's training environment (Kriegeskorte 2015; Lake et al., 2015; Peterson et al., 2018) and it is not clear whether similar results would obtain for perception of more abstract and out-of-distribution stimuli such as sketches of objects and unrecognizable shapes.

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 line drawings of recognizable objects and abstract shapes?
2. Do the internal representations/features acquired by DNNs merely recapitulate other better-understood kinds of visual features, or do they capture aspects of perceived 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 (Jozwik et al., 2017), 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 judgments 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 (Fan et al., 2018), by considering which features best explain and predict the similarities that humans perceive amongst line drawings of recognizable objects and unrecognizable abstract shapes. 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. For instance, beyond neural-network-based features, people might be sensitive to overall similarity in shape, information encoded in different spatial frequencies, and the parts appearing in the object. 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 cognition. 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 both sketches of objects and other arbitrary, unfamiliar shapes. If so, mechanisms embodied in DNNs are *sufficient* to explain the human ability to cope with abstract visual depictions. Alternatively it may be that DNNs, because they are trained on photographs of real objects, acquire features that can represent perceived similarities amongst sketches of real objects, but do not extend well to unrecognizable shapes; or that the features acquired by DNNs are insufficient to explain the structure that people discern amongst drawings of either objects or unrecognizable shapes 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 perceived similarity amongst drawings.

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. The criteria for the similarity matching is intentionally unspecified: participants are free to use their own subjective judgments, based on whatever information they deem useful. Collection of many such judgments across many different participants then encompasses the variety of criteria people are generally inclined to use to adjudicate similarity. Triplet judgments are then used to embed the sketches within a low-dimensional space so that the Euclidean distance between pairs of sketches relates to the probability that the two items will be selected as "more similar" relative to some arbitrary third image (Jamieson et al., 2015). The resulting embeddings thus encode a low-dimensional *human* representational space for the images.

To determine which features govern the organization of this space, we then conducted two analyses. The first used regression techniques to predict the coordinates of the various drawings in the human-derived embedding space from other representational spaces derived from five different DNNs, from other kinds of visual features, or from both together. Comparison of model fit and regression coefficients

across these analyses then shed light on the three core questions raised above. The second analysis investigated how well human judgments on the triplet task could be predicted from the various different representational spaces, either alone or in combination. This analysis allowed us to assess which features, independently or together, are sufficient to explain behavioral decisions about perceived similarities amongst sketches. Within this general framework, experiment 1 focused on line drawings of four common object categories–birds, dogs, chairs, and cars–while experiment 2 focused on drawings depicting complex but unrecognizable abstract shapes (specifically the stimuli from (Schmidt & Fleming, 2016)).

Experiment 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 (Tversky, 1989), 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 four 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.

Behavioral methods

Participants A total of 85 participants were recruited via Amazon Mechanical Turk (mTurk) using CloudResearch (36 Female, 47 Male, two other; mean age = 38.69 years). 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 (Fan et al., 2020) for our similarity judgment study. These drawings were made in Pictionary-style *reference game*, where a sketcher and a guesser were simultaneously shown the same set of four images. The sketcher was tasked with drawing one of the four images and the guesser had to guess which of the four images the sketcher was tasked to draw. Each image belonged to one of four categories—birds, dogs, cars, or chairs—and each category had eight unique exemplars. Additionally, in some trials, the target image belonged to the same basic-level category as the three distractors leading to more detailed drawings by the sketcher, while on other trials all four images belonged to different categories leading the sketcher to make simpler drawings. We sampled two drawings from each condition (2) × category (4) × 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 (Mukherjee et al., 2019). 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 judgment task (Jamieson et al., 2015) implemented using the SALMON online tool for collecting triplet queries and fitting embeddings (https://github.com/stsievert/salmon). On each trial, participants viewed three drawings: a *target* positioned at the top of the screen two *options* positioned below it. They were instructed to select which of the two 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 (Fig. 2) and were used to estimate mean inter-subject agreement for the task. Based on prior work

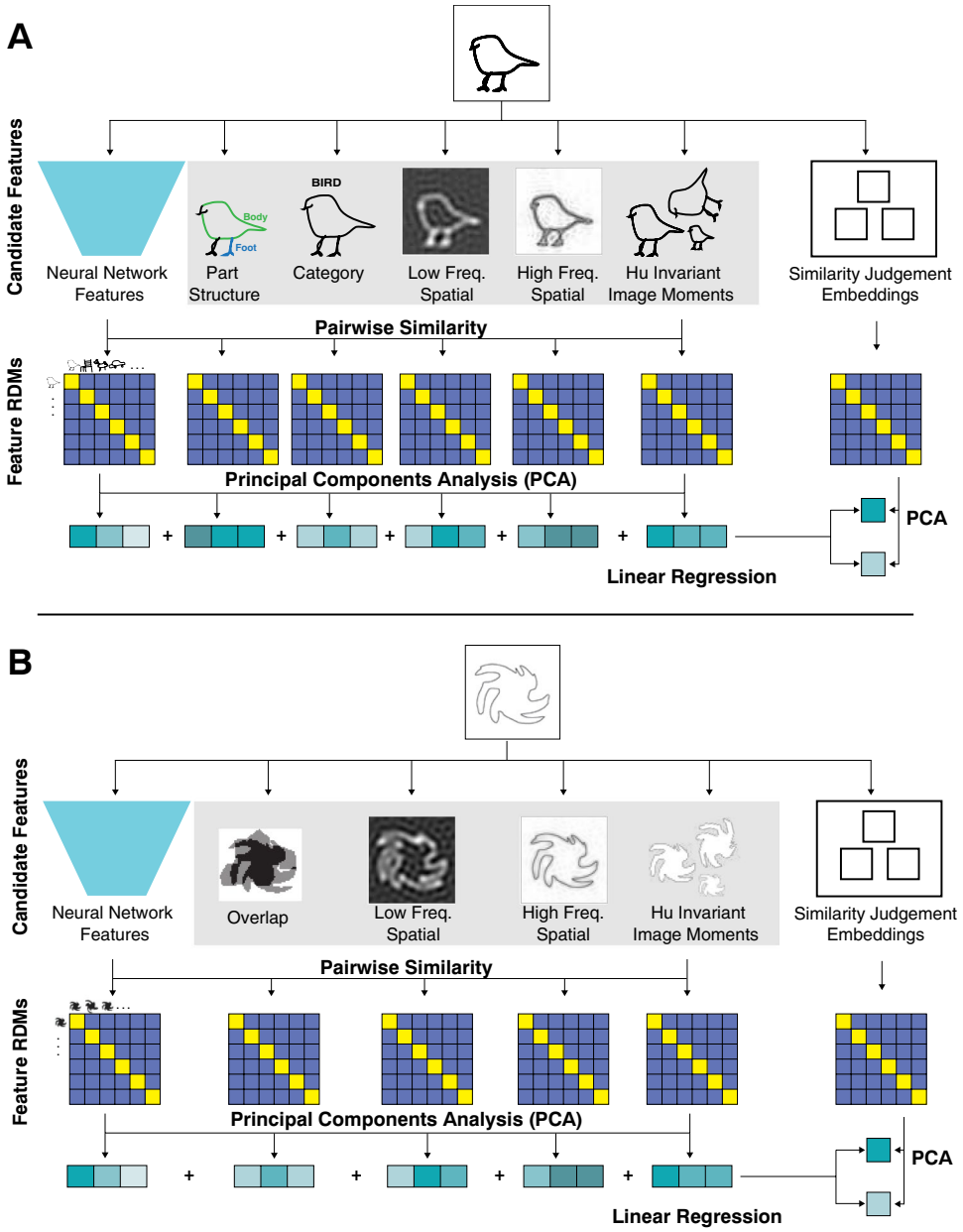

Fig. 1 Procedure for fitting linear regression models to predict human judgment embeddings from candidate features. **A.** In Experiment 1, features were constructed using part-structure, category, low spatial frequency, high spatial frequency, and shape information. Additionally, latent feature activations were extracted from five 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 five 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 judgment 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. (Color figure online)

using this paradigm, participants with a mean response time less than 1,500 ms 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

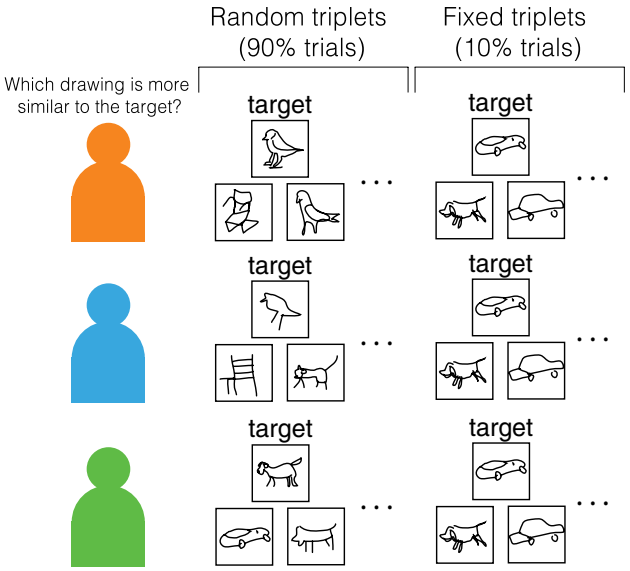

Fig. 2 Structure of the triplet similarity judgment task. Each participant (here, represented with different colors) completed 200 trials, indicating which of two options was most similar to a target drawing in each trail. 180 trials sampled triplets randomly from the set of all possible triplets. The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

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 judgment-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 (Tamuz et al., 2011). 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 result was a 2D embedding shown in Fig. 3A that predicted human decisions for held-out items with accuracy of 72.70%, comparable to interparticipant agreement of 73.10% (one-sample t test, $p = .62$) for the same triplets.

Neural network feature activations Neural network features were extracted using the THINGSVision Python Toolbox (Muttenthaler & Hebart, 2021) and focusing on five different DNNs including (1) AlexNet, a convolutional neural network (Krizhevsky et al., 2017) that was one of the first to achieve near human-level performance at image categorization; (2)

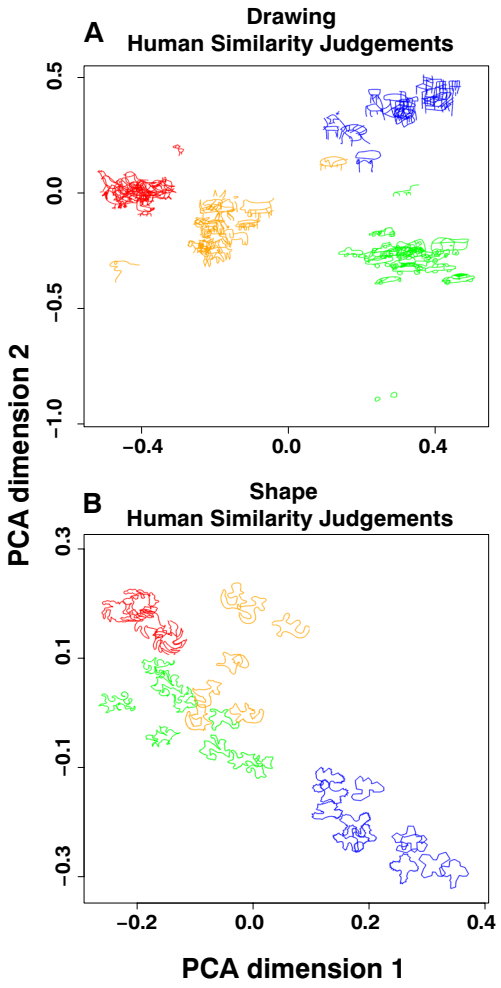

Fig. 3 Visualization of the locations of (top) line drawings and (bottom) abstract shapes along the first and second principal components computed from human similarity judgments. In the top panel, drawings of living objects are separated from nonliving objects along Dimension 1. (Color figure online)

VGG-19 (Simonyan & Zisserman, 2014), a deeper convolutional neural network with 19 layers; (3) ResNet-18 (He et al., 2016), 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), specifically ViT base with patch size 32 (Dosovitskiy et al., 2020), a (non-convolutional) Transformer-based neural network (Vaswani et al., 2017) trained for image classification; and (5) CLIP-ViT, a multimodal variant of the same vision transformer as (4) 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 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 (Cadieu et al., 2014; Yamins et al., 2014). The three variants we studied differ in two respects. First, (2) and (3) possess many more convolutional layers (i.e., 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 multiway 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 (Vaswani et al., 2017) 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 (Dosovitskiy et al., 2020). 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 (Tuli et al., 2021), 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 language-image pretraining," or CLIP.

To extract model internal representations, each drawing was first transformed to a standard 224 × 224 pixel size. Since the drawings are grayscale and most models expect a 3D tensor, the same 224 × 224 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. We focused on these deep layers because several prior studies have found that such representations better capture human behavior for both photographs and sketches of objects (Battleday et al., 2021; Hong et al., 2016; Jozwik et al., 2017; Singer et al., 2022). 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 five 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 the following:

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 four 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 judgments. 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 nonexpert 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 (Navon, 1977; Tversky, 1989). Indeed, classic structural descriptive theories have posited that visual representations are built from the constituent parts that make up an object (Biederman, 1987). Furthermore, people are capable of ascribing meaningful labels to the constituent parts (Jozwik et al., 2017). 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). 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 (Booth & Rolls, 1998; Karimi-Rouzbahani et al., 2017a, 2017b). 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 (Huang & Leng, 2010) using the openCV library. Hu moments, specifically, are a set of seven 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 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 judgment 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—

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

where X and Y are the two 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 overrepresented in the downstream analyses.

Results

How well do DNN-based embeddings explain human-perceived similarities amongst stimuli? 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 Fig. 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 the first component aligned with the direction of greatest variation and that the second component 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 Fig. 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 Fig. 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

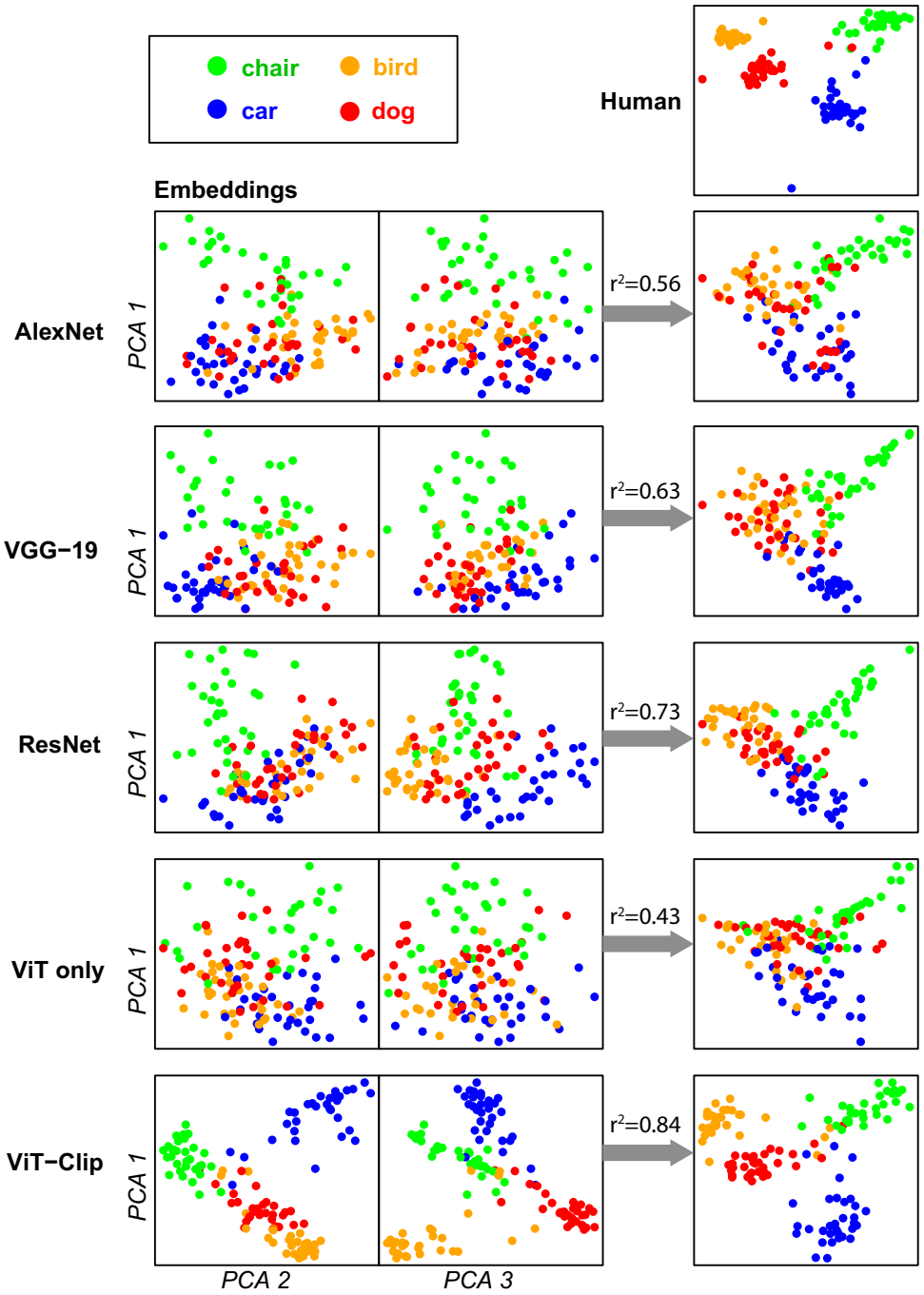

Fig. 4 Sketch embeddings in the regression analyses. The top right panel shows the human-based embeddings rotated to ensure the two components that constitute the dependent measure in the regressions are orthogonal to one another. Within each remaining row, the left plots show the 3D embeddings generated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

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 ($p < .001$ for all contrasts against null hypothesis), indicating that all architectures capture structure that is nonarbitrarily 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 Fig. 4. The different models varied somewhat in this metric,

but the CLIP-trained transformer model captured the most variance ($r^2 = .84$), reliably better than the next-best ResNet model ($p < .001$). 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—that is, without CLIP—did not cleanly separate semantic classes, and showed the worst accuracy predicting human-based embedding coordinates.

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 the left column of Table 1. All candidate 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 91% of the variance in the human-derived embedding 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 may seem to indicate that human judgments are dominated by information about semantic category. Yet the human-derived embedding in Fig. 3 also shows substantial variation amongst different exemplars of each category: The different sketches of items in a given category are not embedded identically, but form a sort of cloud. Thus, for instance, sketches appearing in the lower left of the chair cluster lie at some remove from those in the upper right (within-category distance), and are somewhat closer to the dog sketches (between-category distance). Such variation may reflect random variation arising from stochasticity in the triplet data, or it may capture the influence of other kinds of information beyond category membership alone.

Table 1 The amount of variance in human perceived similarity in drawings explained by each non-DNN candidate feature for the full embedding space (left column) and mean variance explained within categories (right column)

Feature	R^2 (full embedding)	Mean R^2 (within cat. only)
Category	0.91***	0.00
Parts	0.80***	0.28**
Low freq. spatial	0.22***	0.50***
High freq. spatial	0.17***	0.34**
Hu moments	0.16***	0.23**

Note. For each feature type, two 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. For the within-category column, separate models were fit for each category, and R^2 values were averaged across models. *** indicates significance at the .001 level, and ** indicates significance at the .01 level

To adjudicate these possibilities, we replicated the analysis, but fitting separate regressions for each of the four categories. The right column of Table 1 shows the proportion of within-category variance explained by each candidate space, averaged over the four categories. Since all category members have the same category label, the category features transparently do not explain any within-category variation; however, each of the other feature types do explain significant within-category variation–indicating that human-perceived similarities are not solely driven by category information, but also reflect other kinds of visual structure including, potentially, parts, spatial frequency information, and overall shape (Hu moments).

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.

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 nonzero, 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

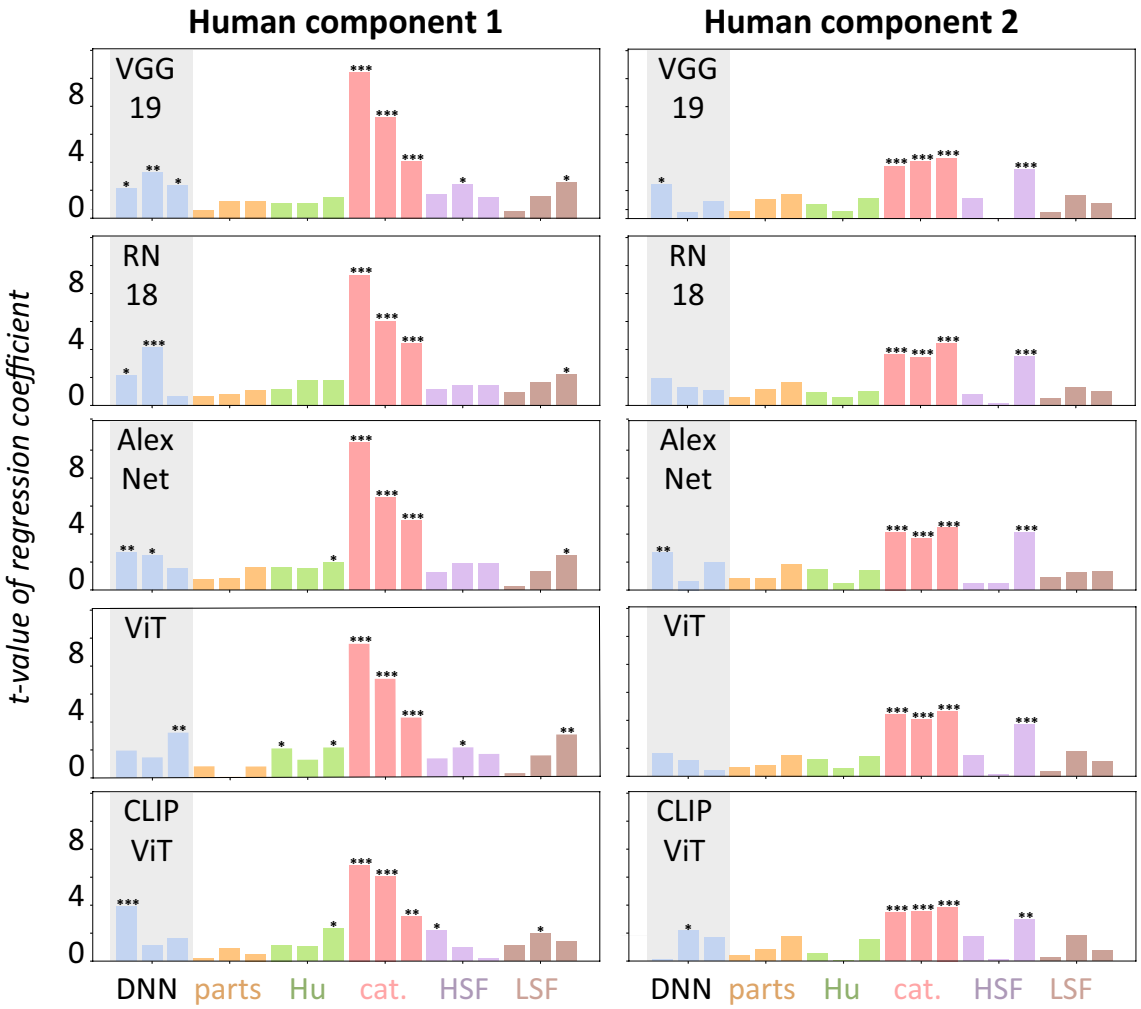

Fig. 5 Regression coefficients from Experiment 1. Rows shows *t* values on regression coefficients predicting Components 1 (left) or 2 (right) of the human embeddings from a combination of candidate features and neural network features extracted from five different architectures. Asterisks indicate coefficients that reliably improve model fit with $*p < .01$, $**p < .01$, or $***p < .001$. DNN = deep neural network; parts = part-based vectors; Hu = Hu moments; HSF = high spatial frequency; LSF = low spatial frequency. (Color figure online)

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 ($p < .05$ 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 using two different metrics.

First, recall that each human participant judged a fixed set of 20 "validation" triplets. Thus we had 85 judgments on each of the 20 triplets, and for each could compute (a) which option was most often chosen across subjects and (b) what proportion of participants agreed with that "majority vote" decision. This in turn provided an estimate of intersubject agreement that provides a benchmark for evaluating different representational spaces: A representation that predicts decisions at the level of the intersubject agreement performs as well as the average individual participant. We therefore predicted responses on the validation triplet trials from each candidate space by simply looking to see, within the corresponding RDM, which of the two option sketches

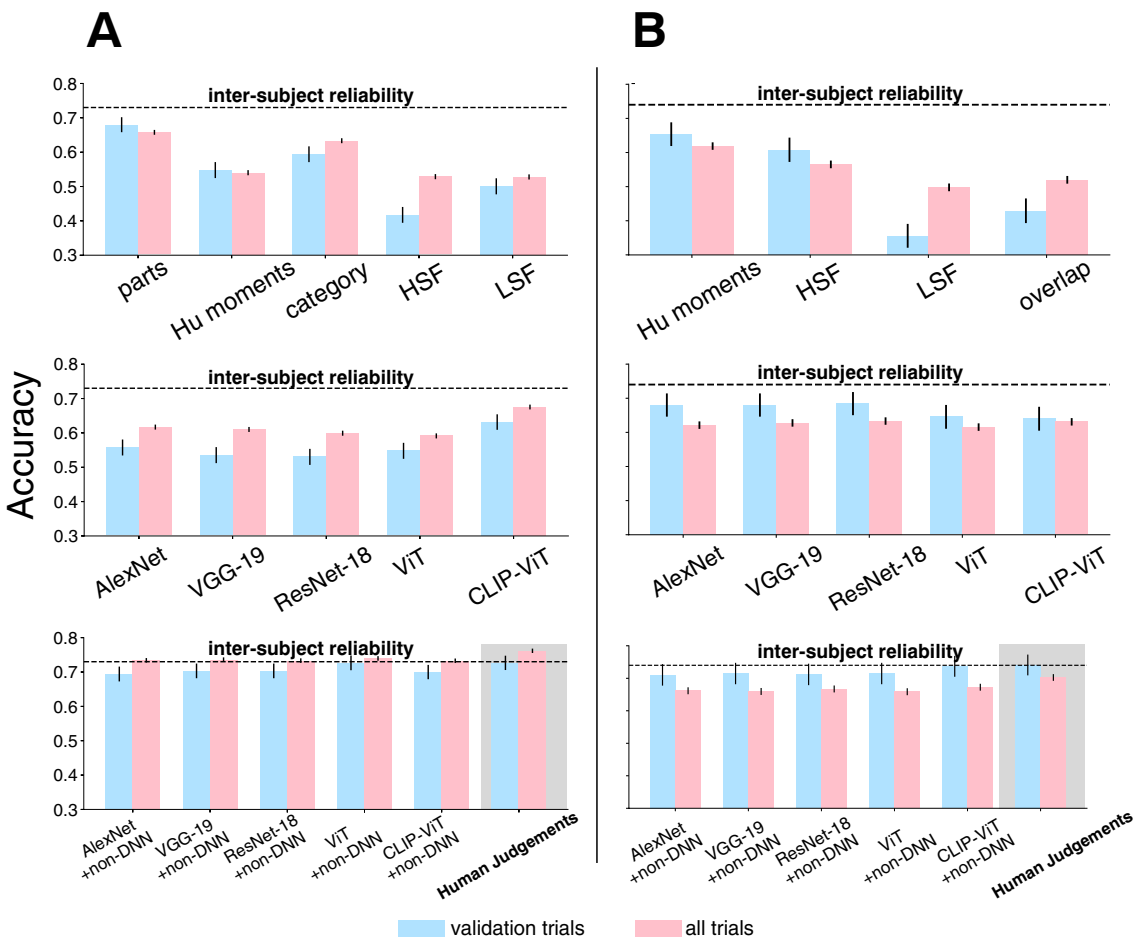

Fig. 6 Accuracy of predicted human similarity decisions for drawings (A) and shapes (B) for "validation" triplet judgment trials shown to all participants (blue bars) and for all trials across all participants (red bars). Row 1 shows the predictive accuracy of psychologically motivated candidate features. Row 2 shows the predictive accuracy of neural network features. Row 3 shows the predictive accuracy of estimated human similarity judgment embeddings from DNN and non-DNN candidate features and from embeddings generated directly from the human judgments (gray background) The dashed line for each plot corresponds to intersubject reliability computed based on the validation triplet trials. Error bars indicate 95% confidence intervals. (Color figure online)

was least dissimilar (by cosine distance) to the target sketch in the full high-dimensional space. For each candidate representation, the predicted responses were then compared with human decisions and counted as "correct" when the model choice matched the human choice and incorrect otherwise. Judgments were predicted for a total of (20 trials × 85 participants) 1,700 items across participants.

One drawback of this approach is that validation data were only collected for the 20 triplet items. For this reason, we also computed prediction accuracy for each representational space across all triplet data collected (85 participants × 200 triplets = 17,000 observations). Results for the validation set and the full set are both shown in Fig. 6.

The dotted horizontal line indicates the mean intersubject agreement computed based on the validation trials, which represents an upper limit on how well any predictive model can do.

Because the validation trials were not used to compute the human embedding, they provide a true independent holdout set for evaluating the quality of this embedding, which in turn provides a basis of comparison for evaluating the other feature types. While all candidate neural network representations predict human responses better than chance, no representation on its own shows predictive accuracy equal to the intersubject agreement on the validation items. 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.

Red bars show predictive accuracy on the full set of triplet judgments collected from human participants. Note that, because these triplets were the basis for computing the

human-derived embeddings, the predictive accuracy of the human embeddings on these triplets is likely inflated due to overfitting; with this caveat, we provide the data as a relative point of comparison for the other feature types. The higher predictive accuracy on all trials vs validation trials indicates that the particular triplet items used for validation were more difficult on average than the set of all items shown to participants. Nevertheless, neural network predictive accuracy remained reliably lower than the intersubject agreement on these items.

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 somewhat similar structure, the category-based RDMs are derived from one-hot vectors, and so do not express any within-category structure (as shown in the previous analysis), nor any broader structure across categories. Prediction accuracies tabulated across all 17,000 triplets accord well with those estimated from the validation triplets alone.

Importance of category information in each analysis Recall that, in the first analysis when looking across the full embedding space, category membership on its own accounted for a remarkable 91% of the variance in perceived similarities amongst stimuli, raising the possibility that human decisions are based only on semantic category membership. Yet when the category features are used directly to predict decisions on the validation triplets, the accuracy was much lower (Fig. 6). To understand this seeming discrepancy, consider that, in the regressions, the variance to be explained arises from the squared distance of each item to the center of the embedding space. Most of this variance arises from the distance of each cluster to the center of the space; comparably little of the unexplained variance arises from within-cluster distances.

Category features can therefore explain much of the overall variation in the embedding solely by predicting the centroid of each category in the space, without accounting for any within-category structure. The same features explain zero variation in the embedding if one looks at each category separately.

Because category features do not capture any within-category structure and represent items from different categories as equally distal to one another, they are not helpful for guiding triplet decisions on a majority of trials. If all three items are from the same category, they will all have the same category features, and there is no basis for choosing one option over the other—the choice will be a coin flip. Likewise, if the target item and the two options all come from different categories, all three will be completely nonoverlapping in their category features, and there is no basis for choosing one option over another. Category features are only useful if one and only one option item is in the same category as the target. Since there are equal numbers of items in each of the four categories and triplet items are sampled with uniform probability, the likelihood of a triplet meeting this condition (and accounting for sampling without replacement) is a little less than 0.38. Thus on more than 62% of triplets, category features provide no information about which option is a better match to the target, and decisions must be a coin flip. Other kinds of features that do capture within- and between-category similarities have a basis for generating nonrandom predictions on all triplets. The fact that some such feature sets show better predictive accuracy than category features, then, suggests that human similarity decisions for these stimuli are not guided solely by perceived category membership.

Human judgments for triplets, in contrast, show significantly higher consistency than the category-based features predict–suggesting that they are informed by more than just semantic category information. To directly test this possibility, we assessed whether regression models fit separately for each semantic category could reliably predict item human-derived-embedding coordinates *within* each category. Since our goal was to assess whether there exists systematic within-category structure in the human-derived embedding, this analysis focused on a single model architecture (VGG-19).

Specifically, for each category separately and in two independent regressions, we fit models to predict item coordinates along Dimension 1 and Dimension 2 of the human-derived embedding. Predictors included the 3D coordinates for embeddings derived from VGG-19 representations, part features, Hu moments, low-spatial-frequency and high-spatial-frequency features. Category membership was omitted since this feature is identical for all category members. Since each category has 32 sketch exemplars and there are 15 possible predictors, we used a forward stepwise approach that entered potential predictors as simple effects into the developing model and retained those that significantly improved model fit. Adjusted r-squared values for each model are shown in Fig. 7. The fitted models accounted for more than 40% of the variation on each dimension within three of the categories (cars, chairs, and dogs, $p < .003$ for all tests against null), and for 30% along the second dimension for the fourth (birds; $p < .03$). The stepwise procedure selected a combination of neural-network and alternative features for all models except the bird category, where it selected only alternative features for Dimension 2.

These results suggest two important conclusions. First, human similarity judgments for these stimuli are not driven solely by semantic category membership; instead, our participants discerned reliable similarities and differences amongst sketches within the same semantic category.

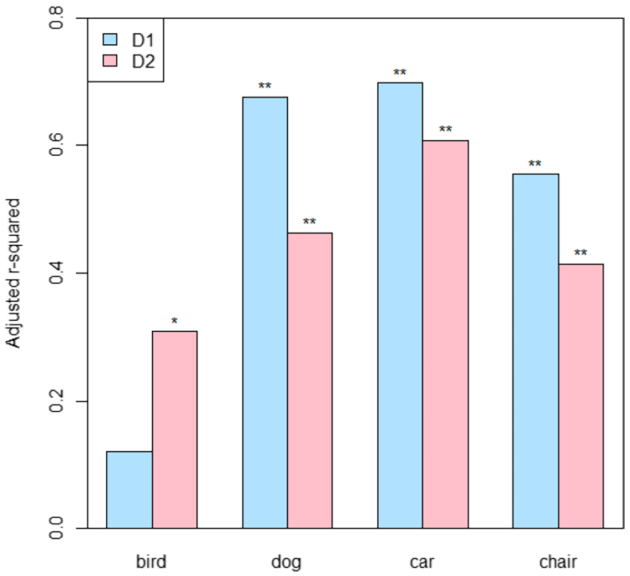

Fig. 7 Adjusted *r*-squared values for regression models predicting item coordinates along the first and second dimension of the human-derived embeddings, fit separately for each category. The fitted models explained significant variation along at least one dimension for all categories and accounted for more than half the variance along both dimensions for three of the four categories. (Color figure online)

Second, these within-category distances are best predicted via a linear combination of features drawn from both neural-network-based and other machine-vision methods—no single feature set was sufficient to explain this non-semantic structure.

Combining regression and triplet prediction Finally, we assessed whether the features under consideration are sufficient (in linear combination) to explain human-perceived similarities amongst the sketches as measured by the triplet task. Using the regression models fit in the first analysis, we generated *predicted coordinates* of the sketches in the human embedding space and from these computed the corresponding expected Euclidean distances between all image pairs.

The resulting RDM was then used to predict human decisions on the validation triplet set (again by choosing whichever option was nearest the target in the resulting space for each triplet). The results are shown in Fig. 6A (bottom) for predictions from regressions using each DNN embedding together with embeddings from other candidate representations. All models predicted human decisions at a level of accuracy similar to the intersubject 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 strongly influenced by semantic category membership, but also reflect other kinds of visual information captured by combinations of DNN-based and other features. While regression models built on category-based embeddings explained 91% of the variance in human-derived similarity spaces, category features alone were only somewhat better than chance at predicting triplet decisions.

Regression models reliably predicted within-category coordinates of sketches when fit independently for each; and predictions of regression models combining DNN and other features predicted triplet judgments at the level of intersubject agreement suggesting that a linear combination of these features is sufficient to explain human judgments in the triplet task.

The internal representations in DNNs capture the important categorical structure to the extent they cluster images by semantic category. As shown in Fig. 4, each model expresses at least some such structure, but the transformer architecture trained with CLIP shows the clearest clustering by category, and also yielded the best predictions of human-perceived structure amongst the different neural networks. No DNN, however, was sufficient on its own to explain human perceptual decisions for these stimuli. Finding a predicted stimulus embedding that predicted human triplet judgments at ceiling required a combination of both DNN and other features. 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 intersubject agreement.

These observations accord with prior studies of perceived similarities amongst photographs of objects, which likewise found that such structure is strongly but not completely influenced by semantic category membership (Jozwik et al., 2017; Mur et al., 2013; Peterson et al., 2016). 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 ViT-CLIP model shows 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. 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. Alternatively, it may be that the superior performance arises from the much larger training corpus used in ViT-CLIP, or from other differences between models.

A remaining question concerns the degree to which our behavioral results obscure the utility of DNNs or other feature sets to model human-perceived visual similarities in drawings, given that Experiment 1 used sketches of real items. The strong influence of semantic category on the human-derived sketch embeddings may arise 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. The DNNs we have considered were trained to assign inputs to semantic categories, and in this sense the representations they acquire may somewhat reflect semantic information—but certainly no model encompasses the rich range of semantic knowledge possessed by human participants. On this view, the use of object sketches may underestimate the utility of DNN-based features. On the other hand, because all models were trained to classify photographs of real objects—including examples of chairs, cars, birds, and dogs—it may be that the use of sketches depicting these categories inflated the apparent utility of DNN-based features for modeling human-perceived visual similarities. Experiment 2 adjudicates these different possibilities by replicating the procedures of Experiment 1, but using drawings of unrecognizable abstract shapes.

Experiment 2

Experiment 2 followed the same design as Experiment 1, but instead using line drawings depicting abstract shapes–specifically the set of 64 line drawings devised by Schmidt and Fleming (2016). These show bounded but visually complex shapes that are not recognizable as real-world objects (see Fig. 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 Forty participants were recruited via Amazon Mechanical Turk (mTurk) using CloudResearch (14 Female, 26 Male; mean age = 36.25 years). 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 four base shapes (Schmidt & Fleming, 2016). 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 525 × 525 pixel canvas.

Procedure for triplet judgment 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 less than 1,500 ms 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 five 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, nonmeaningful shape drawings. Before extracting neural network feature activations, the

images were resized to be 224 × 224 pixels to conform with each model's expectations.

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 Fig. 8. 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 embedding coordinates from model embeddings all account for significant variance ($p < .001$ for all contrasts to null), with the CLIP-trained transformer

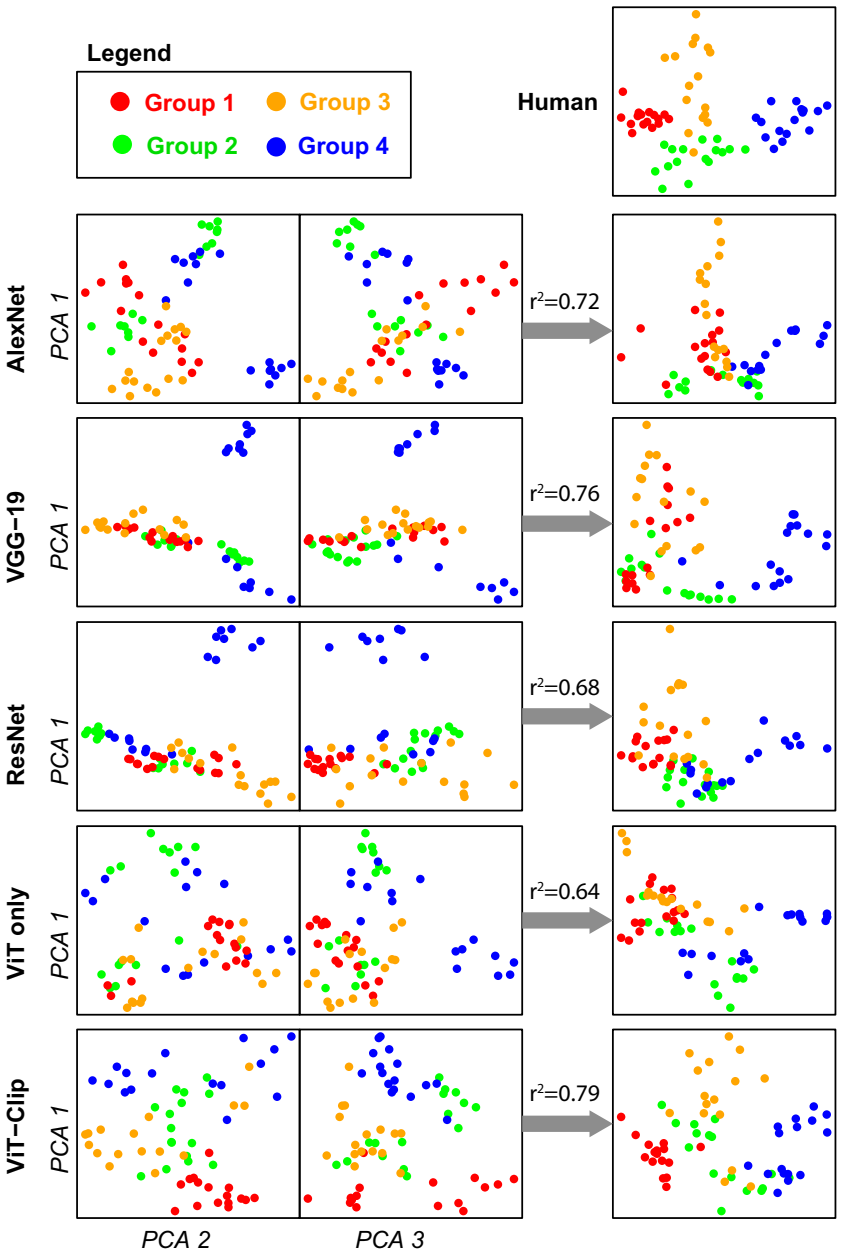

Fig. 8 Shape embeddings in the regression analyses. The top right panel shows the human-based embeddings rotated to ensure the two components that constitute the dependent measure in the regressions are orthogonal to one another. Within each remaining row, the left plots show the 3D embeddings generated from each DNN, and the right plot shows the predicted coordinates of the sketches within the human-based space after fitting regression models. (Color figure online)

Table 2 The amount of variance in human perceived similarity in abstract shapes explained by each non-DNN candidate feature

Feature	R^2	p value
Hu moments	.63	<.001
High freq. spatial	.34	<.001
Low freq. spatial	.48	<.001
Overlap	.41	<.001

Note. For each feature, two 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

again accounting for the most (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 ($p < .001$ 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%).

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 ($p < .001$ 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 Fig. 9. While all metrics account for significant 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.

To assess whether these results reflect the dimension reduction step, and to evaluate whether the features are sufficient in combination to explain human perceptual decisions, we again used the original RDMs for each vector space to predict human judgments on the validation items from the triplet task (i.e., the held-out items from which we can compute intersubject agreement). 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 Fig. 6B. Relative to the results with sketches, the DNN feature spaces alone show higher accuracy predicting human judgments for these non-semantic 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 perceptual decisions.

Red bars again show predictive accuracy on the full set of triplets. In this case, performance was systematically worse than for the held-out validation trials, indicating that the validation items were somewhat easier on average than other triplets. We again note that the human-derived embeddings can only be reliably compared with other features on the held-out validation trials, since the remaining full set of triplets was used to fit the human embedding—thus, the predictive accuracy of human-derived embeddings on these items is likely inflated.

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. Finally, stimulus embeddings generated via regressions combining DNN and other features predicted triplet judgments at the level of intersubject agreement for all DNN feature sets (Fig. 6B, bottom). As with sketches of real objects, a linear combination of DNN-based and alternative feature sets was sufficient to explain human-perceived similarities amongst these stimuli, though neither class of features was sufficient on its own.

Discussion of Study 2

From Study 1, it was unclear whether the use of sketches depicting real objects obscured or inflated the utility of

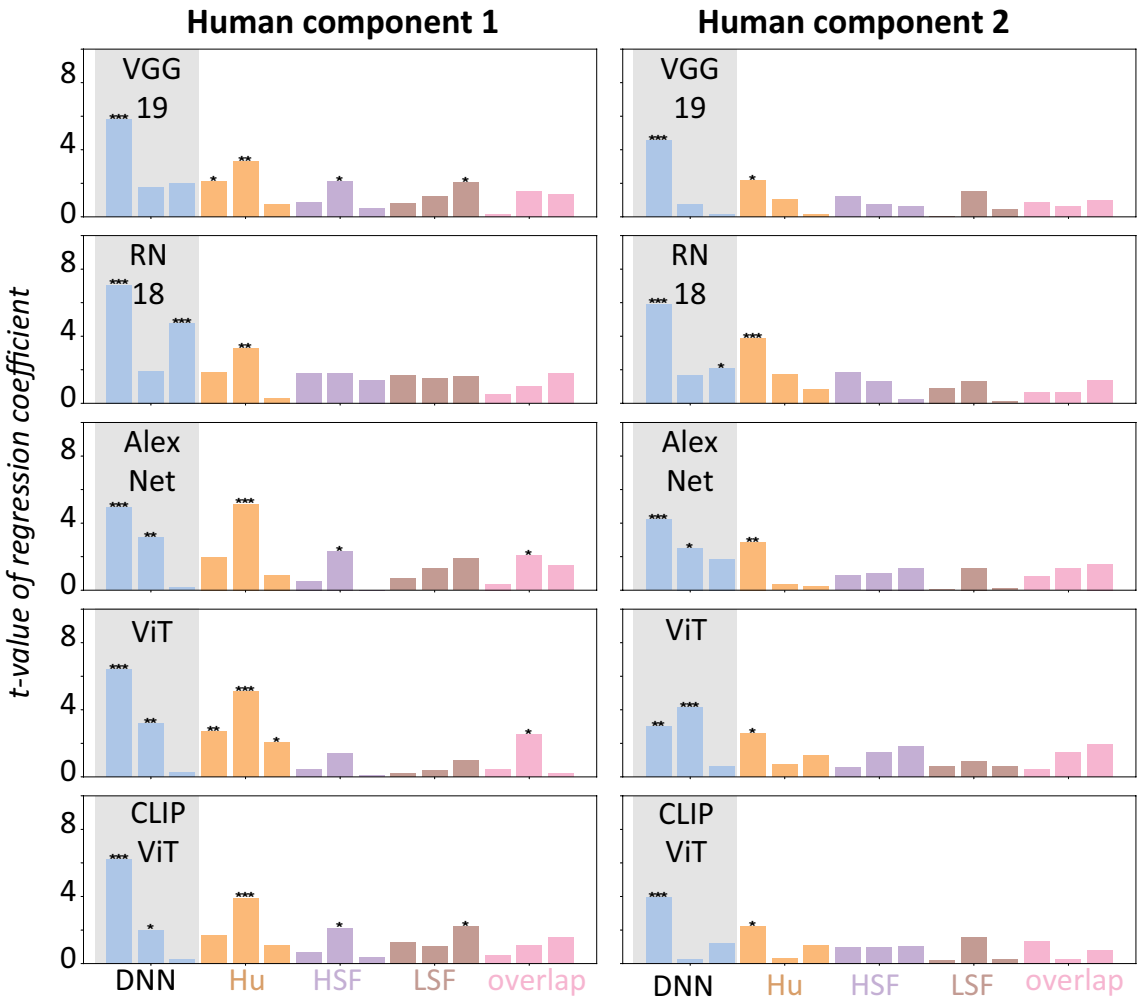

Fig. 9 Regression coefficients from Experiment 2. Rows shows *t* values on regression coefficients predicting Components 1 (left) or 2 (right) of the human embeddings from a combination of handcrafted features and neural network features extracted from five different architectures. Asterisks indicate coefficients that reliably improve model fit with $*p < .01$, $**p < .01$, or $***p < .001$. DNN = deep neural network; Hu = Hu moments; HSF = high spatial frequency; LSF = low spatial frequency; Overlap = degree of shape overlap. (Color figure online)

DNN-based visual features for explaining human-perceived similarity structure. Study 2 suggests that, when semantic information is not available to inform similarity judgments about drawings, DNNs capture substantially more information about human-perceived similarities, beyond that 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. Still, no DNN-based features were sufficient on their own to explain human perceptual similarity decisions for these items–ceiling-level prediction again required a combination of DNN- and non-DNN-based features.

The best-performing DNN-based features 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 likely appeared in

Table 3 The amount of unique variance explained by DNN features in ensemble models with all other candidate features

Feature	ΔR^2	*F*-statistic	*p* value
Human Judgments 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 Judgments 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

Note. Unlike in the case of drawings, DNN features explain a larger part of the variance

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 better captured by the CLIP-based internal representations.

The difference may, again, simply reflect the much larger training corpus for ViT-CLIP. Another possibility is that CLIP training aids 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. Adjudicating these possibilities will require a more apples-to-apples comparison of models trained on the same corpora and with the same architecture, but differing in the use of contrastive losses like CLIP.

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 and can discern similarity of form from drawings even for abstract shapes. 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. Most prior work in this vein has focused on perceived similarities amongst photographs of objects (Jozwik et al., 2017; Peterson et al., 2016). Efforts that have looked at the performance of deep neural networks on simple silhouettes (Baker et al., 2018; Kubilius et al., 2016) or drawings (Singer et al., 2022) have not contrasted DNNs to simpler feature spaces, or compared models varying in architecture and training methods. For both sketches of real objects and line drawings depicting unrecognizable shapes, we used human behavior in a triplet-judgment task to map a low-dimensional space capturing perceived similarities amongst stimuli. We then assessed whether internal representations extracted from various DNNs or other features spaces can explain the resulting structure.

For sketches of real items, we found that human similarity judgments were strongly influenced by the depicted item's basic-level semantic categories. 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. Yet despite this strong impact of semantics, we also found evidence that human similarity judgments are influenced by other visual aspects of a given sketch: inter-item distances within each category were reliably predicted by a linear combination of DNN-based and other features; and only regression models combining feature types were sufficient to fully explain human decisions on the triplet task. The strong impact of semantics coupled with significant contributions from other feature types raised the possibility that the use of real-object sketches might obscure or inflate the utility of DNN-based features for understanding perceived visual similarities.

To test this possibility, Experiment 2 conducted a parallel analysis for drawings of unrecognizable shapes. In this case, DNNs captured 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 interitem 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. Yet such representations were still insufficient on their own to fully explain such structure–as with Experiment 1, only models that combined DNN- and non-DNN-based features predicted human decisions at ceiling level.

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 drawings of objects and unfamiliar shapes?*

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 using regression methods, 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. 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 nontrivial 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. The result is surprising insofar as DNNs are primarily trained on photographs of real objects, and not drawings of abstract, unrecognizable shapes.

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

Our results suggest that the training task and/or corpus size 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. The good performance of ViT-CLIP may arise from its vastly larger training corpus, or from the CLIP training procedure, which promotes affinity in representation between images and their verbal descriptions. In so doing, this loss may promote 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. Adjudicating these possibilities will require comparison of better-matched models.

Broader implications The possibility that the superior performance of ViT-CLIP arises from the CLIP training procedure itself is intriguing, because it resonates with a well-known perspective on semantic representation in the mind and brain, namely the *hub-and-spokes* approach (McClelland & Rogers, 2003; Patterson et al., 2007; Rogers & McClelland, 2004; Rogers et al., 2004). The hub and spokes model proposes that different receptive and expressive information channels in the brain—vision, language, action, hearing—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 (Jackson et al., 2021; Lambon Ralph et al., 2017), 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.

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 (Lambon Ralph et al., 2007; McClelland & Rogers, 2003; Rogers & McClelland, 2004), the large-scale connectivity of the cortical semantic network (Binney et al., 2012; Chen et al., 2017; Jackson et al., 2021), functional imaging of neuro-semantic processing (Chen et al., 2017; Rogers et al., 2006, 2021), and results of transcranial magnetic stimulation (Pobric et al., 2010). Second, because ViT-CLIP representations yielded better agreement with human-derived similarities even for novel object shapes, it may be that encouraging agreement between vision and language representations of real stimuli promotes acquisition of visual features that better capture human perception generally, even for novel shapes. 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.

For these reasons, it will be important for future work to tease out the causes of the superior ViT-CLIP behavior, by comparing models matched for training corpus and architecture but varying in use of a contrastive loss. Such a comparison is beyond the scope of this paper, but recent work in this vein suggests that ViT-CLIP's generally superior behavior is not solely attributable to the large corpus used for training. For instance, Mayilvahanan et al. (2023) recently showed that CLIP models trained on smaller corpora relative to the CLIP model we tested here and with systematic holdouts in the validation image set show equally good out-of-distribution classification accuracy for both photographs and sketches. This work did not consider how well the resulting models capture human-perceived similarity structure, but strongly suggest that CLIP training leads to advantages beyond just the larger and more diverse training corpus.

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 (Morgenstern et al., 2021) provide a promising avenue to extend our approaches to the domain of naturalistic images. Coupled with novel methods for image-computable part-structure (Tiedemann et al., 2022), 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 (He et al., 2017; Li et al., 2022) and drawings (Li et al., 2018).

Lastly, while neural network vision models are capable of learning rich representations from their visual input, both ViT and CNN models can rely on features that humans do not use to guide their categorization judgments, such as background texture (Geirhos et al., 2018; Hermann et al., 2020; Tuli et al., 2021). This property might make it difficult for them to represent line drawings, where shape is more important for classification relative to photographs. We have considered 5 models that characterize the representations learned on standard "visual diets" (i.e., ImageNet and OpenAI's proprietary multimodal dataset), future work can seek to evaluate the degree to which training on more diverse datasets, such as Stylized-ImageNet (Geirhos, 2023), improves human-model alignment.

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 judgments.

References

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLOS Computational Biology, 14*(12), e1006613.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2021). From convolutional neural networks to models of higher-level cognition (and back again). *Annals of the New York Academy of Sciences, 1505*(1), 55–78.

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

Binney, R. J., Parker, G. J., & Lambon Ralph, M. A. (2012). 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.

Booth, M., & Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex (New York, NY: 1991), 8*(6), 510–523.

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F. G., Hummel, J., Heaton, R. F., Evans, B. D., Mitchell, J., & Blything, R. (2022). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences, 46*, 385. https://doi.org/10.1017/S0140525X22002813

Cabe, P. A. (1976). Transfer of discrimination from solid objects to pictures by pigeons: A test of theoretical models of pictorial perception. *Perception & Psychophysics, 19*(6), 545–550.

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLOS Computational Biology, 10*(12), e1003963.

Chen, L., Lambon Ralph, M. A., & Rogers, T. T. (2017). A unified model of human semantic knowledge and its disorders. *Nature Human Behaviour, 1*(3), 0039.

Conwell, C., Prince, J. S., Alvarez, G. A., & Konkle, T. (2021). What can 5.17 billion regression fits tell us about artificial models of the human visual system? *SVRHM 2021 Workshop@ NeurIPS*.

Cox, M. V. (2013). *Children's drawings of the human figure*. Psychology Press.

DeLoache, J. S., Strauss, M. S., & Maynard, J. (1979). Picture perception in infancy. *Infant Behavior and Development, 2*, 77–89.

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

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv Preprint.* arXiv:2010.11929

Fan, J. E., Yamins, D. L., & Turk-Browne, N. B. (2018). Common object representations for visual production and recognition. *Cognitive science, 42*(8), 2670–2698.

Fan, J. E., Hawkins, R. D., Wu, M., & Goodman, N. D. (2020). Pragmatic inference and visual abstraction enable contextual flexibility during visual communication. *Computational Brain & Behavior, 3*(1), 86–101.

Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., & Hullman, J. (2021). The science of visual data communication: What works. *Psychological Science in the Public Interest, 22*(3), 110–161.

Ganea, P. A., Pickard, M. B., & DeLoache, J. S. (2008). Transfer between picture books and the real world by very young children. *Journal of Cognition and Development, 9*(1), 46–66.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *ArXiv Preprint.* arXiv:1811.12231

Geirhos, R. (2023). *Stylized-ImageNet*.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6904–6913). IEEE.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. *Proceedings of the IEEE international Conference on Computer Vision* (pp. 2961–2969). IEEE.

Hermann, K., Chen, T., & Kornblith, S. (2020). The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems, 33*, 19000–19015.

Hochberg, J., & Brooks, V. (1962). Pictorial recognition as an unlearned ability: A study of one child's performance. *The American Journal of Psychology, 75*(4), 624–628.

Hoffmann, D. L., Standish, C. D., García-Diez, M., Pettitt, P. B., Milton, J. A., Zilhão, J., Alcolea-González, J. J., Cantalejo-Duarte, P., Collado, H., de Balbín, R., Lorblanchet, M., Ramos-Muñoz, J., Weniger, G.-C., & Pike, A. W. G. (2018). U-th dating of carbonate crusts reveals neandertal origin of Iberian cave art. *Science, 359*(6378), 912–915. https://doi.org/10.1126/science.aap7778

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

Huang, Z., & Leng, J. (2010). Analysis of Hu's moment invariants on image scaling and rotation International Conference on Computer. *2010 2nd International Conference on Computer Engineering and Technology, 7*, V7-476.

Jackson, R. L., Rogers, T. T., & Lambon Ralph, M. A. (2021). Reverse-engineering the cortical architecture for controlled semantic cognition. *Nature Human Behaviour, 5*(6), 774–786.

Jamieson, K. G., Jain, L., Fernandez, C., Glattard, N. J., & Nowak, R. D. (2015). Next: A system for real-world development, evaluation, and application of active learning. *NIPS*, 2656–2664.

Jang, Y., Song, Y., Yu, Y., Kim, Y., & Kim, G. (2017). Tgif-qa: Toward spatio-temporal reasoning in visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2758–2766). IEEE.

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology, 8*, 1726.

Karimi-Rouzbahani, H., Bagheri, N., & Ebrahimpour, R. (2017a). Hard-wired feed-forward visual mechanisms of the brain compensate for affine variations in object recognition. *Neuroscience, 349*, 48–63.

Karimi-Rouzbahani, H., Bagheri, N., & Ebrahimpour, R. (2017b). 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.

Kobayashi, M., Kakigi, R., Kanazawa, S., & Yamaguchi, M. K. (2020). Infants' recognition of their mothers' faces in facial drawings. *Developmental Psychobiology, 62*(8), 1011–1020.

Konkle, T., & Alvarez, G. A. (2020). Instance-level contrastive learning yields human brain-like representation without category-supervision. *BioRxiv*, 2020–06.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modelling biological vision and brain information processing. *bioRxiv*, 029876. https://doi.org/10.1101/029876

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM, 60*(6), 84–90.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLOS Computational Biology, 12*(4), e1004896.

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep neural networks predict category typicality ratings for images. In D. C. Noelle, R. Dale, A. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th annual meeting of the Cognitive Science Society* (pp. 1243–1248). Cognitive Science Society.

Lambon Ralph, M. A., Lowe, C., & Rogers, T. T. (2007). 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.

Lambon Ralph, M. A., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience, 18*(1), 42–55.

Li, L., Fu, H., & Tai, C.-L. (2018). *Fast sketch segmentation and labeling with deep learning*. IEEE Computer Graphics and Applications.

Li, F., Zhang, H., Liu, S., Zhang, L., Ni, L. M., Shum, H.-Y., et al. (2022). Mask DINO: Towards a unified transformer-based framework for object detection and segmentation. *ArXiv Preprint.* arXiv:2206.02777

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *European Conference on Computer Vision* (pp. 740–755). Springer.

Mayilvahanan, P., Wiedemer, T., Rusak, E., Bethge, M., & Brendel, W. (2023). *Does CLIP's generalization performance mainly stem from high train-test similarity?* arXiv preprint arXiv:2310.09562.

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience, 4*(4), 310–322.

Morgenstern, Y., Hartmann, F., Schmidt, F., Tiedemann, H., Prokott, E., Maiello, G., & Fleming, R. W. (2021). An image-computable model of human visual shape similarity. *PLOS Computational Biology, 17*(6), e1008981.

Mukherjee, K., Hawkins, R. D., & Fan, J. E. (2019, July). *Communicating semantic part information in drawings.* In: Poster presented at the 41st Annual Meeting of the Cognitive Science Society.

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human object-similarity judgments reflect and transcend the primate-it object representation. *Frontiers in Psychology, 4*, 128.

Muttenthaler, L., & Hebart, M. N. (2021). THINGSVision: A Python toolbox for streamlining the extraction of activations from deep neural networks. *Frontiers in Neuroinformatics, 15*, 45. https://doi.org/10.3389/fninf.2021.679838

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology, 9*(3), 353–383.

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., DiCarlo, J. J., & Yamins, D. L. (2018). Task-driven convolutional recurrent models of the visual system. *ArXiv Preprint.* arXiv:1807.00053

Orhan, E., Gupta, V., & Lake, B. M. (2020). Self-supervised learning through the eyes of a child. *Advances in Neural Information Processing Systems, 33*, 9960–9971.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience, 8*(12), 976–987.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2016). Adapting deep network features to capture psychological representations. *ArXiv Preprint.* arXiv:1608.02164

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science, 42*(8), 2648–2669.

Pobric, G., Jefferies, E., & Ralph, M. A. L. (2010). Category-specific versus category-general semantic impairment induced by transcranial magnetic stimulation. *Current Biology, 20*(10), 964–968.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning* (pp. 8748–8763). Authors.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT Press.

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., & Patterson, K. (2004). Structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological review, 111*(1), 205.

Rogers, T. T., Hocking, J., Noppeney, U., Mechelli, A., Gorno-Tempini, M. L., Patterson, K., & Price, C. J. (2006). Anterior temporal cortex and semantic memory: Reconciling findings from neuropsychology and functional imaging. *Cognitive, Affective, & Behavioral Neuroscience, 6*(3), 201–213.

Rogers, T. T., Cox, C. R., Lu, Q., Shimotake, A., Kikuchi, T., Kunieda, T., Miyamoto, S., Takahashi, R., Ikeda, A., Matsumoto, R., & Lambon Ralph, M. A. (2021). Evidence for a deep, distributed and dynamic code for animacy in human ventral anterior temporal cortex. *Elife, 10*, e66276.

Sangkloy, P., Burnell, N., Ham, C., & Hays, J. (2016). The Sketchy Database: Learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG), 35*(4), 119.

Schmidt, F., & Fleming, R. W. (2016). Visual perception of complex shape-transforming processes. *Cognitive Psychology, 90*, 48–70.

Schmidt, J. A., McLaughlin, J. P., & Leighten, P. (1989). Novice strategies for understanding paintings. *Applied Cognitive Psychology, 3*(1), 65–72.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international Conference on Computer Vision* (pp. 618–626). IEEE.

Sexton, N. J., & Love, B. C. (2022). Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science Advances, 8*(28), eabm2219.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *International Conference On Machine Learning* (pp. 3145–3153). Authors.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint.* arXiv:1409.1556

Singer, J. J., Seeliger, K., Kietzmann, T. C., & Hebart, M. N. (2022). 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.

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2020). Diverse deep neural networks all predict human it well, after training and fitting. *BioRxiv*.

Tamuz, O., Liu, C., Belongie, S., Shamir, O., & Kalai, A. T. (2011). Adaptively learning the crowd kernel. *ArXiv Preprint.* arXiv:1105.1033

Tanaka, M. (2007). Recognition of pictorial representations by chimpanzees (*pan troglodytes*). *Animal Cognition, 10*(2), 169–179.

Tiedemann, H., Schmidt, F., & Fleming, R. W. (2022). Superordinate categorization based on the perceptual organization of parts. *Brain Sciences, 12*(5), 667.

Tuli, S., Dasgupta, I., Grant, E., & Griffiths, T. L. (2021). Are convolutional neural networks or transformers more like human vision? *ArXiv Preprint.* arXiv:2105.07197

Tversky, B. (1989). Parts, partonomies, and taxonomies. *Developmental Psychology, 25*(6), 983.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Vinker, Y., Pajouheshgar, E., Bo, J. Y., Bachmann, R. C., Bermano, A. H., Cohen-Or, D., Zamir, A., & Shamir, A. (2022). Clipasso: Semantically-aware object sketching. *ArXiv Preprint.* arXiv:2202.05822

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, 111*(23), 8619–8624.

Yang, J., & Fan, J. E. (2021). Visual communication of object concepts at different levels of abstraction. *ArXiv Preprint.* arXiv:2106.02775

Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., & Savarese, S. (2018). Taskonomy: Disentangling task transfer learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3712–3722). IEEE.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences, 118*(3), e2014196118.

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