SEVA: Leveraging sketches to evaluate alignment between human and machine visual abstraction

Kushin Mukherjee[1,*] Holly Huey[2,*] Xuanchen Lu[2,*] Yael Vinker[3], Rio Aguina-Kang[2], Ariel Shamir[4], Judith E. Fan[2,5]

University of Wisconsin-Madison[1]
University of California, San Diego[2]
Tel-Aviv University[3]
Reichman University[4]
Stanford University[5]

Abstract

Sketching is a powerful tool for creating abstract images that are sparse but meaningful. Sketch understanding poses fundamental challenges for general-purpose vision algorithms because it requires robustness to the sparsity of sketches relative to natural visual inputs and because it demands tolerance for semantic ambiguity, as sketches can reliably evoke multiple meanings. While current vision algorithms have achieved high performance on a variety of visual tasks, it remains unclear to what extent they understand sketches in a human-like way. Here we introduce `SEVA`, a new benchmark dataset containing approximately 90K human-generated sketches of 128 object concepts produced under different time constraints, and thus systematically varying in sparsity. We evaluated a suite of state-of-the-art vision algorithms on their ability to correctly identify the target concept depicted in these sketches and to generate responses that are strongly aligned with human response patterns on the same sketch recognition task. We found that vision algorithms that better predicted human sketch recognition performance also better approximated human uncertainty about sketch meaning, but there remains a sizable gap between model and human response patterns. To explore the potential of models that emulate human visual abstraction in generative tasks, we conducted further evaluations of a recently developed sketch generation algorithm [91] capable of generating sketches that vary in sparsity. We hope that public release of this dataset and evaluation protocol will catalyze progress towards algorithms with enhanced capacities for human-like visual abstraction.

1 Introduction

Abstraction is key to how humans understand the external world. Abstraction enables distillation of individual sensory experiences into compact latent representations that support learning of new concepts [80, 45, 66, 30] and efficient communication about these concepts with others [23, 85, 34, 28]. For example, while no two roses are identical, people can rapidly infer what properties make a flower a *rose* and not some other kind of flower from just a few examples [99, 50], especially when these examples are selected to support such strong inferences [32, 77].

[1/*] Equal contribution

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

1.1 Human Visual Abstraction as Key Target for AI

Visual abstraction enables humans to express what they know about the visual world by creating external representations that highlight the information they judge to be most relevant in any given context—for instance, pictures that highlight the visual features that are diagnostic of a *rose* in a botanical field guide [24, 23, 92]. Critically, there are many different ways to depict even the same object—from a detailed illustration to a simple sketch. The Spanish artist Pablo Picasso famously demonstrated this point in *The Bull* (1945), a series of 11 lithographs of bulls, each sparser than the last (Fig. 1). While some of the drawings in this series look more realistic and others more stylized, all of these images remain evocative of a *bull* (and perhaps other similar animals, such as a *moose* or *buffalo*) to most human viewers.

Drawing is one of the most accessible, enduring, and versatile techniques that humans in many cultures use to encode ideas and emotions in visual form [41, 1, 31]. Even without special training, humans can robustly produce and understand simple line drawings or sketches of familiar visual concepts [24, 81, 43]. The ability to leverage drawings to understand and convey key aspects of the visual world emerges early in childhood [62, 5, 44] and improves throughout development with children's expanding conceptual knowledge [19, 59, 42]. Moreover, failures to produce and recognize drawings of objects are associated with semantic dementia [10, 73], suggesting links between a robust capacity for visual abstraction and the organization of semantic knowledge in the brain.

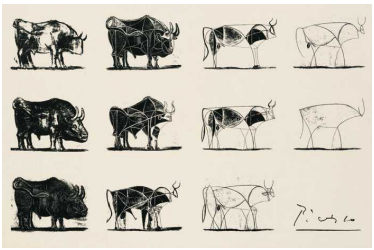

Figure 1: Pablo Picasso. *The Bull*, 1945.

In addition to drawings that represent objects and scenes, other abstract human-made visualizations (e.g., maps, diagrams, charts, graphs) serve important functions in many domains, including all branches of science and engineering [89, 90, 39, 13, 11]. Given the ubiquity and importance of such visualizations in modern life, developing computational models that achieve human-like understanding of freehand sketches is an important milestone. Such computational models of *human* visual abstraction stand to not only advance our understanding of human intelligence, but to also make AI systems more robust and general [40, 63, 27]. For example, prior work has found that incorporating principles based on the structure and function of the human visual system have led to vision models that are more robust (e.g., to adversarial attacks) [6, 55, 25].

1.2 Desiderata for Evaluating Alignment Between Human and Machine Visual Abstraction

The past several years have seen remarkable progress in the development of increasingly performant general-purpose vision algorithms [78, 36, 20, 70], with some of the most prominent algorithms also capable of emulating key aspects of how the primate brain encodes natural visual inputs [102, 48, 110, 47]. Over the same period, artificial vision systems have also been steadily achieving higher performance on tasks involving abstract visual inputs, including sketch categorization [21, 108, 4, 106], sketch segmentation [54, 104], sketch-based image/shape retrieval [22, 74, 105, 82, 9], among others [88, 100, 60, 16]. Moreover, these models have been found to predict human behavior on sketch recognition tasks to some degree [24, 23]. However, these otherwise high-performing vision algorithms struggle to simultaneously achieve robust understanding of visual inputs across multiple levels of abstraction [3, 79, 23]. Moreover, one study found that current vision models trained on natural images still fall short of the representational capabilities of the inferotemporal cortex, a key brain region supporting object categorization, in generalizing to new image distributions, including sketches [2]. Nevertheless, more recently developed models trained on substantially larger and varied datasets show promise on both tasks involving images with different visual styles [70, 107, 76] and tasks that go beyond recognition, including sketch generation [91, 69, 97] and sketch-guided image generation [61, 109, 96, 93, 56].

At present, it remains unclear to what degree any state-of-the-art models achieve *human-like* understanding of line drawings that vary in their degree of abstraction, much less the full range of abstract images that humans regularly engage with. Gaining further clarity on this question requires meeting two key challenges: *first*, creating a dataset containing drawings of a wide variety of object concepts that also systematically vary in their degree of abstraction; and *second*, developing evaluation protocols that can be used to estimate the degree to which any model emulates *human-like* understanding of this suite of drawn images.

Dataset. Meeting the first challenge requires going beyond existing sketch datasets [21, 74, 43, 18, 22, 54, 105, 82, 26, 95, 75, 51–53, 67, 111, 57]. While most of these datasets span a reasonably wide range of visual concepts (i.e., ranging from 125 concepts in *Sketchy* to 345 in *Quickdraw*) and some of them contain fine-grained information (e.g., stroke information and photo-sketch pairing), none of them systematically varied how detailed individual sketches could be, one of the most straightforward ways of inducing variation in semantic abstraction [8, 23, 103]. Our paper addresses the gap by providing sketches with controlled levels of detail, while encompassing the variety and granularity present in existing datasets (Table 1).

Evaluation protocol. Meeting the second challenge requires going beyond simple accuracy-based model performance metrics alone. Instead, it is critical to measure detailed patterns of human behavior on the same sketch understanding tasks to evaluate how well any model emulates these behavioral *patterns*, following recent work in the computational neuroscience of vision [71, 7, 68].

1.3 SEVA: A Novel Sketch Benchmark for Evaluating Visual Abstraction in Humans and Machines

In recognition of the above desiderata, here we introduce `SEVA` (**S**ketch-based **E**valuations of **V**isual **A**bstraction), a new sketch dataset and benchmark for evaluating alignment between human and machine visual abstraction.

Dataset. Our dataset contains approximately 90K human-generated sketches of a wide variety of visual objects that also systematically vary in their level of detail, and thus the variety of meanings they evoke. Each sketch is associated with one of 2,048 object instances belonging to one of 128 object categories selected from the `THINGS` dataset [38]. We achieved variation in sketch detail by imposing constraints on how much time humans (N=5,563 participants) had to produce each sketch (i.e., 4s, 8s, 16s, 32s).

Evaluation protocol. Leveraging these human-generated sketches, we systematically evaluated how well a diverse suite of 17 state-of-the-art vision models generate classification responses that align with those produced by humans (N=1,709 participants) tasked with identifying the most appropriate concept label for each sketch. Of these participants, 579 participants also participated in the sketch production study but were not shown any of their own sketches during this study. We evaluated human-model alignment using three different metrics: (1) top-1 classification accuracy, reflecting raw sketch recognition performance; (2) Shannon entropy of the response distribution, reflecting the degree of uncertainty about the target label; and (3) *semantic neighbor preference*, reflecting the degree to which models and humans generated off-target responses that were semantically related to the target label.

Summary of key findings. We found that sparser human sketches produced under more severe time pressure (e.g., 4 seconds) exhibited greater *semantic ambiguity*—in other words, both humans and models assigned a greater variety of labels to them than to the more detailed sketches that took more time to make (e.g., 32 seconds). Furthermore, we found that models that better predicted human sketch recognition performance also better approximated human uncertainty about sketch meaning, but none of the models came close to approximating human response patterns to human-generated sketches at any level of detail. To explore the potential of models that emulate human visual abstraction in generative tasks, we conducted further evaluations of `CLIPasso`, a recently developed sketch generation algorithm [91] capable of generating sketches that vary in sparsity (measured by number of constituent strokes in a sketch). We discovered that the most detailed `CLIPasso`-generated sketches converged with human sketches of the same object concepts, as measured by the distribution of labels humans assigned to sketches made by both agent types; however, sparser `CLIPasso` sketches diverged from human sketches of the same object concepts, reflecting a gap between how `CLIPasso` and human participants attempted to preserve sketch meaning under more severe production constraints.

2 Methods

2.1 Human Sketch Production

A core contribution of this work is a new dataset containing human-generated sketches of a wide range of visual object concepts that also systematically span multiple levels of semantic abstraction. We created this dataset by crowdsourcing these sketches online, following prior work [24, 105, 74, 23, 34, 43]. Each sketch in the dataset was recorded as a bitmap image as well as a collection of stroke coordinates, thus preserving the precise cursor movements a participant enacted to create the sketch.

Figure 2: Humans and CLIPasso generated approximately 90K sketches under various production constraints.

Participants. 5,563 participants (2,870 male; M_{age} = 36.7 years) were recruited from Prolific and compensated $15.50/hour for their participation. Data from 104 of these sessions were excluded from subsequent analyses due to technical issues (e.g., images did not load). All participants provided informed consent in accordance with the UC San Diego IRB.

Object concepts. We included 128 concrete real-world object categories (e.g., "lion", "banjo", "car") sourced from the THINGS dataset. We used the THINGS dataset [38, 37] because it is a well validated set of concrete, real-world visual object categories designed to support interoperability among large-scale studies in human visual cognition and cognitive neuroscience. For each of these 128 object concepts, we randomly sampled 16 object instances represented by color photographs, which served to visually ground the human sketch production task. As such, each sketch in our dataset is uniquely associated with one of these 2,048 object instances, and our final sample size was determined by our predefined goal of obtaining at least 10 human sketches of each of these instances.

Sketch production task. In each session, participants produced sketches of 16 different object categories, randomly sampled from the full set of 128 object categories. On each trial, they were cued with a color photograph (500px $\times$ 500px) of an object paired with its concept label. Each participant was randomly assigned to one of four conditions, defined by the maximum amount of time participants could take to produce their sketches: 4 seconds, 8 seconds, 16 seconds, or 32 seconds (Fig. 2, *left*). Such random assignment of participants to condition ensures that estimates of differences between conditions will not, in expectation, be biased by individual differences in sketching behavior. Participants drew on a digital drawing canvas (500px $\times$ 500px) using whatever input device they already had available (e.g., mouse, stylus) and were able to undo their most recent

Dataset Name	Dataset Contents	# Classes	Stroke Info?	Photo Cue?	Abstraction?
TU-Berlin [21]	20K sketches	250	✓		
QuickDraw [43]	50M sketches	345	✓		
QuickDrawExtended [18]	330K sketches, 204K photos	110			
SPG [54]	20K sketches w/ stroke grouping	25	✓		
SBSR [22]	1.8K sketches, 1.8K 3D models	161			
QMUL [105, 82]	1.3K sketches, 1.3K photos	3		✓	
Sketchy [74]	75K sketches, 12K photos	125	✓	✓	
SketchyCOCO [26]	14K sketches, 14K photos	17	✓	✓	
SEVA	**90K sketches, 2048 photos**	**128**	✓	✓	✓

Table 1: Comparison between SEVA and prior sketch datasets.

stroke or completely clear their canvas if needed. They were encouraged to make their drawings as recognizable as they could at the concept level and to use the photograph only to remind them of what individual objects belonging to that category generally look like. A countdown timer indicated how many seconds they had left to produce their drawing. Each trial ended either when time ran out or when the participant indicated that they wished to continue to the next trial, but participants were encouraged to use the full time available to produce as recognizable of a drawing as they could. At the beginning of the session, participants were explicitly instructed not to include any background context (e.g., grass in a drawing of a "horse"), arrows, or text. Participants also completed one practice trial (that we did not include in analyses) at the beginning of the session to familiarize themselves with the drawing interface. Our final dataset contains 89,797 sketches after filtering out invalid sketches (e.g., blank canvases).

2.2 Human Sketch Understanding

A key component of any evaluation of how well current vision algorithms emulate human visual abstraction is measurement of human behavior in tasks relying on visual abstraction. Here we focus on characterizing what meanings humans extract from the collected sketches, providing the basis for our subsequent empirical evaluation of how well any state-of-the-art vision model approximates human response patterns when presented with the same sketches.

Participants. 1,709 participants (776 male; M_{age} = 39.2 years) were recruited from Prolific and compensated $\$15.50$/hour for their participation. Data from 21 of these sessions were excluded from subsequent analyses due to technical issues. Our predefined criterion for stopping data collection was acquisition of at least 12 recognition judgments for each sketch.

Sketch recognition task. In each session, participants provided labels for 64 sketches randomly sampled from a fixed set of 8,192 sketches, approximately 10% of the full human sketch dataset. The specific set of 8,192 sketches included in this experiment was determined by randomly sampling one sketch cued by each object instance from each drawing-time condition (i.e., 16 instances/concept $\times$ 128 categories $\times$ 4 drawing-time conditions = 8,192). On each trial, participants were presented with a single sketch (300px $\times$ 300px) and a text field where they could provide their best guess concerning the concept the sketch was intended to convey. As soon as they began typing, a drop-down menu appeared with suggested word completions. This drop-down menu contained the entire set of 1,854 labels in the THINGS dataset and only responses that matched one of these 1,854 labels were accepted. Because many words have multiple meanings, the labels contained in this dropdown menu were also accompanied by disambiguating text (e.g., to distinguish *mouse (animal)* from *mouse (computer)*). If participants were unsure of which label best applied to the sketch, they were encouraged to provide additional guesses (up to 5 per sketch). Collecting multiple labels on each drawing trial was important because it enabled us to more thoroughly sample the distribution of meanings that each sketch evoked for human participants (i.e., which labels came to mind and how often they did so). At the beginning of the session, participants completed a practice trial (that we did not include in analyses) to familiarize themselves with the labeling interface.

2.3 Machine Sketch Understanding

We propose a generic protocol for evaluating machine sketch understanding that can be applied to any vision algorithm using our sketch dataset. In this paper, we conduct evaluations of a wide range of state-of-the-art vision models with the goal of demonstrating the feasibility of our protocol and guiding future model development.

Model Suite. Specifically, we evaluated 17 vision models spanning a wide range of architectures and training methods (Table 2), all of which have been demonstrated to achieve high performance on object recognition on standard datasets, such as ImageNet [17]. We also made sure to include variants of standard ConvNet and Transformer models that have gained traction within the field of computational cognitive neuroscience for their potential to close the gap between biological and artificial vision [46, 25, 49, 65].

Model	Architecture	Training Paradigm	Dataset
VGG-19 [78]	VGG-19	supervised	ImageNet
Inception-V3 [84]	Inception-V3	supervised	ImageNet
ResNet-50 [36]	ResNet-50	supervised	ImageNet
ViT-B [20]	ViT-B	supervised	ImageNet
Swin-B [58]	Swin-B	supervised	ImageNet
MLPMixer-B [87]	MLPMixer-B	supervised	ImageNet
CORnet-S [49]	CORnet-S	supervised	ImageNet
Harmonization [25]	ViT-B	supervised	ImageNet + Human Feature Importance [25]
ECOSET [65]	ResNet-50	supervised	ECOSET [65]
SimCLR [14]	ResNet-50	self-supervised	ImageNet
MoCo-v3 [15]	ViT-B	self-supervised	ImageNet
DINO [12]	ViT-B	self-supervised	ImageNet
MAE [35]	ViT-B	self-supervised	ImageNet
CLIP [70]	ViT-B	self-supervised	WebImageText [70]
IPCL [46]	AlexNet	self-supervised	ImageNet
Noisy Student [98]	EfficientNet-b4	semi-supervised	ImageNet + JFT [83]
SWSL [101]	ResNet-50	semi-supervised	ImageNet + YFCC-100M [86] + IG-1B-Targeted [101]

Table 2: Model suite annotated by backbone architecture, training paradigm, and training dataset.

Evaluation Protocol. The goal of our evaluation protocol was to measure how well any of these vision models approximated human sketch recognition behavior when presented with the same sketches. Because these models contain different latent representations of widely varying dimensionalities, we measured machine sketch-recognition behavior by extracting activation patterns from each model's final convolutional or attention block and training linear classification-based readouts on these activation patterns. That is, for each model we independently fit 1,854-way logistic regression classifiers using 5-fold stratified cross-validation to predict the "ground-truth" concept label associated with each sketch. [1]

These predicted labels were aggregated across the 16 sketches of the same concept (e.g., *lion*) and from the same drawing-time condition (e.g., 4 seconds) to yield a model's response distribution for that *type* of sketch. The top-1 classification accuracy was determined by computing the relative frequency of the "ground-truth" concept label in this response distribution. We also computed the Shannon entropy of this response distribution to estimate the degree of semantic ambiguity exhibited by this type of sketch. Further, we derived a measure of the degree to which even the *non*-ground-truth labels generated by each model were semantically related to the ground-truth label, which we term the *semantic neighbor preference* score. This semantic neighbor preference score falls in the range $[0, 1]$ and is highest when labels that are more semantically related to the ground-truth label appear more frequently than more semantically distant labels, is close to 0.5 when labels appear with uniform probability, and is minimized when labels that are more semantically distant appear most frequently.

To compare model classification outputs with human responses, we derived an analogous response distribution from the human labels obtained in the human sketch recognition experiment. That is, we aggregated all labels assigned by human participants to all 16 sketches of the same concept (e.g., *lion*) and from the same drawing-time condition (e.g., 4 seconds) to construct a response distribution for each type of sketch. We then computed the same three metrics above (i.e., top-1 classification accuracy, response entropy, semantic neighbor preference) using the human response distributions.

[1] Because these classification-based readouts were only trained on sketches of 128 object concepts subsetted from the `THINGS` dataset, the probabilities assigned to the remaining 1,726 labels were set to zero during training. As such, none of these models produced any of these other labels at test time, but humans in the sketch recognition experiment *could* select these other labels. This difference between how sketch recognition behavior was elicited from humans and models led us to focus on relative measures of performance when evaluating human-model alignment.

time	accuracy$_{mean}$	accuracy$_{sem}$	entropy$_{mean}$	entropy$_{sem}$	SNP$_{mean}$	SNP$_{sem}$
4 seconds	.031	.003	1.958	.011	.628	.008
8 seconds	.082	.004	1.814	.013	.701	.009
16 seconds	.139	.006	1.690	.014	.750	.010
32 seconds	.199	.007	1.555	.015	.787	.010

Table 3: Human sketch understanding under each draw duration constraint. Columns represent means and standard errors of the mean for top-1 accuracy, response entropy, and semantic neighbour preference (SNP).

2.4 Machine Sketch Production

To explore the potential of models that emulate human visual abstraction in generative tasks, we also include evaluations of `CLIPasso`, a recently developed sketch generation algorithm [91] capable of generating sketches that vary in sparsity.

Generating machine sketches. Specifically, we leveraged `CLIPasso` to generate 8,192 sketches conditioned on the same 2,048 object instances we used in the human sketch production experiment such that each sketch was constrained to consist of either 4, 8, 16, or 32 pen strokes (Fig. 2, *right*). `CLIPasso` generates sketches by optimizing the parameters of a set of curves (i.e., start/end points; control points), each representing a single pen stroke, to be similar to target image. This optimization is guided by a pretrained implementation of CLIP [70], a large model trained using contrastive learning on vast quantities of text-image pairs. Similarity to the target image is defined based on the distance between CLIP's embedding of the target image and its embedding of the sketch, where these embeddings reflect combinations of feature activations from multiple intermediate layers of CLIP.

Measuring human understanding of machine sketches. We evaluated human recognition performance on these `CLIPasso`-generated sketches by recruiting 1,481 participants (730 male; M_{age} = 41.05 years) on Prolific to complete the same sketch recognition task described earlier. Data from 7 of these sessions were excluded from subsequent analyses due to technical issues.

3 Results

Humans produce sparser sketches under stronger time constraints. We first sought to validate the effect of manipulating the maximum time that human participants had to draw on how detailed their sketches were. We estimated how detailed a sketch was by counting the number of strokes it contained (Fig. 3) and then fit a mixed-effects linear regression model predicting the number of strokes as a function of drawing-time condition (i.e., 4s, 8s, 16s, 32s), with random intercepts for object concept. We found that drawings produced under the 4s limit contained the fewest strokes on average, whereas those produced under the 32s limit contained the greatest number of strokes ($\beta = .29$, $SE = 4.95 \times 10^{-3}$, $p < .001$). These results confirm that restricting the amount of time human participants had to produce their sketches led to systematic differences in how detailed their sketches were.

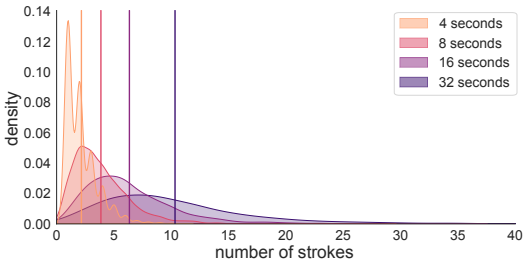

Figure 3: Distributions of number of strokes for each drawing time condition in the human sketch production task. Vertical lines indicate means.

Sparser sketches are more semantically ambiguous for models and humans. Having verified that we had successfully manipulated the level of detail in human sketches, we next sought to evaluate how well current vision models extract semantic information from them at each level of detail. Our general approach was to fit mixed-effect linear regression models to estimate the effect of sparsity on 3 key metrics: (1) top-1 classification accuracy, (2) entropy of response distributions, and (3) semantic neighbor preference. Regression models included random intercepts and slopes for object concepts. Figure 4 shows the performance of each vision model with respect to these metrics for sketches produced under different time constraints.

We found that models generally achieved higher top-1 classification accuracy for more detailed sketches than sparser ones ($\beta = 9.57 \times 10^{-2}$, $t = 20.25$, $p < .001$).

We further found the entropy of the models' response distribution was lower for detailed sketches than for sparser sketches ($\beta = -.03$, $t = -30.19$, $p < .001$), suggesting greater uncertainty about the best label to apply to sparser sketches. Even when sketches were more ambiguous, however, models generated labels that were semantically related to the ground-truth label, as measured by our *semantic neighbor preference* score, with more detailed sketches eliciting a greater proportion of semantically related labels ($\beta = 2.5 \times 10^{-2}$, $t = 8.43$, $p < .001$). These patterns were mirrored in human sketch recognition behavior, with more detailed sketches being associated with higher top-1 classification performance ($\beta = 6.08 \times 10^{-2}, t = 9.57, p < .001$), a tighter distribution of responses (lower entropy) ($\beta = -.14, t = -12.29, p < .001$), and greater semantic neighbor preference ($\beta = 5.41 \times 10^{-2}$, $t = 20.24, p < .001$).

Different models display distinct patterns of sketch recognition behavior. Although all vision models were sensitive to the effect of our drawing-time manipulation, we found that there were reliable differences in classification accuracy between models ($\chi^2(16) = 3455.3$, $p < .001$). Moreover, some models generated a greater diversity of responses than others, as measured by the entropy of their response distribution (Fig. 4B, $\chi^2(16) = 89698$, $p < .001$). Finally, models varied in the degree to which they generated non-ground-truth labels that were semantically related to the ground-truth label ($\chi^2(16) = 318.46$, $p < .001$). Taken together, these results indicate that these models, all high-performing, display systematic differences in how they extract semantic information from sketches.

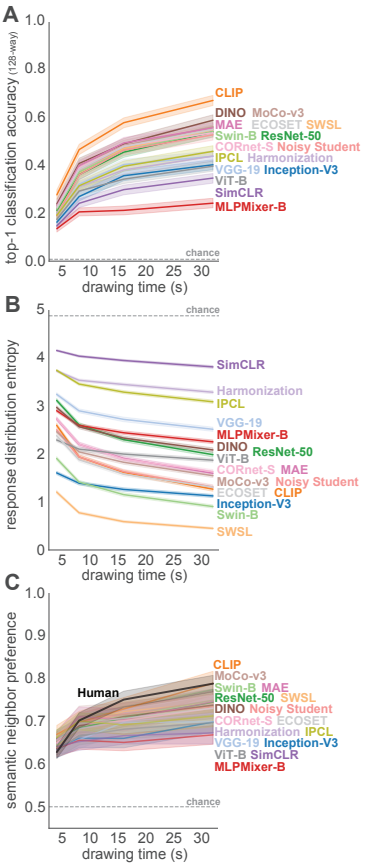

Figure 4: Effect of drawing time constraints on sketch understanding in different vision models.

A large gap remains between human and model sketch understanding. While both humans and models are affected by the amount of detail in sketches, it is not yet clear to what degree their response patterns are well aligned. We evaluated human-model alignment scores using the same three metrics (i.e., top-1 classification accuracy, entropy, semantic neighbor preference) by estimating the degree to which model performance on different *types* of sketches (e.g., lions drawn in 4 seconds or less) covaried systematically with human performance on the same *types* of sketches. For example, a model is considered well aligned with humans with respect to recognition performance if it achieves high top-1 classification accuracy on the same types of sketches that humans succeed in classifying *and* if it achieves low accuracy on the types of sketches that humans fail to classify. Similarly, a model is considered well aligned with humans with respect to semantic ambiguity if it produces a response distribution with high entropy for the same types of sketches that humans are also highly uncertain about *and* if it produces a low-entropy response distribution for the types of sketches that humans systematically agree on (regardless of whether this agreement is concentrated on the correct label). We found that models generaly displayed some degree of alignment to humans on top-1 classification accuracy ($\beta = 7.70 \times 10^{-2}, t = 6.77, p < .001$), response entropy ($\beta = 1.17 \times 10^{-1}$, $t = 28.52$, $p < .001$), and semantic neighbor preference ($\beta = 6.66 \times 10^{-2}$, $t = 6.21$, $p < .001$). Moreover, we found that different vision models aligned with humans to varying degrees (top-1 classification accuracy: $\chi^2(16) = 134.81, p < .001$; response entropy: $\chi^2(16) = 725.78, p < .001$; semantic neighbor preference: $\chi^2(16) = 5.05$, $p = .99$). Nevertheless, a sizable gap remains between the most aligned models and a human-human consistency baseline for *all* metrics (Fig 5; top-1 classification accuracy: $t = 952.19$, $p < .001$; response entropy: $t = 184.21$, $p < .001$; semantic neighbor preference: $t = 389.56$, $p < .001$). Finally, we observe that while examining classification accuracy and entropy yield similar rankings over which models are best aligned to humans, these metrics appear to capture non-redundant sources of information about human and model sketch understanding (Fig 5D).

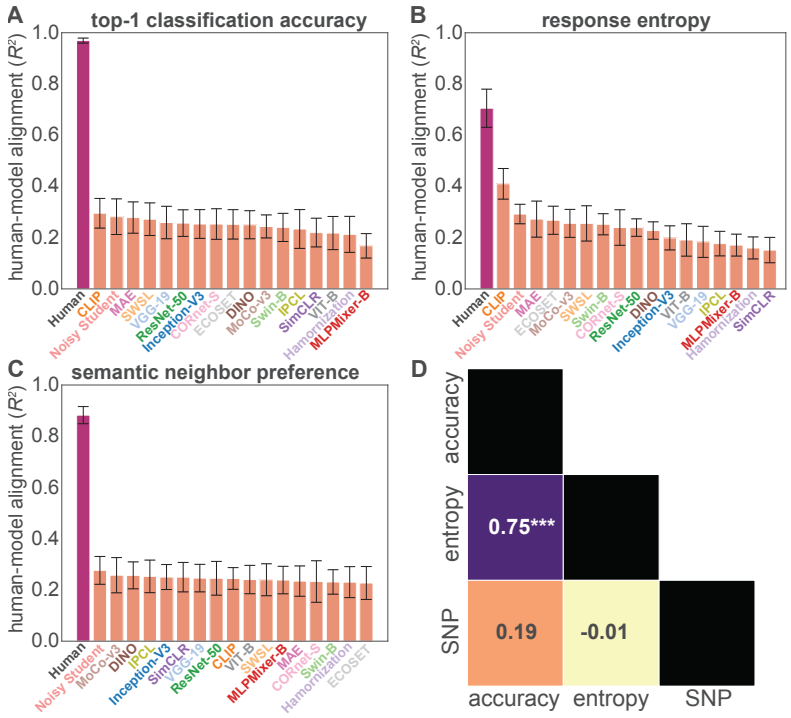

Figure 5: Human-model alignment on (A) top-1 classification accuracy, (B) response entropy, and (C) semantic neighbour preference. Leftmost red bars in each plot correspond to baseline human-human consistency on each metric. Error bars indicate bootstrapped 95% confidence intervals. (D) Spearman ρ correlations between the rank-ordering of vision models with respect to their alignment to human performance on each metric.

A CLIP-based sketch generation algorithm emulates human sketches under some conditions.

While sketch understanding is a critical aspect of visual abstraction, the ability to *produce* sketches spanning different levels of abstraction is no less important. Although there remains a gap between human and model sketch understanding, a CLIP-based vision model [70] was among the most performant and best aligned to human sketch understanding. Insofar as the major bottleneck to being able to generate more human-like sketches is achieving more human-like understanding of sketches and other images [24], a generative model leveraging CLIP's latent representation may be a promising approach. Consistent with this possibility, we found that `CLIPasso` generated sketches of concepts at each abstraction level were about as recognizable as the human-generated sketches of those same concepts at the same abstraction level (Fig. 6A; $adjusted\ R^2 = .64$). Moreover, we found that the more detailed `CLIPasso` sketches were especially human-like in that they evoked a similar set of meanings to their human-generated counterparts.

To measure the degree to which human and `CLIPasso` sketches converged with respect to the object labels that they elicited from human viewers, we computed the Jensen-Shannon distance (JSD) between the label distributions of human and `CLIPasso` sketches for each concept at each of the 4 levels of abstraction. In Fig. 6 B. we show the average label divergence at each level of abstraction. We found that humans and `CLIPasso` sketches were least divergent in terms of their perceived meaning when they were depicted in greater detail or were *less abstract* ($\beta = -0.69$, $t = -6.86$, $p < .001$). At lower levels of detail and visual fidelity, human and `CLIPasso` sketches elicited more diverging responses. Thus, while `CLIPasso` more closely approximates human sketch production behavior at greater levels of detail, there remains a large gap between how `CLIPasso` and human participants attempted to preserve sketch meaning under more severe production constraints. Taken together, while `CLIPasso` marks significant progress towards human-like sketch production its ability to produce highly abstract sketches in a human-like manner remains limited.

4 Conclusion

Recent advances in machine vision have enabled new opportunities to understand the computational mechanisms that underlie human-like visual abstraction—how humans create and interpret a wide variety of images (from detailed illustrations to schematic diagrams) to convey what they perceive and know about the world. Here we introduce SEVA, a new dataset and benchmark containing approximately 90K human and machine generated sketches spanning multiple levels of abstraction, to evaluate progress towards alignment between human and machine visual abstraction. Critically, our model evaluation protocol is focused not just on performance but on how well these models approximate *human-like* sketch production and understanding. Initial benchmarking of a set of 17 vision models on SEVA following this protocol revealed that even the most performant and well-aligned models deviate from human behavior in systematic ways. We also find that current generative models of sketching are able to sketch in human-like ways, but only in limited settings.

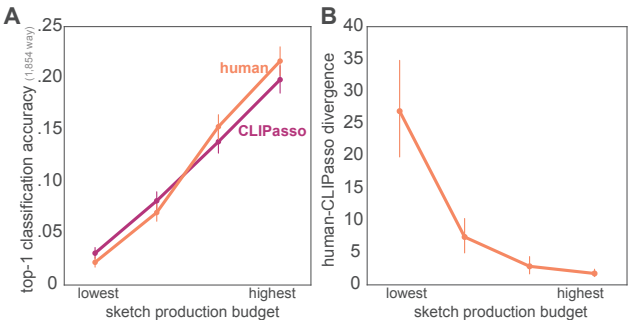

Figure 6: (A) Human top-1 recognition performance on CLIPasso and human sketches at different production budgets (draw duration for human sketches and number of strokes for CLIPasso sketches). (B) Divergence in human-responses to the same type of human and CLIPasso sketches as a function of sketch production budget.

We hope that SEVA will accelerate progress towards unified computational theories that explain how humans are capable of generating and understanding such a wide variety of abstract visual representations. Specifically, our dataset could be used to adjudicate between vision models that are high-performing on tasks involving natural images [64, 29] by characterizing their alignment with how humans understand images generated in a very different manner, including freehand sketches. In addition, SEVA is distinctive among existing human sketch datasets in that it contains detailed measurements of how humans make moment-to-moment decisions about where to place each stroke, when aiming to produce a sketch of a visible object, under different constraints. As such, we expect SEVA to be a key resource for advancing the state-of-the-art in sketch generation [69, 94, 91, 33, 72], and thus lead to computational models of generalized visual abstraction.

5 Limitations

We note several limitations of the current study that would be important to address in future work. First, our sample of participants was limited to English-speaking individuals based in the United States. As such, the current study cannot speak to potential differences in sketch-production behavior across geographical and cultural contexts. However, future work that recruits from a broader cross-section of individuals, including those located in the "Majority World," will be vital for understanding those potential sources of variation in human sketch production and comprehension. Second, we recruited participants via Prolific, a widely used crowdsourcing platform in human behavioral research, without regard to any previous artistic training they had received. As such, our study generally reflects sketch production behavior among individuals without substantial expertise in the visual arts. Third, following prior work [24, 23, 34, 105, 43, 59], the human sketches in our dataset were obtained using a web-based digital drawing interface, where most participants used a mouse or trackpad to produce their sketches (89.96%) and some participants used a touchscreen (6.94%) or a stylus (1.67%). Investigation of the impact of input device on sketch production would be a fruitful avenue for follow-up work. Fourth, because we did not obtain non-digital sketches (e.g., drawn on paper with a pen/pencil), our data cannot speak to differences between digital and non-digital sketches. Fifth, our study of machine sketch production included just one model, *CLIPasso*, limiting the conclusions that can be drawn about sketch generation algorithms in general. Future work that evaluates a broader suite of algorithms would thus be valuable.

Acknowledgements

We thank members of the Cognitive Tools Lab at Stanford and the Stanford Neuro AI Lab for helpful feedback, discussions, and support. This work was supported by an NSF CAREER Award #2047191 to J.E.F.. J.E.F is additionally supported by an ONR Science of Autonomy award.

All code and materials available at:
`https://github.com/cogtoolslab/visual_abstractions_benchmarking_public2023/`

Broader Impacts

Visual abstraction is as ubiquitous as it is central to our understanding of the visual world. Humans can recognize depictions of objects at varying levels of fidelity to their real-world counterparts. This ability to recognize and render concepts in abstract forms is crucial for visual knowledge transmission in the form of graphs, diagrams, and symbols. While computer vision has steadily made progress towards algorithms that can recognize objects in scenes, distinguish among different instances of an object, or even answer questions about those objects in natural language, it remains unclear if these systems have the capability to understand visual concepts at multiple levels of abstraction in a human-like manner. `SEVA` marks a step towards evaluating current vision algorithms on their sensitivity to depictions of a wide variety of common object concepts at varying levels of abstraction. Critically, we also provide results on *human* performance on sketch understanding for sketches of varying sparsity, setting a clear benchmark for future vision algorithms. Our initial study of state-of-the-art vision algorithms shows that even the most high-performing of models still fall short of human consistency baselines. With these vision algorithms increasingly being deployed in many aspects of everyday life, it is crucial to measure how human-like they are in this most natural of human abilities. We hope that `SEVA` will be used by the community to help close the gaps in current vision algorithms' alignment to human behavior that we have identified in this paper.

References

[1] Maxime Aubert, Adam Brumm, Muhammad Ramli, Thomas Sutikna, E Wahyu Saptomo, Budianto Hakim, Michael J Morwood, Gerrit D van den Bergh, Leslie Kinsley, and Anthony Dosseto. Pleistocene cave art from sulawesi, indonesia. *Nature*, 514(7521):223–227, 2014.

[2] Ayu Marliawaty I Gusti Bagus, Tiago Marques, Sachi Sanghavi, James J DiCarlo, and Martin Schrimpf. Primate inferotemporal cortex neurons generalize better to novel image distributions than analogous deep neural networks units. In *SVRHM 2022 Workshop@ NeurIPS*.

[3] Nicholas Baker and Philip J Kellman. Abstract shape representation in human visual perception. *Journal of Experimental Psychology: General*, 147(9):1295, 2018.

[4] Pedro Ballester and Ricardo Araujo. On the performance of googlenet and alexnet applied to sketches. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[5] MD Barrett and PH Light. Symbolism and intellectual realism in children's drawings. *British Journal of Educational Psychology*, 46(2):198–202, 1976.

[6] Ruairidh M Battleday, Joshua C Peterson, and Thomas L Griffiths. Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature communications*, 11(1):5418, 2020.

[7] Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021.

[8] Itamar Berger, Ariel Shamir, Moshe Mahler, Elizabeth Carter, and Jessica Hodgins. Style and abstraction in portrait sketching. *ACM Transactions on Graphics (TOG)*, 32(4):1–12, 2013.

[9] Ayan Kumar Bhunia, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9779–9788, 2020.

[10] Sasha Bozeat, Matthew A Lambon Ralph, Kim S Graham, Karalyn Patterson, Helen Wilkin, Josephin Rowland, Timothy T Rogers, and John R Hodges. A duck with four legs: Investigating the structure of conceptual knowledge using picture drawing in semantic dementia. *Cognitive neuropsychology*, 20(1):27–47, 2003.

[11] Mackinlay Card. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[13] Min Chen, Helwig Hauser, Penny Rheingans, and Gerik Scheuermann. *Foundations of data visualization*. Springer, 2020.

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[15] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[16] Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. What can human sketches do for object detection? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15083–15094, 2023.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[18] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2179–2188, 2019.

[19] Moira R Dillon. Rooms without walls: Young children draw objects but not layouts. *Journal of Experimental Psychology: General*, 150(6):1071, 2021.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[21] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Transactions on graphics (TOG)*, 31(4):1–10, 2012.

[22] Mathias Eitz, Ronald Richter, Tamy Boubekeur, Kristian Hildebrand, and Marc Alexa. Sketch-based shape retrieval. *ACM Transactions on graphics (TOG)*, 31(4):1–10, 2012.

[23] Judith E Fan, Robert D Hawkins, Mike Wu, and Noah D Goodman. Pragmatic inference and visual abstraction enable contextual flexibility during visual communication. *Computational Brain & Behavior*, 3:86–101, 2020.

[24] Judith E Fan, Daniel LK Yamins, and Nicholas B Turk-Browne. Common object representations for visual production and recognition. *Cognitive science*, 42(8):2670–2698, 2018.

[25] Thomas Fel, Ivan Felipe, Drew Linsley, and Thomas Serre. Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[26] Chengying Gao, Qi Liu, Qi Xu, Limin Wang, Jianzhuang Liu, and Changqing Zou. Sketchycoco: Image generation from freehand scene sketches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5174–5183, 2020.

[27] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31, 2018.

[28] Dedre Gentner and Christian Hoyos. Analogy and abstraction. *Topics in cognitive science*, 9(3):672–693, 2017.

[29] Tal Golan, Prashant C Raju, and Nikolaus Kriegeskorte. Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, 117(47):29330–29337, 2020.

[30] Robert L Goldstone, Alan Kersten, and Paulo F Carvalho. Concepts and categorization. 2013.

[31] Ernst Hans Gombrich. *The story of art*, volume 12. Phaidon London, 1995.

[32] Hyowon Gweon, Joshua B Tenenbaum, and Laura E Schulz. Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, 107(20):9066–9071, 2010.

[33] David Ha and Douglas Eck. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*, 2017.

[34] Robert D Hawkins, Megumi Sano, Noah D Goodman, and Judith E Fan. Visual resemblance and interaction history jointly constrain pictorial meaning. *Nature Communications*, 14(1):2199, 2023.

[35] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[36] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[37] Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12:e82580, 2023.

[38] Martin N Hebart, Adam H Dickter, Alexis Kidder, Wan Y Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I Baker. Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, 14(10):e0223792, 2019.

[39] Mary Hegarty. The cognitive science of visual-spatial displays: Implications for design. *Topics in cognitive science*, 3(3):446–474, 2011.

[40] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[41] D. L. Hoffmann, C. D. Standish, M. García-Diez, P. B. Pettitt, J. A. Milton, J. Zilhão, J. J. Alcolea-González, P. Cantalejo-Duarte, H. Collado, R. de Balbín, M. Lorblanchet, J. Ramos-Muñoz, G.-Ch. Weniger, and A. W. G. Pike. U-th dating of carbonate crusts reveals neandertal origin of iberian cave art. *Science*, 359(6378):912–915, 2018.

[42] Holly Huey and Bria Long. Developmental changes in the semantic part structure of drawn objects. In *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*, 2022.

[43] J Jongejan, H Rowley, T Kawashima, J Kim, and N Fox-Gieg. The quick, draw! dataset, 2017.

[44] Annette Karmiloff-Smith. Constraints on representational change: Evidence from children's drawing. *Cognition*, 34(1):57–83, 1990.

[45] Charles Kemp, Andrew Perfors, and Joshua B Tenenbaum. Learning overhypotheses with hierarchical bayesian models. *Developmental science*, 10(3):307–321, 2007.

[46] Talia Konkle and George A Alvarez. Beyond category-supervision: instance-level contrastive learning models predict human visual system responses to objects. *bioRxiv*, pages 2021–05, 2021.

[47] Talia Konkle and George A Alvarez. A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, 13(1):1–12, 2022.

[48] Nikolaus Kriegeskorte. Deep neural networks: a new framework for modelling biological vision and brain information processing. *biorxiv*, page 029876, 2015.

[49] Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. Brain-like object recognition with high-performing shallow recurrent anns. *Advances in neural information processing systems*, 32, 2019.

[50] Brenden M Lake and Steven T Piantadosi. People infer recursive visual concepts from just a few examples. *Computational Brain & Behavior*, 3:54–65, 2020.

[51] Bo Li, Yijuan Lu, Afzal Godil, Tobias Schreck, Benjamin Bustos, Alfredo Ferreira, Takahiko Furuya, Manuel J Fonseca, Henry Johan, Takahiro Matsuda, et al. A comparison of methods for sketch-based 3d shape retrieval. *Computer Vision and Image Understanding*, 119:57–80, 2014.

[52] Bo Li, Yijuan Lu, Chunyuan Li, Afzal Godil, Tobias Schreck, Masaki Aono, Martin Burtscher, Hongbo Fu, Takahiko Furuya, Henry Johan, et al. Shrec'14 track: Extended large scale sketch-based 3d shape retrieval. In *Eurographics workshop on 3D object retrieval*, volume 2014, pages 121–130, 2014.

[53] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[54] Ke Li, Kaiyue Pang, Jifei Song, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Honggang Zhang. Universal sketch perceptual grouping. In *Proceedings of the european conference on computer vision (ECCV)*, pages 582–597, 2018.

[55] Xiao Li, Ziqi Wang, Bo Zhang, Fuchun Sun, and Xiaolin Hu. Recognizing object by components with human prior knowledge enhances adversarial robustness of deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[56] Gao Lin, Liu Feng-Lin, Chen Shu-Yu, Jiang Kaiwen, Li Chunpeng, Yukun Lai, and Fu Hongbo. Sketchfacenerf: Sketch-based facial generation and editing in neural radiance fields. *ACM Transactions on Graphics*, 2023.

[57] Fang Liu, Changqing Zou, Xiaoming Deng, Ran Zuo, Yu-Kun Lai, Cuixia Ma, Yong-Jin Liu, and Hongan Wang. Scenesketcher: Fine-grained image retrieval with scene sketches. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 718–734. Springer, 2020.

[58] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[59] Bria Long, Judith Fan, Zixian Chai, and Michael C Frank. Parallel developmental changes in children's drawing and recognition of visual concepts. 2021.

[60] Xuanchen Lu, Xiaolong Wang, and Judith E Fan. Learning dense correspondences between photos and sketches. 2023.

[61] Yongyi Lu, Shangzhe Wu, Yu-Wing Tai, and Chi-Keung Tang. Image generation from sketch constraint using contextual gan. In *Proceedings of the European conference on computer vision (ECCV)*, pages 205–220, 2018.

[62] Georges-Henri Luquet. Le dessin enfantin.(bibliothèque de psychologie de l" enfant et de pédagogie.). 1927.

[63] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[64] Nesma Mahmoud, Hanna Antson, Jaesik Choi, Osamu Shimmi, and Kallol Roy. Stress and adaptation: Applying anna karenina principle in deep learning for image classification. *arXiv preprint arXiv:2302.11380*, 2023.

[65] Johannes Mehrer, Courtney J Spoerer, Emer C Jones, Nikolaus Kriegeskorte, and Tim C Kietzmann. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8):e2011417118, 2021.

[66] Gregory Murphy. *The big book of concepts*. MIT press, 2004.

[67] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.

[68] Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9617–9626, 2019.

[69] Shuwen Qiu, Sirui Xie, Lifeng Fan, Tao Gao, Jungseock Joo, Song-Chun Zhu, and Yixin Zhu. Emergent graphical conventions in a visual communication game. *Advances in Neural Information Processing Systems*, 35:13119–13131, 2022.

[70] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[71] Rishi Rajalingham, Elias B Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J DiCarlo. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, 38(33):7255–7269, 2018.

[72] Leo Sampaio Ferraz Ribeiro, Tu Bui, John Collomosse, and Moacir Ponti. Sketchformer: Transformer-based representation for sketched structure. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14153–14162, 2020.

[73] Timothy T Rogers and Karalyn Patterson. Object categorization: reversals and explanations of the basic-level advantage. *Journal of Experimental Psychology: General*, 136(3):451, 2007.

[74] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016.

[75] Ravi Kiran Sarvadevabhatla, Sudharshan Suresh, and R Venkatesh Babu. Object category understanding via eye fixations on freehand sketches. *IEEE Transactions on Image Processing*, 26(5):2508–2518, 2017.

[76] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[77] Patrick Shafto, Noah D Goodman, and Thomas L Griffiths. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, 71:55–89, 2014.

[78] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[79] Johannes JD Singer, Katja Seeliger, Tim C Kietzmann, and Martin N Hebart. From photos to sketches-how humans and deep neural networks process objects across different levels of visual abstraction. *Journal of vision*, 22(2):4–4, 2022.

[80] Linda B Smith, Susan S Jones, Barbara Landau, Lisa Gershkoff-Stowe, and Larissa Samuelson. Object name learning provides on-the-job training for attention. *Psychological science*, 13(1):13–19, 2002.

[81] Joan G Snodgrass and Mary Vanderwart. A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology: Human learning and memory*, 6(2):174, 1980.

[82] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *Proceedings of the IEEE international conference on computer vision*, pages 5551–5560, 2017.

[83] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[84] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[85] Michael Henry Tessler and Noah D Goodman. The language of generalization. *Psychological review*, 126(3):395, 2019.

[86] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[87] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

[88] Aditay Tripathi, Rajath R Dani, Anand Mishra, and Anirban Chakraborty. Sketch-guided object localization in natural images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 532–547. Springer, 2020.

[89] Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, USA, 1986.

[90] Barbara Tversky. Some ways that maps and diagrams communicate. In *Spatial cognition II: Integrating abstract theories, empirical studies, formal methods, and practical applications*, pages 72–79. Springer, 2000.

[91] Yael Vinker, Ehsan Pajouheshgar, Jessica Y Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. Clipasso: Semantically-aware object sketching. *arXiv preprint arXiv:2202.05822*, 2022.

[92] Ivan Viola and Tobias Isenberg. Pondering the concept of abstraction in (illustrative) visualization. *IEEE transactions on visualization and computer graphics*, 24(9):2573–2588, 2017.

[93] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.

[94] Alexander Wang, Mengye Ren, and Richard Zemel. Sketchembednet: Learning novel concepts by imitating drawings. In *International Conference on Machine Learning*, pages 10870–10881. PMLR, 2021.

[95] Fei Wang, Shujin Lin, Hefeng Wu, Hanhui Li, Ruomei Wang, Xiaonan Luo, and Xiangjian He. Spfusionnet: Sketch segmentation using multi-modal data fusion. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1654–1659. IEEE, 2019.

[96] Jiayun Wang, Jierui Lin, Qian Yu, Runtao Liu, Yubei Chen, and Stella X Yu. 3d shape reconstruction from free-hand sketches. In *European Conference on Computer Vision*, pages 184–202. Springer, 2022.

[97] Qiang Wang, Haoge Deng, Yonggang Qi, Da Li, and Yi-Zhe Song. Sketchknitter: Vectorized sketch generation with diffusion models. In *The Eleventh International Conference on Learning Representations*, 2022.

[98] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

[99] Fei Xu and Joshua B Tenenbaum. Word learning as bayesian inference. *Psychological review*, 114(2):245, 2007.

[100] Peng Xu, Timothy M Hospedales, Qiyue Yin, Yi-Zhe Song, Tao Xiang, and Liang Wang. Deep learning for free-hand sketch: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):285–312, 2022.

[101] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019.

[102] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

[103] Justin Yang and Judith E Fan. Visual communication of object concepts at different levels of abstraction. *arXiv preprint arXiv:2106.02775*, 2021.

[104] Lumin Yang, Jiajie Zhuang, Hongbo Fu, Xiangzhi Wei, Kun Zhou, and Youyi Zheng. Sketchgnn: Semantic sketch segmentation with graph neural networks. *ACM Transactions on Graphics (TOG)*, 40(3):1–13, 2021.

[105] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 799–807, 2016.

[106] Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Sketch-a-net: A deep neural network that beats humans. *International journal of computer vision*, 122:411–425, 2017.

[107] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022.

[108] Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, and Xiaochun Cao. Sketchnet: Sketch classification with web images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1105–1113, 2016.

[109] Song-Hai Zhang, Yuan-Chen Guo, and Qing-Wen Gu. Sketch2model: View-aware 3d modeling from single free-hand sketches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6012–6021, 2021.

[110] Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.

[111] Changqing Zou, Qian Yu, Ruofei Du, Haoran Mo, Yi-Zhe Song, Tao Xiang, Chengying Gao, Baoquan Chen, and Hao Zhang. Sketchyscene: Richly-annotated scene sketches. In *Proceedings of the european conference on computer vision (ECCV)*, pages 421–436, 2018.