MODELVSBABY: A DEVELOPMENTALLY MOTIVATED BENCHMARK OF OUT-OF-DISTRIBUTION OBJECT RECOGNITION

A PREPRINT

Saber Sheybani
Department of Informatics
Indiana University Bloomington
Bloomington, IN 47405
sheybani@iu.edu

Sahaj Singh Maini
Department of Computer Science
Indiana University Bloomington
Bloomington, IN 47405
sahmaini@iu.edu

Aravind Dendukuri
Data Science Program
Indiana University Bloomington
Bloomington, IN 47405
ardend@iu.edu

Zoran Tiganj
Department of Computer Science
Indiana University Bloomington
Bloomington, IN 47405
ztiganj@iu.edu

Linda B. Smith
Department of Psychological and Brain Sciences
Indiana University Bloomington
Bloomington, IN 47405
smith4@iu.edu

May 21, 2024

ABSTRACT

Deep neural networks have recently become remarkable computational tools for thinking about human visual learning. Recent studies have explored the effects of altering naturalistic images and compared the responses of both humans and models, providing valuable insights into their functioning and how deep neural networks can shape our understanding of human learning. Critically, much of human visual learning happens throughout early development. Yet, well-controlled benchmarks comparing AI models with young humans are scarce. Here, we present a developmentally motivated benchmark of out-of-distribution (OOD) object recognition. Our benchmark, ModelVsBaby, includes a set of OOD conditions that have long been studied in the vision science literature, and are expected to be sensitive to the development of OOD object-recognition in humans: silhouette, geon, occluded, blurred, crowded background, and a baseline realistic condition. Along with the stimuli, we release a unique dataset of the responses of 2-year-old children to the stimuli. Our preliminary analyses of the dataset show several interesting patterns: 2-year-olds achieve 80% accuracy in the silhouette condition, nearly as well as in the realistic condition (chance=12%). They also perform well above chance, near 60% accuracy, on the other challenging conditions. We also evaluate image-text association (CLIP) models trained on varying amounts of internet-scale datasets. The model performances show that with enough data, all conditions are learnable by artificial learners. However, Realistic and Silhouette are learned with fewer training data similar to humans. Our benchmark stimuli and infant responses, provide an essential steppingstone for building computational models that are aligned with humans both in terms of the learning outcomes as well as the learning trajectory. This endeavor can furnish creating better models of visual development as well as improving the efficiency of AI systems for practical applications. Future work may use the benchmark stimuli to test more age groups, and provide a detailed comparison of models of various flavors in terms of "developmental alignment".

Keywords object recognition · visual development · representation learning

1 Introduction

Humans demonstrate robust object recognition abilities by adulthood. How do their brains achieve this remarkable feat? As an information encoding system, which aspects of the visual input do human brains capture that allow for

such robust category recognition when encountering new object instances? What aspects of the input are retained in the representation, especially in information theoretic terms? Verbal theories in cognitive science point to different aspects of the visual input being captured by the visual system for object recognition. These theories suggest that the visual system stores objects in the form of contours Peterson and Gibson [1994], category-specific features like bills of ducks Ullman [2007], Ullman et al. [2002], simplified geometric 3D shapes with category-specific part-whole relations Biederman [1987], or even raw pictures Tarr [1995]. By adulthood, humans can recognize objects given any of the above information in the input. However, little is known about the nature of visual representations in infants during the development of object recognition abilities (recent work provides evidence for skeletal-based 3D shape recognition in 6-12mo infants Ayzenberg and Lourenco [2022]). Do some forms of recognition appear before others in humans? As a remarkable language of hypotheses for learning Golan et al. [2023], we consider deep neural network models trained on large scale datasets with little to no innate biases. Such models broadly allow asking to what extent fitting to naturalistic inputs with little to no inductive biases can produce biologically relevant learning outcomes. Here we ask if various forms of recognition in models are ordered by the amount of training data / i.e., visual experience in models.

To investigate the above questions and gain more insights into the developmental trajectory of humans and models Huber et al. [2023], we create ModelVsBaby, the first comprehensive developmentally-motivated object recognition benchmark. We include 5 OOD conditions informed by the aforementioned theories of object-recognition in humans: *Silhouettes*, *Geons*, *Features*, *Blurred*, and a baseline *Realistic* condition. Along with the stimuli, we release a unique dataset of the responses of 2-year-old children to the stimuli. In an investigation of our motivating questions, we compare the object recognition abilities of 2 year-olds with the performance of artificial learners given different amounts of data. We first find that all OOD conditions, some previously regarded as requiring innate mechanisms, can be learned from the latent structure of natural images, given a sufficient quantity of visual experience. Second, we show that the Realistic and Silhouettes conditions are learned with less visual experience in both babies and multimodal image-text association models, compared to the other OOD conditions. This suggests that the more abstract forms of representations such as geons are less accessible to learners and are learned by both babies and models as they acquire more visual experience and find deeper patterns in the latent structure of their visual inputs. Finally we introduce ImageNet-Baby8, a dataset for finetuning any image or video-based model, to allow evaluating them on ModelVsBaby. We hope that ModelVsBaby will be utilized by the research community and contribute to advancing the field of computational modeling of visual development.

2 Benchmark Description: Human experiments

Participants : Sixty-eight toddlers (age range: 24-30 months) participated in this study. All children were tested individually in a controlled laboratory setting, accompanied by a trained experimenter who guided them through the test notebooks. The experiment was designed to resemble a naturalistic parent-child play session, as young children often struggle to follow task instructions in the same manner as adults. Parents provided informed consent for their child's participation, and the study was approved by the Institutional Review Board.

Stimuli and Procedure : The experiment consisted of 16 trials, each presented in a separate test notebook. In each trial, children were shown a page containing six distinct objects and were asked to point to a specific target object named by the experimenter (see Figure 1). The experimenter, acting as a caregiver, asked the child, "Look, there is a <object name>. Where is the <object name>?" The child's response was recorded as correct if they pointed to the target object and incorrect if they pointed to the wrong object, moved their finger indecisively on the page, or refused to respond. In some trials, two instances of the correct object were present on the page. In these cases, after the child's initial response, the experimenter asked, "Is there another one?" and recorded the child's additional response. The objects on each page of the notebook were selected from a set of individual object images. This set includes five instances from each of eight basic object categories: dog, donkey, duck, cup, chair, hat, car, and airplane. Additionally, each instance is represented in five forms of object representation, motivated by theories of object recognition in humans: *Realistic, Silhouettes, Geons, Blurred, and Features*. In total, the stimulus set comprises 200 unique individual object images. For each condition, two notebooks were created for use in the six-alternative forced choice task. To construct each notebook (e.g., the silhouettes notebook number 2), each of the 16 trials was filled with six silhouette objects from the individual objects set. In eight out of the 16 trials, two instances from the target category were used, while in the other eight trials, only one instance of the target category was included. To control for potential order effects and to ensure that exposure to one condition did not influence performance in another, each child participated in only one of the experimental conditions. The current dataset consists of 1,724 trials collected from the 68 participating toddlers, with each child contributing data to a single experimental condition.

3 Evaluating Models on ModelVsBaby

The six-alternative forced-choice recognition task can be reproduced for image-computable models in a variety of forms, depending on the interface of the models.

3.1 Qualitative Evaluation: t-SNE plots

Any image embedding model encodes an image in terms of the dimensions of an abstract high-dimensional space. The dimensions of this space are often not amenable to interpretation or visualization. However, a two dimensional projection of the embeddings of a set of stimuli can inform how the model broadly organizes them with respect to each other.

Therefore, the tSNE or MDS embeddings of the last layer of the encoder in response to a select group of stimuli is often used to provide a qualitative account of the embedding space. Figure 4 shows the tSNE projections of 100 randomly selected ModelVsBaby stimuli for a SWSL-IG1B model (retrieved from the PyTorch Image Models package Wightman [2019], under the name $resnext10 - 32x4d.fb - swsl - ig1b - ft - in1k$).

3.2 Quantitative Evaluation: Multimodal image-text models

For image-text models, we use the following procedure:

In each trail, given the text prompt s, and 6 image prompts x_i:

1. compute the text embedding $z(s)$ for the text prompt.
2. compute the image embedding $z(x_i)$ for each image prompt x_i.
3. compute the pairwise similarity between the text prompt and each image $z(s) and z(x_i)$ using cosine similarity.
4. the image with the maximum similarity to the text is used as the model response.

Figure 3 panel B-F shows the result of evaluating a range of CLIP models from the Datacomp repository Foundations [2023]. To better reproduce human decision making, a similarity threshold for the decision may be used in future work.

3.3 Quantitative Evaluation: Unimodal image embedding models

While our human experiments are best reproduced with image-text models, many models in NeuroAI are unimodal, trained only on images Zhuang et al. [2021] or videos Orhan [2023]. There are two steps for evaluating encoder models: 1) extract category decisions from single images for ModelVsBaby stimuli. 2) optionally, one may attempt to simulate the experiment conditions further by taking the competing stimuli into account.

Extracting category decisions The common approach in deep learning for evaluating pretrained encoder models is to finetune a linear classifier on top of their features. With the high dimensionality of the representations of these models (often in the order of 1000 or more), this approach requires the benchmark to include thousands of samples in order to show the advantage of the pretrained feature over a random feature vector of the same size. On the other hand, the time and resource constraints of lab experiments often are not amenable to recording benchmark responses of this scale from human subjects. This issue is even more pronounced when considering young humans and animals as participants, essential to the study of visual development.

As a workaround, Geirhos et al. [2021] suggests establishing a mapping between ImageNet1k categories, commonly known by encoder models, and basic categories commonly used in human object recognition research. This approach is called Direct Mapping of ImageNet categories and can also be used for our benchmark. To further prevent the potential impact of the mapping procedure, we have curated a finetuning set specifically for learning the categories of ModelVsBaby We also note that a leave-one-out cross validation scheme would leave too many shortcuts for small scale benchmarks to be effective.

Taking the competing stimuli into account To apply the force-choice condition of the human experiments, we suggest the following procedure:

For a given trail, with the target category s, and 6 image prompts x_i:

1. extract decision logit vector of the model for each image prompt Y_i.
2. the image with the maximum logit element corresponding to the target category is used as the model decision.

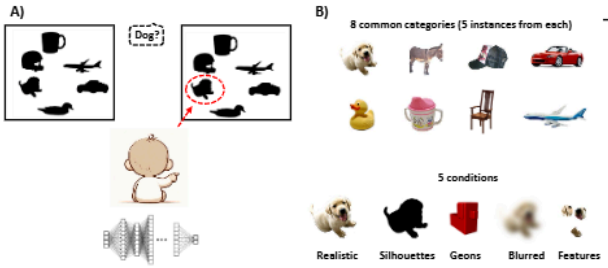

Figure 1: Benchmark task and stimuli. A) babies or image-computable models participate in a six-alternative forced-choice task of matching an object name to an object image. B) The dataset consists of 5 instances of 8 common object categories, each in 5 conditions.

A decision threshold should be considered for the logit value. If all logit vectors are bellow the threshold, the output is considered *undecided*.

3.4 ImageNet-Baby8: a complementary finetuning set

Humans learn from a continuous stream of many different signals from their sensory organs responding to changes in the environment and the body. Throughout learning, the associative relationships within and between the signals are *learned*.

For instance, object name learning includes both learning the patterns within the visual signals (and other sensory signals) that constitute a given object as well as associating the object name (e.g. in the form of heard object name) with its visual representation. Empirical studies on the egocentric experience of babies report that object names are heard extremely rarely in the scope of everyday life Clerkin et al. [2017] no more than a few times. How do babies learn the complex relationship between the raw visual signals of an object and its names? It is hypothesized that experiencing naturalistic views in general allows learning mid-level visual representations. In addition, experiencing many views of only a few instances of each object category allows learning high level features that generalize to other instances of the same category. The mid-level and high level visual features learned – without hearing object names – provide a foundation for rapid word learning, i.e. associating the high level features of each object to its name.

Therefore to computationally investigate few-shot word learning, models may be trained in two stages: 1) General pretraining on naturalistic views (without co-occurring object names). The model will learn how to associate elements of scenes and objects with each other. 2) Fine tuning on a small dataset of object-name pairs. The model learns to associate the key visual features for each object to its name.

To allow evaluating models with various visual diets on ModelVsBaby, we curated a dataset of the ModelVsBaby categories, with 500 train samples and 100 validation samples for each category.

3.4.1 Curating ImageNet-Baby8

For each of the 8 ModelVsBaby categories, we first find the relevant ImageNet categories using the WordNet database Fellbaum [1998]. We then randomly select 1000 images from ImageNet1k to be considered for inclusion in ImageNet-Baby8. However, we find that some ModelVsBaby categories such as *donkey*, *hat* and *duck* are poorly represented in ImageNet. Furthermore, we aim to include toy object images to better represent the visual experience of children. Therefore, we collect an additional set of internet images using the Flickr API Flickr [n.d.]. Informed by the collection procedure of the ImageNet dataset Deng et al. [2009], we perform the following steps:

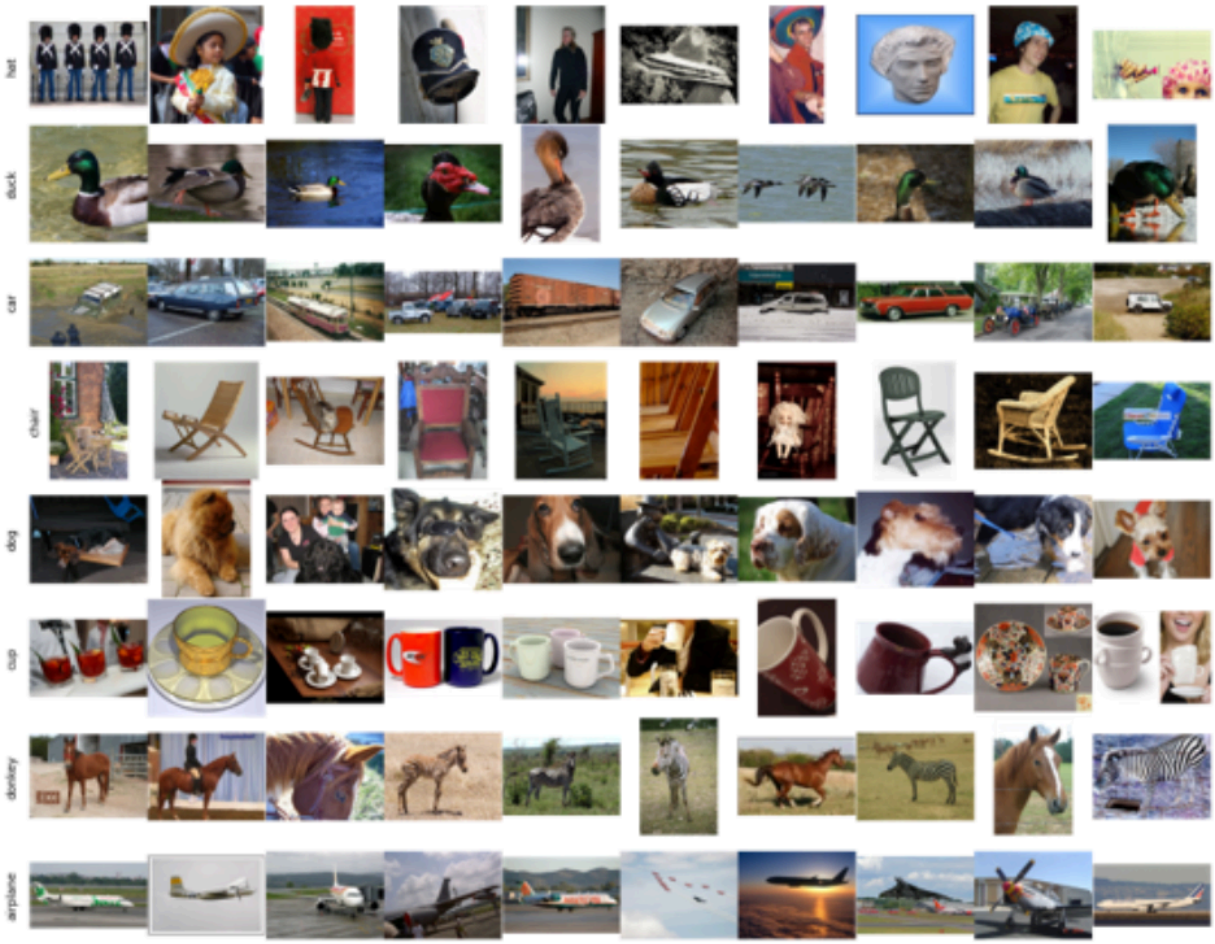

Figure 2: A snapshot of ImageNet-Baby8, a complementary finetuning set for ModelVsBaby.

1. For each category, we download 2000 images with the category name as the search keyword. We filter the results by license to include the Attribution-NonCommerical-Sharealike license.
2. We exclude irrelevant results.
3. We exclude images including human faces (1/2 or more), for privacy concerns.
4. We remove the duplicates using a duplicate detector based on image hashing Buchner [2013].
5. We also include 50 toy images per category using the same procedure.

Overall, we collect 700 samples for each category, with a suggested split of 500-100-100 for train, validation and test.

4 How do babies and image-text models perform

4.1 2-year-olds can reasonably identify object categories based on shape and key features

We collected a dataset of the responses of 24-30 month old children to each of the conditions of the ModelVsBaby benchmark as described in section 2. To find the baby performance, we calculate the categorization accuracy as the percentage of correctly categorized trials (over all trials) for each condition (see figure 3A). In the realistic condition as a baseline, 2yo children correctly identify the object image in around 80% of the trials. In the OOD conditions, the 2yo performance is surprisingly high, with the silhouette condition being nearly the same as the realistic condition and the geons, blurred and features being close to 60% which is well above chance. This shows that the object representation in the young child brain already goes beyond superficial features and makes advantageous use of category-relevant

information existing in contours (silhouettes), low spatial frequencies (blurred), key features (features), and abstract geometric features (geons).

4.2 Image-text models may learn the conditions in the same order as children

We evaluate a series of pretrained CLIP models, retrieved from the OpenClip repository Ilharco et al. [2021] on ModelVsBaby. These models have some differences in size and training recipe, but most importantly Geirhos et al. [2021] have different data diets. We focus on models trained on the following datasets: CommonPool (various sizes and data filtering schemes) Gadre et al. [2024], Datacomp-1BGadre et al. [2024], LAION (2B, 5B)Schuhmann et al. [2022], Conceptual-12mChangpinyo et al. [2021], OpenAI Radford et al. [2021].

The evaluation is performed by computing the CLIP similarity of image-prompt pairs as they appear in the real experiment trials (i.e. zero-shot).

Figure 3B-F shows the object recognition accuracy of the selected models as a function of the size of the training dataset. We see some variability in performance within the models of the same training size. However, when considering the best performing models for each train dataset size (pareto frontier), we see a logarithmic relationship between accuracy and data size as commonly reported in the scaling laws of deep neural networks Hestness et al. [2017], Zhai et al. [2022]. To emphasize this pattern, we fit an exponential curve to the pareto frontier of accuracy and dataset size (shown with a dashed line in 3B-F).

The models trained on the most data perform near perfectly in all task conditions. This is surprising as a condition such as geons is unlikely to be well-represented in the training dataset of these models and confirms the previous observations in the exceptional generalization of large-scale multimodal models Geirhos et al. [2021]. As it pertains to the theories of visual learning in humans and animals, this result suggests that with enough data, all conditions are learnable by a general learning mechanism without inductive biases related to geon or skeletal representations Bowers et al. [2023].

When looking at models with smaller datasets – million-scale models– we observe that similar to children, the performance on the silhouettes condition reaches maximum accuracy with fewer training data. This suggests that contours are more readily learned from naturalistic images (as experienced by both models and humans) in comparison to geons, features and blurry images.

5 How large data models organize the stimuli

How do large data models *see* the ModelVsBaby objects in unusual conditions such as geons? While high-dimensional representation of connectionist models are generally not amenable to human intuition, we may still get a glimpse of the distance between the stimuli in the representation space (as in the last layer of a network). Here we compute the object embeddings of a billion-scale image model (SWSL-IG1B) trained with supervised learning on Instagram-1B Yalniz et al. [2019] and finetuned on ImageNet1k. We then project the 2000-dimensional embeddings into a 2-dimensional plane using the t-SNE method. Figure 4 shows the relative embedding distance between 100 randomly selected objects in ModelVsBaby as encoded by SWSL-IG1B. The bounding box color encodes the true category of the object. While the model retains information about the OOD condition, many different objects of the same category are grouped together.

6 Related Work: Related Benchmarks

In this section, we provide a brief list of related benchmarks from the literature of computer vision and computational neuroscience.

ModelVsHuman. Closest to our work is Geirhos et al. [2021] which collects a large set of OOD conditions and collects adult responses in a lab setting where the participants chose one out of 16 basic categories, after seeing an image for a fraction of a second. They evaluated a large set of models and made remarkable conclusions, most notably the high texture bias (as opposed to a human-like shape bias) in models. Huber et al. [2023], motivated by similar goals to ours, evaluates 4-15 year olds on a subset of ModelVsHuman and shows that older children, have a high shape bias similar to adults, unlike most ANNs.

ToyBox. Motivated by recreating the naturalistic patterns of embodied visual experience, Wang et al. [2018] created an egocentric video dataset called Toybox that contains egocentric (i.e., first-person perspective) videos of common household objects and toys being manually manipulated to undergo structured transformations, such as rotation, translation, and zooming.

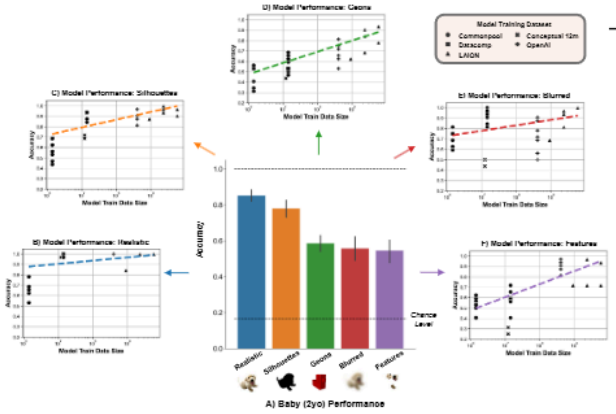

Figure 3: Performance of 2 year old babies and CLIP models. A) Bar heights indicate mean across trials, error bars indicate 95% ci. B-F) The performance of a series of CLIP models trained on various portions of internet datasets. Each point indicates one model. The dashed line indicates the envelope of the best models train dataset size, found by fitting a logarithmic curve to pareto frontier of data size - accuracy

OOD versions of ImageNet for evaluating models. Motivated by evaluating the robustness of computer vision models, many datasets have been developed. Numerous digital distortions such as various kinds of noise, blur, brightness and contrast manipulation, and pixelation. Hendrycks and Dietterich [2019], Hendrycks et al. [2021a] addresses real perturbations such as distribution shifts resulting from image capture process, object occlusion, orientation, zoom, and scale, and non-synthetic blurry images. Motivated by evaluating shape bias (as opposed to texture bias), the authors also create artistic renditions of object classes from the original ImageNet dataset. Finally, Hendrycks et al. [2021b] creates a dataset of natural adversarial examples, collected as the images commonly misclassified by ImageNet models. In other work, Abbas and Deny [2023] create a synthetic dataset of images of objects in unusual orientations, and evaluate the robustness of a collection of 38 recent and competitive deep networks for image classification

7 Conclusion

We present the first developmentally-motivated benchmark of OOD object recognition to allow the evaluation of computational models of visual development and facilitate the advancement of the emerging field of developmental machine learning Smith and Slone [2017]. Our benchmark stimuli and infant responses, provide an essential steppingstone for building competitive computational models of human visual development. We show that physical versions of all hypothesized forms of internal representations are recognizable by young learners and by image-text association models given sufficient data. We also provide evidence that contour-based recognition emerges earlier in development, already comparable to realistic views at 2 years of age. Evaluating image-text association models (CLIP) shows a similar pattern: silhouettes are recognized well with fewer amounts of training data, compared to geons, blurred, and key features. This suggests that contour representations –silhouettes – are readily discoverable in the visual experiences of toddlers and in the training data of the models. Future work may use the benchmark stimuli to test more age groups, more difficult tasks, and provide a detailed comparison of models of various flavors in terms of "developmental alignment"

Figure 4: Low-dimensional projection of the representations of ModelVsBaby stimuli as encoded by a large-data model. The color of the bounding box encodes ground truth object category.

8 Data and Code Availability

The benchmark stimuli, baselines, finetuning set, and a Python interface is publicly available on Open Science Framework (OSF) at `https://osf.io/wbrd4/`.

References

Mary A Peterson and Bradley S Gibson. Object recognition contributions to figure-ground organization: Operations on outlines and subjective contours. *Perception & Psychophysics*, 56(5):551–564, 1994.

Shimon Ullman. Object recognition and segmentation by a fragment-based hierarchy. *Trends in cognitive sciences*, 11 (2):58–64, 2007.

Shimon Ullman, Michel Vidal-Naquet, and Erez Sali. Visual features of intermediate complexity and their use in classification. *Nature neuroscience*, 5(7):682–687, 2002.

Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2): 115, 1987.

Michael J Tarr. Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin & Review*, 2:55–82, 1995.

Vladislav Ayzenberg and Stella Lourenco. Perception of an object's global shape is best described by a model of skeletal structure in human infants. *elife*, 11:e74943, 2022.

Tal Golan, JohnMark Taylor, Heiko Schütt, Benjamin Peters, Rowan P Sommers, Katja Seeliger, Adrien Doerig, Paul Linton, Talia Konkle, Marcel Van Gerven, et al. Deep neural networks are not a single hypothesis but a language for expressing computational hypotheses. *Behavioral and Brain Sciences*, 46, 2023.

Lukas S Huber, Robert Geirhos, and Felix A Wichmann. The developmental trajectory of object recognition robustness: children are like small adults but unlike big deep neural networks. *Journal of vision*, 23(7):4–4, 2023.

Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

ML Foundations. Datacomp: A competition to democratize data. `https://github.com/mlfoundations/datacomp`, 2023.

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.

A Emin Orhan. Scaling may be all you need for achieving human-level object recognition capacity with human-like visual experience. *arXiv preprint arXiv:2308.03712*, 2023.

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021.

Elizabeth M Clerkin, Elizabeth Hart, James M Rehg, Chen Yu, and Linda B Smith. Real-world visual statistics and infants' first-learned object names. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372 (1711):20160055, 2017.

Christiane Fellbaum. *WordNet: An electronic lexical database*. MIT press, 1998.

Flickr. Flickr services. `https://www.flickr.com/services/api/flickr.photos.search.html`, n.d. Accessed: 2023-04-01.

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.

Johannes Buchner. Imagehash. `https://pypi.org/project/ImageHash/`, 2013. Version 4.3.1.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL `https://doi.org/10.5281/zenodo.5143773`. If you use this software, please cite it as below.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021.

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.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

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.

Jeffrey S Bowers, Gaurav Malhotra, Marin Dujmović, Milton L Montero, Christian Tsvetkov, Valerio Biscione, Guillermo Puebla, Federico Adolfi, John E Hummel, Rachel F Heaton, et al. Clarifying status of dnns as models of human vision. *Behavioral and Brain Sciences*, 46, 2023.

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.

Xiaohan Wang, Tengyu Ma, James Ainooson, Seunghwan Cha, Xiaotian Wang, Azhar Molla, and Maithilee Kunda. The toybox dataset of egocentric visual object transformations. *arXiv preprint arXiv:1806.06034*, 2018.

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

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021b.

Amro Abbas and Stéphane Deny. Progress and limitations of deep networks to recognize objects in unusual poses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 160–168, 2023.

Linda B Smith and Lauren K Slone. A developmental approach to machine learning? *Frontiers in psychology*, 8: 296143, 2017.