Deep Vision Models Follow Shepard's Universal Law of Generalization

Daniel L. Carstensen (daniel_carstensen@brown.edu)
Department of Cognitive and Psychological Sciences, Brown University

Serra E. Favila* **(serra_favila@brown.edu)**
Department of Cognitive and Psychological Sciences, Brown University

Steven M. Frankland* **(steven.m.frankland@dartmouth.edu)**
Program in Cognitive Science, Dartmouth College

Abstract

Shepard's (1987) universal law of generalization holds that the probability of generalizing between two stimuli decays as a concave function of their distance in psychological space. While there is widespread evidence for the law in human perception, its relevance to artificial neural networks remains unclear, despite the importance of generalization for these systems. Here, we find that the representational spaces of models that vary in their architecture, objective, and training data yield a concave generalization gradient with respect to human judgments of naturalistic images (Peterson et al., 2018), consistent with Shepard's law. Our results suggest that the representational spaces of deep vision networks serve as compelling, but imperfect, proxies for classic psychological spaces derived from behavioral data. This highlights the strengths and weaknesses of deep vision models as contributors to cognitive theories of perceptual generalization, while adding further evidence for the generality of Shepard's law.

Keywords: generalization; deep learning; embedding spaces; perceptual similarity

Introduction

Generalization reflects a fundamental challenge for information processing systems. Suppose a bird preys on a bumblebee and is consequently stung. In order to avoid being stung again, the bird must identify organisms that resemble this bumblebee based on the perceptual information it receives. Since no two bumblebees look identical, the bird must be able to generalize over these possibilities. It does so at the cost of potentially missing opportunities to prey on visually similar—but innocuous—species that may closely resemble the bumblebee, such as the hoverfly (Edmunds & Reader, 2014).

Similarly, suppose that a convolutional neural network is presented with images of insects and asked to categorize them. Since no two images of a species, such as bumblebees, are exactly the same, the network must generalize over the possible pixel-level inputs to learn a useful representation of the class. Likewise, it must be careful not to learn overly coarse representations that treat visually similar, but functionally distinct, task-relevant categories as equivalent.

The broad scope of the challenge of generalization reinforces the search for unifying principles, first made famous by Shepard and subsequently strengthened by others (Chater & Vitányi, 2003; Frank, 2018; Shepard, 1987; Sims, 2018; Tenenbaum & Griffiths, 2001; Wu, Meder, & Schulz, 2025). Shepard posited that the strength of generalization between two stimuli decays as an invariant, concave upward function of their distance in a hypothetical metric "psychological space." Numerous empirical studies have supported Shepard's law across different types of low-dimensional stimuli and across different species, with generalization gradients usually following an exponential or, sometimes, a Gaussian function (e.g., Cheng, 2000; Ghirlanda & Enquist, 2003; Shepard, 1987). More recently, Marjieh et al. (2024) showed that the universal law holds for naturalistic stimuli by analyzing a large set of human similarity judgments made across sets of natural images.

Since psychological space is not directly observable, studies have relied on non-metric multidimensional scaling (NMDS) of behavioral data, a technique developed by Shepard (1962) and Kruskal (1964). NMDS takes a similarity matrix for a set of stimuli and maps the stimuli into a low-dimensional space, ensuring their pairwise similarities are preserved as accurately as possible and allowing for the extraction of stimulus distances. Here, we take a different approach to deriving psychological space. We ask: could the representational spaces of deep neural networks (DNNs) trained to optimize standard computer vision objectives naturally stand in as proxies for behaviorally-derived psychological spaces?

In recent years, advances in DNNs have significantly enhanced the capabilities of artificial processing systems. In computer vision, DNNs (or deep vision models) have achieved or surpassed human-level performance on a number of perceptual tasks involving naturalistic images (LeCun, Bengio, & Hinton, 2015; Kheradpisheh et al., 2016; Quesada et al., 2024; Russakovsky et al., 2015). While deep vision models are at most imperfect models of biological vision (Wichmann & Geirhos, 2023), growing evidence has revealed parallels between their internal representations (i.e., their embedding spaces) and the psychological and neural representations extracted from humans and non-human primates (Cichy et al., 2016; Khaligh-Razavi & Kriegeskorte, 2014; Muttenthaler et al., 2023; Rajalingham et al., 2018; Sucholutsky et al., 2023; Yamins et al., 2014). Importantly, prior work has shown that the representational geometry of DNN embedding spaces is predictive of human similarity judgments (Jha et al., 2023; Peterson et al., 2016, 2018) and can be finetuned to predict NMDS-derived psychological space (Sanders & Nosofsky, 2018). Nevertheless, standard deep vision models sub-

*Denotes equal contribution

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

stantially differ from human visual cognition in a number of ways, including their respective strategies in object recognition tasks (Bowers et al., 2023; Linsley et al., 2023). Hence, it remains unclear whether the embedding spaces of deep vision models might be suitable proxies for Shepard's psychological space.

Here, we evaluate this possibility by specifically asking whether the distances in DNN-generated embedding spaces predict human similarity judgments as a concave generalization gradient. To do so, we selected a diverse range of deep vision models that varied in architecture, training task, and training dataset, and drew from a large dataset of natural images with corresponding human similarity judgments (Peterson et al., 2018). We then extracted image embeddings from each deep vision model and computed the pairwise distance matrix of the embeddings, yielding a human-evaluated similarity score and a model-derived distance value for each image. Finally, we examined the resulting generalization gradients for each image set and for each model. We find that the internal representations of deep vision models predict human similarity judgments as a concave generalization gradient in a way that is strikingly similar to classic behaviorally derived psychological spaces.

Methods

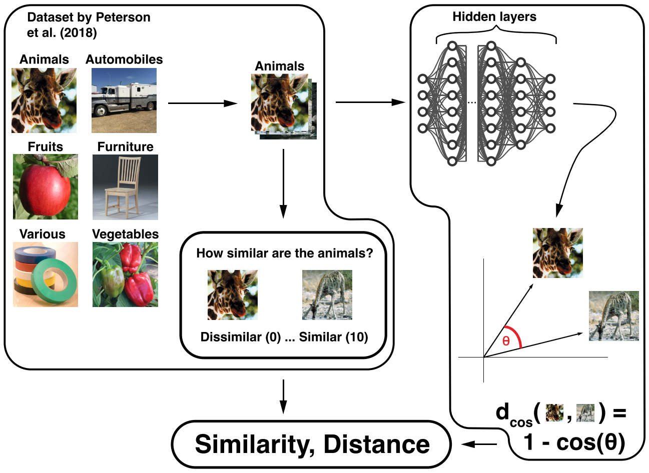


Figure 1: **Overview of the experimental approach.** Human similarity judgments are paired with model-derived distances to test whether deep vision model embeddings adhere to Shepard's universal law of generalization.

Stimuli and Vision Models

We used a dataset collected by Peterson et al. (2018). This dataset includes six sets of 120 images from the general categories "animals," "automobiles," "fruits," "furniture," "various," and "vegetables," and corresponding human pairwise similarity judgments (see Figure 1, left). Each image pair received ten unique human similarity judgments, reported on a scale from 0 ("not similar at all") to 10 ("very similar"). These data were aggregated into an average 120 x 120 similarity matrix for each of the six image sets and then rescaled to range from 0 to 1. Notably, Marjieh et al. (2024) also used the "animals", "fruits", and "vegetables" image sets, establishing the existence of Shepard's law in naturalistic stimuli. Using the same analysis, we replicated these results and tested if they applied to the remaining "automobiles," "furniture," and "various" image sets. We then selected a diverse set of pretrained deep vision model families that varied in architecture, training task, and training dataset (see Table 1). From each model family, we chose several models of varying size and, if applicable, using different architectures. This enabled us to conduct a comprehensive evaluation of embedding spaces across different deep vision models; in total, we tested 24 individual models. We additionally chose pixel-level MSE as a baseline model of low-level perceptual distance.

Table 1. Selected Deep Vision Model Families

Model family	Architecture	Training task	Training data
VGG	CNN	1K classification	ImageNet-1k
ResNet	CNN	1K classification	ImageNet-1k
ViT	Transformer	1K classification	ImageNet-1k
CLIP	CNN/Transformer	Language alignment	WebImageText
OpenCLIP	CNN/Transformer	Language alignment	CC12M/LAION-2B
DINO	CNN/Transformer	Self-supervision	ImageNet-1k
DINOv2	Transformer	Self-supervision	LVD-142M
DreamSim	Transformer	Perceptual alignment	NIGHTS

Embedding Extraction and Generalization Gradient Computation

Our goal was to test if the internal representations of deep vision models served as proxies for psychological space, thus aligning with Shepard's universal law of generalization. To derive generalization gradients, we mapped the human-evaluated similarity score for each image pair to a corresponding model-derived distance value (see Figure 1, right). Though prior work has mainly relied on Euclidean distance as a measure of distance in psychological space (e.g., Marjieh et al., 2024; Shepard, 1962, 1987), we chose to use cosine distance for our analyses because it is standard practice in the field of computer vision and has previously been shown to predict human similarity judgments (e.g., Fu et al., 2023; Radford et al., 2021; Roads & Love, 2021).[1] We first extracted individual image embeddings from the final hidden layer of each model. Embeddings from this layer contain high-level information most related to human categorization (Cohen et al., 2020; LeCun et al., 2015; Sucholutsky et al., 2023). For transformer-based models, we specifically extracted the classification token of the last hidden layer (Dosovitskiy et al., 2021). Then, we computed the cosine distance between each pair of embeddings from the same image set. Finally, we matched the human-evaluated similarity score for each image pair A-B with its associated cosine distance in model embedding space, yielding a similarity-distance tuple

[1] Nevertheless, we also performed the main analyses using Euclidean distance, yielding comparable results (see Figure 3).

(see Figure 1, bottom). To mitigate artificial bias in the distribution of similarity-distance tuples, we excluded trivial self-similarity tuples with a similarity of 1 and a distance of 0 (i.e., the diagonal of the similarity-distance matrix). To allow for cross-model comparisons, we then rescaled distances from each model such that the smallest distance equaled 0 and the greatest distance equaled 1. This produced six sets of similarity-distance generalization gradients corresponding to the six image sets for each model.

Since the sampling density of these gradients varied drastically, with substantially fewer high-similarity pairs than low-similarity pairs, the right tail of the gradient is overemphasized. To counteract this imbalance, we followed the approach taken in Marjieh et al. (2024) and additionally computed binned generalization gradients. Specifically, we divided the similarity-distance tuples into 100 equally spaced bins based on their distance values and computed the average similarity score and distance value for each bin.

Curve Fitting and Evaluation

To test if the embedding spaces of vision models abide by the universal law of generalization, we fit four candidate curves to the raw and binned generalization gradients. As a baseline, we chose a linear model of the form $f(x) = ax + b$. To allow for concave upward or downward fits, we additionally chose a quadratic model of the form $f(x) = ax^2 + bx + c$. Finally, we selected an exponential model of the form $f(x) = ae^{-bx} + c$ and a Gaussian model of the form $f(x) = ae^{-bx^2} + c$. Both exponential and Gaussian functions are models of the universal law that are supported by empirical and theoretical evidence (e.g., Chater & Vitányi, 2003; Cheng, 2000; Frank, 2018; Marjieh et al., 2024; Shepard, 1987). To fit the curves, we used the `model.fit` least squares optimizer from the `lmfit` package (Newville et al., 2024).

To evaluate the fit of the curves to the data, we computed the root mean squared error (RMSE), the coefficient of determination (R^2), and the Bayesian Information Criterion (BIC). We performed 5 x 5-fold cross-validation (i.e., five repetitions of standard 5-fold cross-validation) to obtain robust performance estimates (Burman, 1989; Kim, 2009; Stone, 1974). In particular, we randomly split the similarity-distance tuples from each generalization gradient into five equal-sized folds. Then, we fit the curves on the union of four folds and tested on the remaining fold, repeating this process five times with a different test fold each time. We repeated this procedure five times and averaged the resulting performance metrics across all folds. Since we were particularly interested in whether our non-linear curves approximated the generalization gradients better than the linear curve, we also computed the performance difference between each non-linear curve and the linear curve within each fold, yielding ΔRMSE, ΔR^2, and ΔBIC values.

Distance Alignment and Regression Analysis

To assess if model embedding spaces are viable proxy spaces for psychological space, we extracted the distances between all image pairs in NMDS space and matched each NMDS-derived distance value with its corresponding model-derived distance value. We split the resulting tuples into 100 equally spaced bins according to their model-derived distance values and computed an average NMDS distance value and model distance value for each bin. Then, we fit an ordinary least squares (OLS) model by regressing the NMDS-derived distances onto the model-derived distances for each individual model and recorded the estimated coefficients and Pearson correlation coefficients. Finally, we used a random-effects meta-analysis model to estimate pooled coefficients and explained variance along with 95% confidence intervals (CI).[2]

Residual Analysis

We used the residuals obtained during curve fitting to determine if model embedding spaces failed to adhere to Shepard's law in some regions. First, we examined a quantile-quantile plot (Q-Q plot) comparing the distribution of the residuals to a normal distribution. Additionally, we split the residuals into 10 equal-sized bins according to their corresponding ground-truth similarity value and computed the average residual in each bin. We applied the same evaluation to the residuals from the distance alignment regression analysis. This allowed us to assess if the distribution of the residuals along the gradient showed any systematic biases.

Results

Our goal was to investigate whether the embedding spaces of deep vision models follow Shepard's universal law. We note that for the sake of brevity, the results presented here derive from the binned generalization gradients unless otherwise indicated. Nevertheless, all findings apply to the raw gradients as well, though with reduced performance metrics due to increased variance.

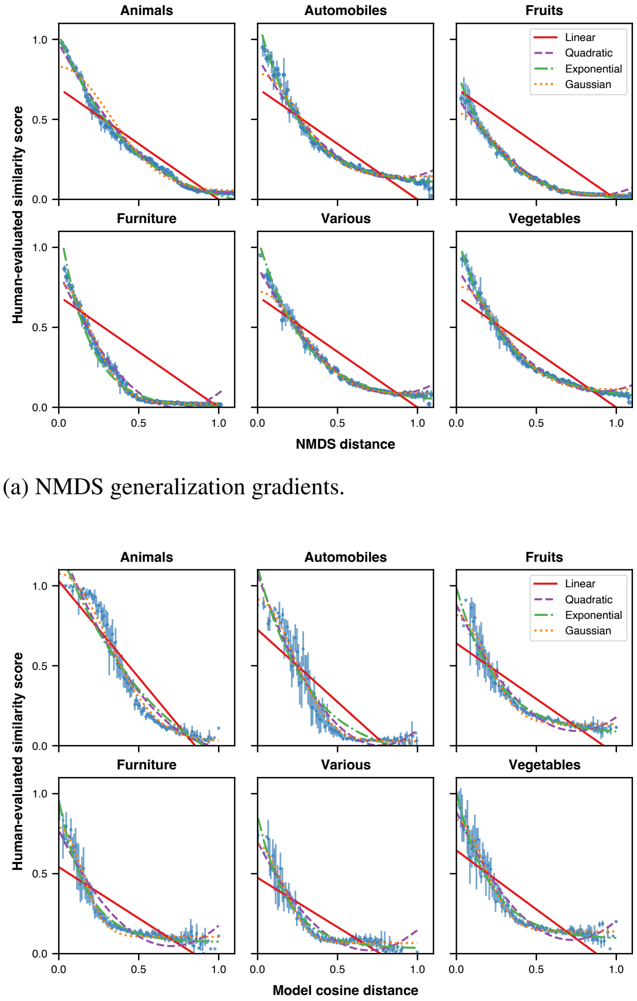
NMDS-Derived Psychological Space Adheres to Shepard's Law

We first constructed a baseline psychological space by mapping human similarity judgments for the six image sets into NMDS space. Our results reproduced the findings of Marjieh et al. (2024) and extended them to the remaining image sets ("automobiles," "furniture," and "various"), all of which exhibited concave generalization gradients and thus aligned with Shepard's law (Figure 2a).

Model-Derived Generalization Gradients Adhere to Shepard's Law

Having established this baseline, we focused our analysis on the model-derived generalization gradients. To test each model's adherence to Shepard's law, we fit four curves (linear, quadratic, exponential, Gaussian) to the binned generalization gradients and recorded RMSE, R^2, and BIC of each fit.

[2] To estimate pooled explained variance, we first computed a pooled Fisher z-transformed correlation coefficient, then inverse-transformed the estimate back to a correlation coefficient, and finally squared the result.



(a) NMDS generalization gradients.

(b) CLIP ViT-B/16 generalization gradients.

Figure 2: **Comparison of human and model-derived generalization gradients.** (a) NMDS-derived generalization gradients from human similarity judgments confirm Shepard's law across six image sets. (b) Model-derived generalization gradients for CLIP ViT-B/16 exhibit similar concavity, indicating alignment between deep vision models and psychological space.

Figure 2b shows the best-fit curves for the best-performing model (CLIP ViT-B/16) across all image sets, highlighting the strong concavity of the gradients. Overall, we found that the non-linear curves outperformed the linear curve across all image sets with the Gaussian curve performing the best (Table 2). Additionally, Table 3 reports the ΔRMSE, ΔR^2, and ΔBIC values for all non-linear curves. Since these differences were computed relative to the linear curve fits, the relative performance metrics for the linear curve are omitted from the table. These results held across all model families and generalized to Euclidean distance-derived generalization gradients (see Figure 3). Notably, no nonlinear curve reliably provided improved fits across all image datasets for gradients computed using the pixel-level MSE model. Finally, we found that over 83% of quadratic curve fits were concave upward with strictly positive second derivatives $f(x)'' = 2a > 0$ with CI $[0.7,\ 0.75]$. Since the quadratic curve allowed both concave upward and concave downward fits, the strong tendency toward positive second derivatives further supports adherence to Shepard's law. Overall, these results confirmed that model-derived generalization gradients aligned with Shepard's law.

Table 2. Performance Metrics of Curves

Curve	Metric	Mean	CI Lower	CI Upper
Exponential	RMSE	0.075	0.074	0.076
Exponential	R^2	0.818	0.787	0.840
Gaussian	RMSE	**0.064**	0.063	0.065
Gaussian	R^2	**0.852**	0.840	0.864
Linear	RMSE	0.102	0.101	0.103
Linear	R^2	0.679	0.653	0.701
Quadratic	RMSE	0.072	0.071	0.073
Quadratic	R^2	0.836	0.822	0.848

Table 3. Performance Metrics of Non-Linear Curves Relative to Linear Curve

Curve	Metric	Mean	CI Lower	CI Upper
Exponential	ΔRMSE	-0.027	-0.028	-0.026
Exponential	ΔR^2	0.139	0.104	0.170
Exponential	ΔBIC	-52.049	-53.719	-50.361
Gaussian	ΔRMSE	**-0.038**	-0.039	-0.037
Gaussian	ΔR^2	**0.174**	0.151	0.198
Gaussian	ΔBIC	**-72.075**	-73.869	-70.250
Quadratic	ΔRMSE	-0.030	-0.031	-0.029
Quadratic	ΔR^2	0.157	0.140	0.177
Quadratic	ΔBIC	-53.878	-55.282	-52.472

Model Embedding Spaces Approximate Psychological Space

Next, we assessed if model embedding spaces are viable proxy spaces for psychological space by regressing the binned NMDS-derived distances onto the binned model-derived distances. We observed a positive correlation between model-derived distances and NMDS-derived distances across all models and image sets. By applying a random-mixed effects meta-analysis, we found that in aggregate model-derived distances were able to explain over 86% of the variance in NMDS-derived distances. Table 4 lists the pooled slope, intercept, and explained variance estimates along with CIs. The pixel-level MSE model did not exhibit high predictive strength, with only 8% variance explained. Hence, model embedding spaces were adequate proxy spaces for psychological space.

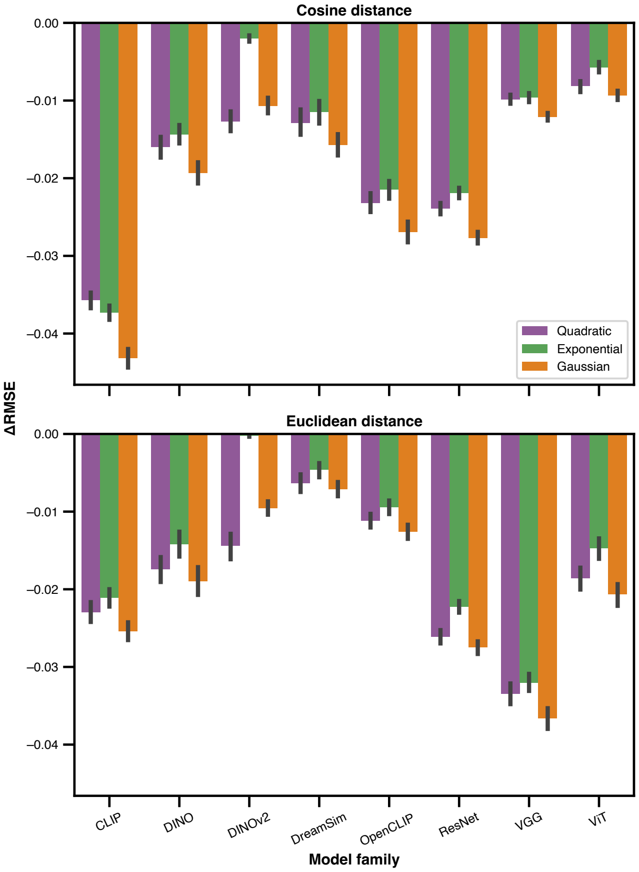


Figure 3: **ΔRMSE of non-linear curves across model families and distance metrics.** Non-linear curves outperformed the linear baseline across all model families regardless of distance metric used. Lower is better.

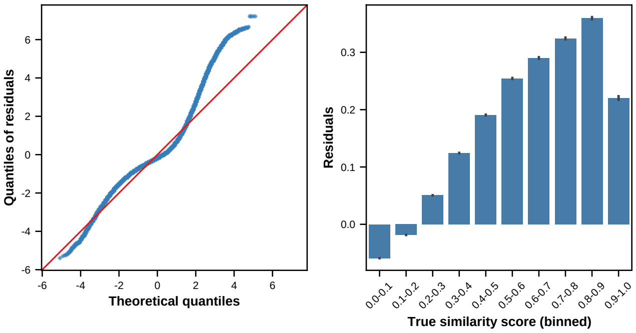


Figure 4: **Assessment of model deviations from Shepard's law.** The Q-Q plot (left) shows that residuals deviate from normality, while residuals binned by similarity score (right) reveal a systematic underestimation of high-similarity pairs, highlighting model limitations.

Table 4. Regression Coefficients and r^2 For NMDS Distances Predicted by Model Distances

	Pooled estimate	CI Lower	CI Upper
Slope	0.800	0.776	0.834
Intercept	0.081	0.051	0.111
r^2	0.868	0.850	0.884

Model Embedding Spaces Fail for High-Similarity Pairs

Finally, we assessed if model embedding spaces systematically deviated from Shepard's law in any region, making use of the curve residuals. Here, we used the residuals corresponding to the curves fitted to raw generalization gradients as opposed to binned gradients.[3] Figure 4 shows the Q-Q plot and the binned average residuals corresponding to the Gaussian curve. We found that residuals were well-behaved for low-similarity image pairs but were skewed for higher-similarity pairs, exhibiting systematic underestimation of similarity for those pairs. For example, two different images of *giraffes* may be farther apart in the network's embedding space than expected from the psychological space. This applied across all models and image sets. We corroborated these findings by examining the residuals of the distance alignment regression which showed the same pattern.

Discussion

In this study, we asked whether the representational spaces of a diverse group of deep vision models naturally serve as proxies for Shepard's "psychological space." We found that regressing human generalization judgments onto the cosine distances of image pairs in model embedding space naturally yielded Shepard's invariant concave upward generalization gradient (Figure 2b). Regression analyses corroborated these findings, showing a strong positive correlation (pooled $r^2 \approx 0.87$) between model-derived distances and NMDS-derived distances, demonstrating that the internal model representations naturally approximated psychological space derived from behavioral data. Nevertheless, deep vision models were not perfect predictors of human generalization judgments: residual analyses revealed systematic distortions for high-similarity image pairs.

In previous studies (e.g., Marjieh et al., 2024; Shepard, 1987), researchers approximated the relevant psychological space by performing NMDS on a matrix of human behavioral judgments. In the present work, we exploited our ability to fully observe the internal representational spaces of deep vision models to directly approximate distances in psychological space. Remarkably, neither the training data nor the objective used to train the models (see Table 1) had any obvious connection to the behavioral task or the corresponding

[3] We used residuals from the raw generalization gradients instead of the binned gradients because residuals specifically capture deviations from the fitted model, rather than reflecting the overall distribution of the raw data. This distinction is important because the raw data distribution can be influenced by uneven sampling, whereas residuals isolate how much each data point differs from the model's expected pattern. By working with residuals, we could retain the full dataset without discarding data points due to binning, which maximized statistical power in our analyses.

data from Peterson et al. (2018).[4] Nonetheless, our results paralleled previous findings that inferred psychological space directly from human behavioral data, suggesting that these components successfully shaped the models' representational spaces without directly relating to the behavioral context we tested them in. This further supports the universality of Shepard's law and suggests considerable flexibility in how representational spaces that exhibit this law can be derived.

Our analyses thus far suggest that the emergence of concave generalization gradients in model embedding spaces is not tightly tied to any particular network architecture, objective function, or training dataset. We specifically tested deep vision models that, while not exhaustive, varied across all of these characteristics (see Table 1); yet their embedding spaces consistently exhibited alignment with Shepard's law. Our findings fall in line with recent studies that have revealed surprising levels of alignment between different deep vision models, both with respect to their representational spaces (Huh et al., 2024) and with respect to their brain predictivity (Conwell et al., 2024). Finally, our findings suggest that the representational spaces of conventional deep vision models already capture a high proportion of variance in psychological space. While it remains to be seen if this pattern transfers to DNNs of other modalities, this highlights the potential of such models to serve as computational proxies for human perceptual generalization after further refinement.

In ongoing and future work, we plan to address several questions left open by the present study. First, while we found concave generalization gradients across all tested models, we observed significant differences between individual models (see Figure 3). It is crucial to understand if and how differences in model architecture, training objective, and training data influence the degree of adherence to Shepard's law. Going forward, we hope to design and train custom models to help us achieve this. A deeper understanding of the model characteristics driving alignment with psychological space and Shepard's law will not only aid in the development of more accurate models of human perception and generalization but will also inform theoretical accounts of this law. For instance, by systematically manipulating the model architecture and/or training procedure, future work could potentially arbitrate between competing accounts of how this law arises (e.g., Frank, 2018; Sims, 2018; Tenenbaum & Griffiths, 2001).

Furthermore, it is unclear why model embedding spaces fail to align with psychological space for high-similarity pairs and whether this deviation is correctable. We aim to employ several approaches to remedy this deviation. For example, it is possible that fine-tuning techniques that align the visual strategies used by deep vision models with human behavioral data (Fel et al., 2022) will also align their embedding spaces for high-similarity pairs. Moreover, large language models (LLMs) are increasingly performant predictors of human behavior (Binz et al., 2024). Hence, it may be possible to treat LLMs as though they were participants in a behavioral experiment and replicate the analyses in Marjieh et al. (2024) with LLM-evaluated similarity judgments. Crucially, in-context learning (Dong et al., 2024) allows LLMs to learn from examples in the provided context (and without adjustment of their weights). This paradigm provides a tool that could replicate the experimental conditions under which human participants made similarity judgments in Peterson et al. (2018) with greater precision. This could yield similarity judgments that more closely mirror the context dependence of human similarity judgments (Medin, Goldstone, & Gentner, 1993; Tversky, 1977).

Finally, we hope to assess the utility of Shepard's law as a novel measure of representational alignment. Often, representational alignment of deep vision models is measured through linear probing or correlation with neural data and/or measures of behavior, such as similarity judgments or recognition task performance, yielding a singular alignment score (e.g., Schrimpf et al., 2018). Recent work has further sought to align the representational geometries of deep vision models to behaviorally derived human similarity spaces (for review see Sucholutsky et al., 2023). Here, Shepard's psychological space could be a useful target of representational alignment. Parameterizing alignment in this space could offer a more fine-grained measure that identifies where DNN representations adhere to human cognitive representations and where they diverge.

In summary, our study demonstrates that deep vision models naturally capture a concave generalization gradient, paralleling the predictions of Shepard's universal law. These findings underscore the potential of model-derived representational spaces as computational proxies for human perceptual generalization, even in the absence of direct training on behavioral data. Looking forward, we plan to refine our understanding of the factors that drive alignment between artificial and psychological spaces and explore the utility of large language models in replicating human similarity judgments. By integrating insights from computational modeling and behavioral experiments, we hope to advance a more unified framework for understanding generalization across both artificial and biological systems.

References

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., ... Schulz, E. (2024). *Centaur: a foundation model of human cognition.* arXiv.

Bowers, J. S., Malhotra, G., Dujmović, M., Llera Montero, M., Tsvetkov, C., Biscione, V., ... et al. (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, *46*, e385.

Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, *76*(3), 503–514.

[4] DreamSim, one of the models we tested, was explicitly trained to predict human similarity judgments. Somewhat surprisingly however, DreamSim did not perform better than other models.

Chater, N., & Vitányi, P. M. B. (2003). The generalized universal law of generalization. *Journal of Mathematical Psychology*, *47*(3), 346–369.

Cheng, K. (2000). Shepard's universal law supported by honeybees in spatial generalization. *Psychological Science*, *11*(5), 403–408.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), 27755.

Cohen, U., Chung, S., Lee, D. D., & Sompolinsky, H. (2020). Separability and geometry of object manifolds in deep neural networks. *Nature Communications*, *11*(1), 746.

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2024). A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, *15*(1), 9383.

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., … Sui, Z. (2024). *A survey on in-context learning.* arXiv.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *The Ninth International Conference on Learning Representations.*

Edmunds, M., & Reader, T. (2014). Evidence for batesian mimicry in a polymorphic hoverfly. *Evolution*, *68*(3), 827–839.

Fel, T., Rodriguez, I. F. R., Linsley, D., & Serre, T. (2022). Harmonizing the object recognition strategies of deep neural networks with humans. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 9432–9446). Curran Associates, Inc.

Frank, S. A. (2018). Measurement invariance explains the universal law of generalization for psychological perception. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(39), 9803–9806.

Fu, S., Tamir, N. Y., Sundaram, S., Chai, L., Zhang, R., Dekel, T., & Isola, P. (2023). DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems* (Vol. 36, pp. 50742–50768). Curran Associates, Inc.

Ghirlanda, S., & Enquist, M. (2003). A century of generalization. *Animal Behaviour*, *66*(1), 15–36.

Huh, M., Cheung, B., Wang, T., & Isola, P. (2024, 21–27 Jul). Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235, pp. 20617–20642). PMLR.

Jha, A., Peterson, J. C., & Griffiths, T. L. (2023). Extracting Low-Dimensional psychological representations from convolutional neural networks. *Cognitive Science*, *47*(1), e13226.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLOS Computational Biology*, *10*(11), e1003915.

Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., & Masquelier, T. (2016). Deep networks can resemble human feed-forward vision in invariant object recognition. *Scientific Reports*, *6*(1), 32672.

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, *53*(11), 3735–3745.

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*(2), 115–129.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Linsley, D., Rodriguez, I. F. R., FEL, T., Arcaro, M., Sharma, S., Livingstone, M., & Serre, T. (2023). Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex. In *Thirty-seventh conference on neural information processing systems.*

Marjieh, R., Jacoby, N., Peterson, J. C., & Griffiths, T. L. (2024). The universal law of generalization holds for naturalistic stimuli. *Journal of Experimental Psychology: General*, *153*(3), 573–589.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*(2), 254–278.

Muttenthaler, L., Dippel, J., Linhardt, L., Vandermeulen, R. A., & Kornblith, S. (2023). Human alignment of neural network representations. In *The eleventh international conference on learning representations.*

Newville, M., Otten, R., Nelson, A., Stensitzki, T., Ingargiola, A., Allan, D., … Persaud, A. (2024). *lmfit/lmfit-py: 1.3.2.* Zenodo.

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

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

Quesada, J., Fowler, Z., Alotaibi, M., Prabhushankar, M., & AlRegib, G. (2024). Benchmarking human and automated prompting in the segment anything model. In *2024 IEEE International Conference on Big Data (BigData)* (pp. 1625–1634). IEEE Computer Society.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 8748–8763). PMLR.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). 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.

Roads, B. D., & Love, B. C. (2021). Enriching ImageNet with human similarity judgments and psychological embeddings. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (p. 3546-3556). IEEE Computer Society.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Sanders, C., & Nosofsky, R. M. (2018). Using deep-learning representations of complex natural stimuli as input to psychological models of classification. In C. Kalish, M. A. Rau, X. J. Zhu, & T. T. Rogers (Eds.), *Proceedings of the 40th annual meeting of the cognitive science society.*

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., … DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, *27*(3), 219–246.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science*, *360*(6389), 652–656.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), 111–133.

Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., … Griffiths, T. L. (2023). *Getting aligned on representational alignment.* arXiv.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–40; discussion 652–791.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.

Wichmann, F. A., & Geirhos, R. (2023). Are deep neural networks adequate behavioral models of human visual perception? *Annual Review of Vision Science*, *9*(1), 501–524.

Wu, C. M., Meder, B., & Schulz, E. (2025). Unifying principles of generalization: Past, present, and future. *Annual Review of Psychology*, *76*(1), 275–302.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8619–8624.