L-WISE: BOOSTING HUMAN IMAGE CATEGORY LEARNING THROUGH MODEL-BASED IMAGE SELECTION AND ENHANCEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

The currently leading artificial neural network (ANN) models of the visual ventral stream – which are derived from a combination of performance optimization and robustification methods -– have demonstrated a remarkable degree of behavioral alignment with humans on visual categorization tasks. Extending upon previous work, we show that not only can these models guide image perturbations that change the induced human category percepts, but they also can enhance human ability to accurately report the *original* ground truth. Furthermore, we find that the same models can also be used out-of-the-box to predict the proportion of correct human responses to individual images, providing a simple, human-aligned estimator of the relative difficulty of each image. Motivated by these observations, we propose to *augment visual learning* in humans in a way that improves human categorization accuracy at *test time*. Our learning augmentation approach consists of (i) selecting images based on their model-estimated recognition difficulty, and (ii) using image perturbations that aid recognition for novice learners. We find that combining these model-based strategies gives rise to test-time categorization accuracy gains of 33-72% relative to control subjects without these interventions, despite using the same number of training feedback trials. Surprisingly, beyond the accuracy gain, the training time for the augmented learning group was also shorter by 20-23%. We demonstrate the efficacy of our approach in a fine-grained categorization task with natural images, as well as tasks in two clinically relevant image domains – histology and dermoscopy – where visual learning is notoriously challenging. To the best of our knowledge, this is the first application of ANNs to increase visual learning performance in humans by enhancing category-specific features.[1]

1 INTRODUCTION

Over the last decade, specific artificial neural network (ANN) models have been shown to be the best image-computable emulators of neural processing along the human and monkey ventral visual stream and its support of a range of human visual tasks, as measured by human behavior. Iterative efforts have developed even better models in this same vein that are more and more accurate emulators. Indeed, among the most contemporary of such models – so called "robustified" deep ANN models (Mądry et al., 2018) – have been shown to allow the design of images to predictably control both ventral stream neural activity (Guo et al., 2022) and human object categorization reports (Gaziv et al., 2024; Croce & Hein, 2020).

Learning to recognize new, unfamiliar categories in images is a task that the human visual system normally excels at. Importantly, despite being a basic task for humans, visual learning of new categories often carries significant practical relevance: for example, medical students and residents devote numerous hours to mastering the diagnosis of various diseases in image modalities such as histology (microscopic images of cells and tissues) and dermoscopy (skin lesions). They achieve this by practicing category recognition on a variety of imaged examples representing different cases. Studies of perceptual learning with simple tasks, such as line orientation discrimination, effectively show a curriculum effect whereby providing easy trials to a novice learner before gradually increas-

[1] We will make our code and data publicly available upon publication.

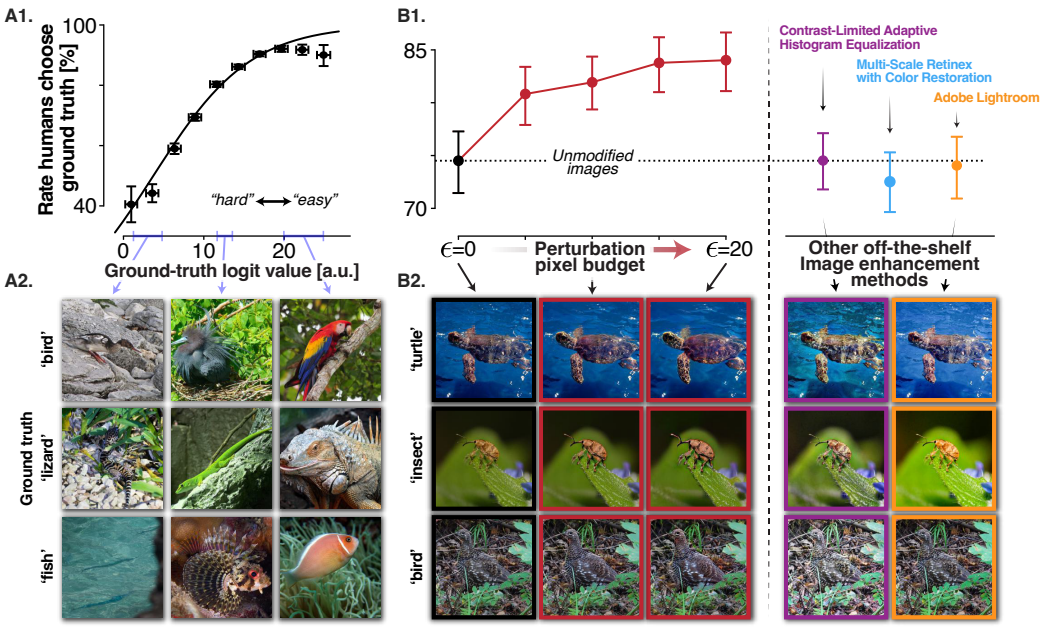

Figure 1: **Robustified ANNs can be used out-of-the-box as image-recognition difficulty estimators and ground-truth percept enhancers.** *We consider a 16-way basic animal classification task:* ***A1*** *Correspondence between human categorization accuracy and the model-computed ground-truth logit activation value. The curve denotes a logistic regression model predicting the probability of a correct response based on the logit value (AUC=0.71 under 10-fold cross validation, $p < 0.001$).* ***A2*** *Example images with varying ground truth logit values (predicted difficulty)* ***B1*** *Perturbing images by ground truth logit maximization increases recognition accuracy progressively with the perturbation ℓ_2-norm pixel budget ϵ. Other off-the-shelf image enhancement methods do not increase categorization accuracy, despite inducing larger perturbations of $\epsilon = 43$, $\epsilon = 106$, and $\epsilon = 26$ on average from left to right.* ***B2*** *Example images: Unmodified (left), enhanced by ground truth logit maximization with pixel budget $\epsilon = 10$ and $\epsilon = 20$, and by off-the-shelf methods (right). All vertical error bars are 95% CI by bootstrap. Horizontal error bars in A1 denote SD over images within each of the logit value bins.*

ing the difficulty promotes faster perceptual learning (Lu & Dosher, 2022). Motivated by these findings, here, we asked whether the human-behavioral and neural alignment that is attributed to these models makes them useful in helping humans learn to recognize new, unknown categories. Enhancing learning with a model, therefore, serves as both another scientific test of the model's validity and a beneficial application for human education.

To test the viability of model-guided boosting of visual category learning in humans, we first established two key empirical observations, summarized in Fig. 1: (i) We found that the human error rate in an object categorization task is highly predicted by the ground truth logit activation of a robustified ANN, rendering it a valid image-recognition difficulty score for humans. (ii) We found that this relationship also holds in reverse – pixel-level perturbations can be guided using the model in a way that increases the ground truth logit activation, generating a "perturbed" version of the image that is *"easier"* to recognize as the ground truth label suggests. Notably, other off-the-shelf image enhancement methods do not result in significant increases in categorization accuracy by humans. We thus propose a method that combines model-based image difficulty prediction and enhancement to generate optimized curricula for novice humans learning challenging image classification tasks.

Our proposed method, "Logit-Weighted Image Selection and Enhancement" (L-WISE), is illustrated in Fig. 2. In the general setting, a novice human learner completes a categorization learning task on an online platform. A labeled image dataset, with unfamiliar category labels, is used to teach the categories by showing examples and providing per-trial feedback after each image category judgment by the learner. Upon completion of the training phase, test accuracy is measured on held-out images in similar trials without feedback (Fig. 2ABC).

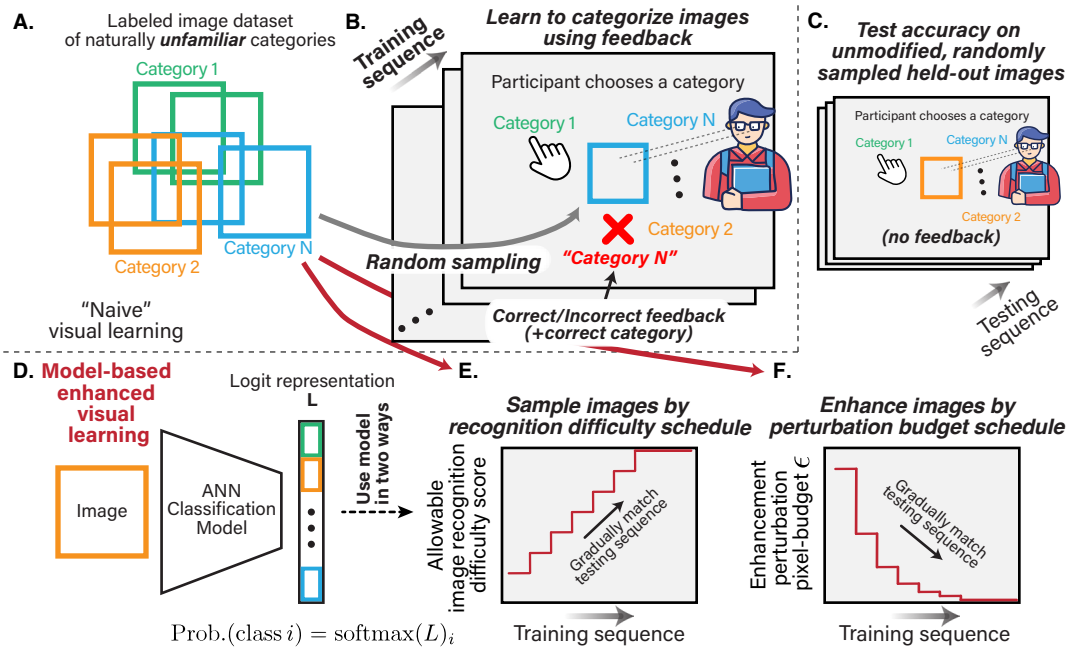

Figure 2: **Robustified ANNs can be used to boost novel-object recognition learning in humans.** *A novice human learner performs an image categorization task, which consists of a training phase (**B**) and a test phase (**C**). Images for both phases are randomly drawn from a labeled image dataset of fine-grained unfamiliar categories (**A**). Feedback (correct/incorrect, with indication of the correct category) is delivered during the training phase only. Our proposed "Logit-Weighted Image Selection and Enhancement" (L-WISE) approach uses an ANN model (**D**) to augment the visual learning curriculum by using the difficulty score to sample images based on a predefined profile of maximal difficulty per trial (**E**), and by enhancing images for easier recognition with an enhancement magnitude of a predefined trial-sequence profile (**F**).*

L-WISE intervenes on this naive visual learning baseline using a robustified ANN model in two ways: (i) it uses the difficulty score to sample images based on a predefined profile of maximal difficulty per trial (Fig. 2DE); (ii) it "perturbs" training images to be easier for recognition with an enhancement magnitude parameter similarly governed by a predefined trial-sequence profile (Fig. 2DF).

Surprisingly, despite the human visual system being well-adapted for rapidly learning new visual categories, we find that L-WISE gives rise to substantial gains of 33-72% in test-time accuracy margins above chance relative to control subjects without these interventions. In addition to improved accuracy, the duration of the training for the augmented learning group was significantly reduced. We further demonstrate the practical utility of our approach in boosting visual category learning across varied image domains and category spaces. In particular, we considered three main categorization tasks: moth species in natural images, dermoscopy images, and histology images.

Notably, our approach is inspired by recent findings demonstrating model-guided perturbations that accurately modulate category percepts *away* from the ground truth label (Gaziv et al., 2024). In this work, we conversely seek to amplify the ground truth percept, i.e., to facilitate the correct categorization of a given image.

Our contributions are several-fold:

- We establish a new state-of-the-art in predicting image recognition difficulty for humans, using a simple approach employing robustified ANN logit activations.
- We show that the leading models of the ventral visual stream can guide image perturbations that enhance human ability to accurately report the original ground truth label.
- We propose a novel model-based visual learning augmentation approach for humans that substantially increases test-time categorization accuracy at a reduced training time. To the best of our

knowledge, this is the first application of image enhancement to augment human visual learning.

• We demonstrate the general applicability of our proposed approach on categorization tasks across a range of image domains – including clinically relevant ones.

2 Related Work

We develop two important capabilities that form the foundation of our approach to assisting learners: (1) state-of-the-art predictions of the classification difficulty of images for humans, and (2) image perturbations that increase human categorization accuracy. Many works have ranked the difficulty of images to design curricula for ANNs (Wang et al., 2021b). Leading approaches include the c-score learning speed proxy (Jiang et al., 2021) and the prediction depth (Baldock et al., 2021) calculated for each image. Mayo et al. (2023) applied these techniques to predict the difficulty of natural images for humans, defined as either the minimum viewing time required to classify a given image correctly, or (as in our work) the proportion of humans who correctly classify it. Here, we show that the logit score associated with the ground truth class from a robustified ANN model more accurately predicts human image difficulty than prior approaches.

Enhancing image quality has been the focus of many previous studies (Qi et al., 2021), ranging from correction of factors such as lighting and contrast (e.g., Zuiderveld (1994); Jobson et al. (1997)) to ANN models that "upsample" images to higher resolution (Anwar et al., 2020). However, very little research has focused on enhancing images to more strongly represent a specific category. Prior works in this vein focused on making images easier for ANN models to classify correctly (Kim et al., 2023; Tussupov et al., 2023) or less vulnerable to subsequent adversarial attacks (Salman et al., 2021; Frosio & Kautz, 2023). Such perturbations, however, do not strongly affect human perception due to misalignment between human and machine perception (Gaziv et al., 2024).

Other studies focused on model-human alignment. Brain-Score directly benchmarks ANN models with respect to neural representation and downstream behavior (Schrimpf et al., 2018); "Harmonization" methods directly drive alignment by an auxiliary objective on ANN-predicted feature importance maps and crowd-sourced ones (Fel et al., 2022). Other works introduce architecture components to account for additional aspects of human vision, such as the dorsal-stream "where" pathway in the brain (Choi et al., 2023).

A key property that enables ANNs to generate human-interpretable image perturbations is that of perceptually aligned gradients, which is closely related to adversarial robustness and can be induced through adversarial training Ganz et al. (2023); Gaziv et al. (2024). Here, we apply adversarially-trained ANNs to enhance images such that they are more strongly associated with their ground truth label by the guiding model and by humans. To the best of our knowledge, we are the first to demonstrate improved human performance on image classification tasks through category-specific image enhancement.

Our primary goal is to apply difficulty prediction and image enhancement to assist human learning. The emerging field of machine teaching (Zhu, 2015) employs machine learning to find or generate optimal "teaching sets" that can be used to train other models or humans. While many such approaches have been successfully applied to training machine learning models (e.g., Liu et al. (2017); Qiu et al. (2023)), few studies have successfully enhanced image category learning in humans and most of these focus on teaching set selection. Singla et al. (2014) propose STRICT, which optimizes the expected decrease in learner error based on how the selected images and their labels constrain a linear hypothesis class in a feature space. Johns et al. (2015) extend a similar approach to select images in an online fashion by modeling the learner's progress. MaxGrad (Wang et al., 2021a) uses bi-level optimization to iteratively refine a teaching set by modeling learners as optimal empirical risk minimizers. Most similar to our work are approaches like EXPLAIN (Mac Aodha et al., 2018), which uses ANN class activation maps (CAMs) to highlight relevant image regions while providing feedback to the learner. EXPLAIN also selects a curriculum of images based on (a) a multi-class adaptation of STRICT, (b) representativeness (mean feature-space distance to other images of the same class), and (c) the estimated difficulty (entropy) of the CAM explanations. Chang et al. (2023) use bounding boxes to highlight image regions attended to by experts and not novices, allowing humans to more accurately match bird or flower images to one species among five shown in a gallery.

Our approach departs from previous studies in several ways. We make explicit estimates of image difficulty with unprecedented accuracy to select easier images for early-stage learners. We are

unique in employing category-specific image enhancement, which is a novel technique in itself, to improve the teaching efficacy of a given set of images. While Mac Aodha et al. (2018) and Chang et al. (2023) help learners by explicitly highlighting *where* learners should attend to in the image, we take a distinct and complementary approach by implicitly highlighting *what* learners must attend to in order to classify images correctly.

3 Overview of approach and experiments

Our approach to improve visual learning in humans is based on two key observations in the leading models of the ventral visual stream: (i) they can predict image recognition difficulty well for humans (Fig. 1A), and (ii) they can be used to perturb images in a way that enhances human ability to accurately report the original ground truth (Fig. 1B). In other words, these models can be used out-of-the-box as *category-recognition difficulty estimators*, and *category-percept enhancers*. As such, we propose using them to augment a visual learning sequence that is used to learn and practice recognition of unfamiliar image classes in various challenging categorization tasks.

Fig. 2 summarizes our approach. In the naive baseline scenario, the novice learner is presented with a sequence of training trials, where in each, a randomly selected image from one of N categories is presented, and the user is tasked to perform an N-way categorization task through an online survey platform. During this training phase, the learner receives correct/incorrect feedback after each trial. Notably, the category labels are made meaningless by assigning them to be a random Greek name unrelated to the task, such that learners start at chance level. At the end of the training phase, the experiment transitions into the test phase, where the same task continues over a held-out set of images, and no feedback is delivered. During the test phase we measure the visual learning outcome of interest, the test accuracy (Fig. 2A-C).

Harnessing the key observations on robustified ANNs, our approach uses a model to optimize the training phase in a way that improves the test accuracy via two mechanisms: (i) sampling training images based on their predicted recognition difficulty, and (ii) enhancing the training images. Both mechanisms have "strength" control knobs which are generally time dependent, namely the maximally allowable difficulty of an image at a given trial, and the corresponding enhancement strength. The latter is approximated via the ℓ_2-norm pixel-budget ϵ. Using these mechanisms, the user of our approach can flexibly define arbitrary time-dependent profiles for image selection and enhancement (Fig. 2D-F). In this study, we focused on a linear ramp profile for the allowable image difficulty at a given time and on exponential tapering of the enhancement ϵ. Intuitively, this should correspond to an easy-to-challenging traversal during the training phase. Notably, extensively optimizing these profiles was not our goal.

Incorporating these mechanisms led to significant gains in the test-time accuracy of human participants, while also requiring less time to complete the training phase (which includes a constant number of trials). This result was robustly obtainable across the varied image domains and category spaces tested. In particular, we considered three categorization tasks: moth species in natural images, dermoscopy images (skin lesion diagnosis), and histology images (benign lesions vs pre-cancerouss lesion diagnosis). We next describe the model and explain the two model-based mechanisms that drive the boost in visual learning.

3.1 Training task-specific robustified models

To obtain robustified models for a task specific category space, we adversarially-trained ResNet-50 ANNs (He et al., 2016) on the ImageNet-1K (Deng et al., 2009) and iNaturalist 2021 (Van Horn et al., 2021) datasets (separately) using the same technique from Mądry et al. (2018). To adapt the resulting model to the three categorization tasks of interest, we conducted additional adversarial fine-tuning on the smaller datasets: a small subset of moth species images from iNaturalist, the HAM10000 skin lesion dermoscopy dataset (Tschandl et al., 2018), and the MHIST colon histology dataset (Wei et al., 2021).

3.2 Predicting category recognition difficulty

We propose an elegant way to predict the human categorization error rate on a given image, which suggests a new image-recognition difficulty score: the logit activation (pre-softmax) at the ground truth category output. The higher this logit value is, the lower the human categorization error rate. We established this relationship through user study reports in a basic natural image categorization task with 16 animal categories (Fig. 1A). We found this robustified model-based metric to be the current state-of-the-art in predicting human error rates (see Appendix Fig. S4).

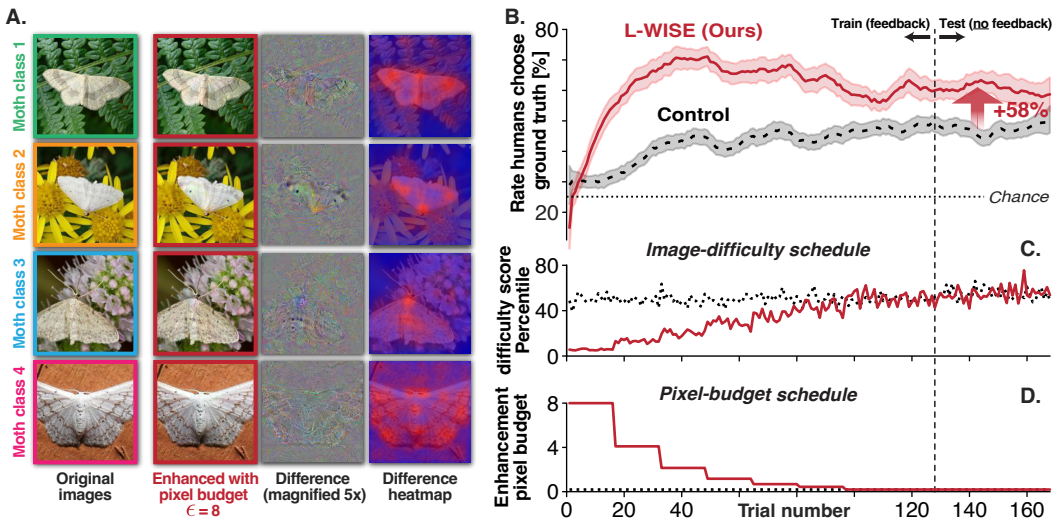

Figure 3: **Novice learners who had their curriculum augmented by our method showed improved test-time categorization accuracy for previously unfamiliar categories.** *Empirical results for a 4-way fine-grained moth species categorization task. Panel* ***A*** *shows examples of the four moth classes in the task, side-by-side with their model-based enhanced versions at the highest pixel budget used in our experiments ($\epsilon = 8$). While subtle, one notable difference is the distinctive wing spots of moth class 2, which are enlarged in the enhanced version of the image. Also shown are difference images showing the (5x magnified) difference between original and enhanced versions, and heat maps with more red coloration in regions of larger changes from enhancement.* ***B*** *compares the average smoothed accuracy of participants in the L-WISE group and a control group. Shaded areas denote 95% CI by bootstrap. The test accuracy gain of the L-WISE group relative to the control group is statistically significant (chi-square, $p < 0.001$).* ***C,D*** *show the trial-dependent empirical profiles of the average image difficulty percentile of selected images, which (noisily) step-wise increases, and the perturbation pixel budget for enhancement (ϵ), which step-wise decreases. The profiles of the baseline controls (black dotted lines) are uniform, indicating randomly-selected images with no enhancement.*

3.3 Generating image perturbations to enhance category percept

Given a bona fide metric of image difficulty, backpropagation on the model from output back to pixel space allows us to perturb an image to be easier to recognize with respect to its ground truth label. To this end, we impose a criterion to maximize the ground-truth logit activation, while also constraining the perturbations to be within a predefined ℓ_2-norm pixel-budget. This approach is analogous to Gaziv et al. (2024) but is designed to enhance the ground truth percept, rather than guiding away from it.

4 Results

We used robustified ANNs to both enhance images and predict the difficulty of images across multiple domains. We applied each of these techniques to improve the final test performance (on original, unmodified images) of novices learning challenging image classification tasks.

4.1 Models can both predict image recognition difficulty and reduce it

We tested the effects of image enhancement via maximization of the ground truth logit from robustified ANNs on human image categorization accuracy. We successfully demonstrate that we can enhance images by maximizing the ground truth logit from a robustified ANN (ResNet-50) using gradient descent in image pixel space. As the size of the perturbations grows (ℓ_2-norm pixel budget ϵ), human participants become increasingly accurate on a 16-way animal photograph classification task derived from ImageNet (Fig. 1B1, chance = 1/16). While mean accuracy on the original, unmodified ($\epsilon = 0$) images was 0.75, mean accuracy on enhanced images was as high as 0.84 at $\epsilon = 20$. The accuracy gains from enhancement appear to reach a saturation point as the perturbations grow larger. The improvements in accuracy are also somewhat dependent on the starting ground truth logit score, as shown in Appendix Fig. S8: accuracy gains are significantly higher for "difficult" images than for "easy" images. Baseline enhancement algorithms Contrast-Limited Adaptive His-

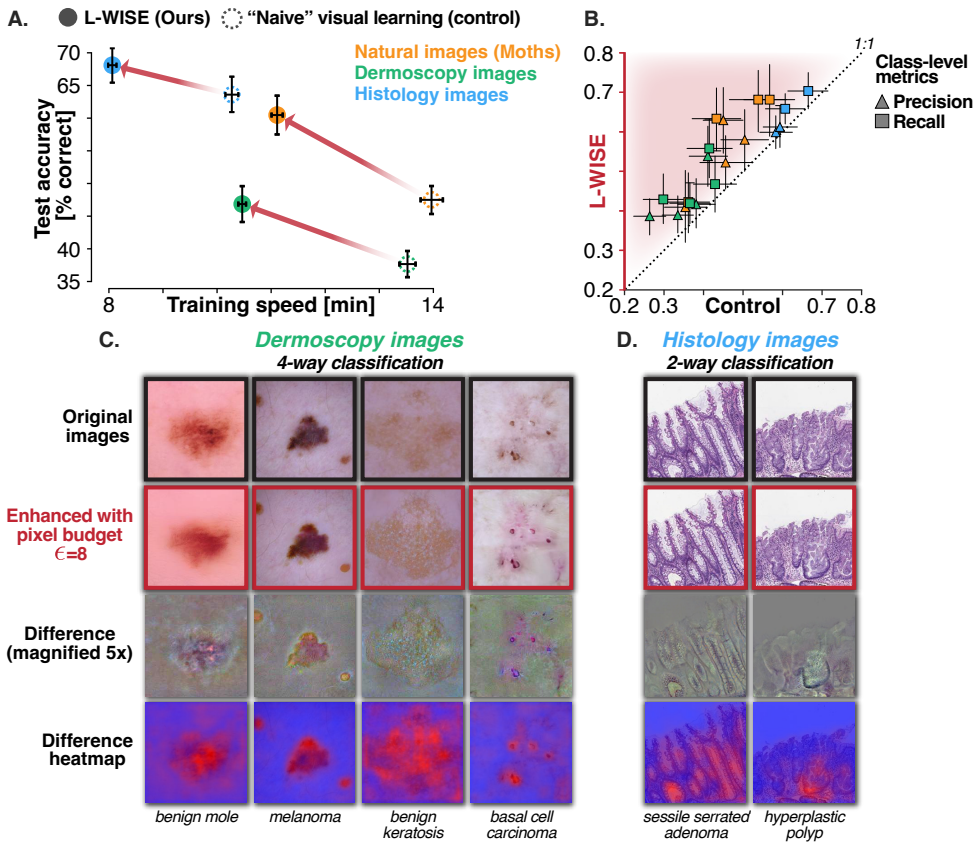

Figure 4: **Our approach can boost accuracy and time efficiency of image category learning for humans across varied image domains, including clinically-relevant ones.** *Panel **A** compares the mean test-time accuracy and training-phase duration of human participants who learned a moth photo, dermoscopy, or histology classification task, randomized to L-WISE and control groups. All differences between L-WISE and the control group are statistically significant (chi-squared test, $p < 0.05$). Panel **B** shows how precision and recall in L-WISE and control groups, with each point representing a specific class in one of the three tasks. All error bars show 95% bootstrap confidence intervals. Each class from the dermoscopy and histology tasks is illustrated in panels **C** and **D** respectively, similarly to the moth classes in Fig. 3A.*

togram Equalization (CLAHE), Multi-Scale Retinex with Color Restoration (MSRCR), and Adobe Photoshop Lightroom's "Auto" enhancement feature (LR) had no significant effect on performance, despite inducing image perturbations of considerably larger ℓ^2 norm on average than the ℓ^2 pixel budget ϵ values we used for model-based enhancement.

We also demonstrate that the robustified model's ground truth logit $L_{\text{gt}}(x)$ is strongly correlated with the rate at which humans choose the ground truth category associated with image x in a 16-way basic animal classification task (Fig. 1A). We used robustified ResNet-50 to calculate L_{gt} for each of the 2,400 distinct natural images used in the task, and applied logistic regression to predict binary correct vs. incorrect responses to individual image trials (pooling responses to original images and MSRCR, CLAHE, and LR control images to increase sample size). The logistic regression model (Fig. 1A) used L_{gt} to predict the correctness of the trial responses with **AUC** $= 0.71$ (10-fold cross-validation, $p < 0.001$). Notably, we found this simple approach to be better aligned with human error rates than the c-score, a state-of-the-art metric at predicting the difficulty of images for humans (Mayo et al., 2023) (see Appendix Fig. S4). We also demonstrate difficulty prediction and image enhancement with XCiT vision transformers (see Appendix Figs. S5-S7).

4.2 L-WISE IMPROVES BOTH LEARNING SPEED AND TEST ACCURACY FOR HUMANS

We applied both image difficulty prediction and image enhancement in a novel framework that designs curriculum image sequences for novices learning challenging image classification tasks.

	Idaea moth photos		Skin lesion dermoscopy	
	Mean Acc.	Training Time	Mean Acc.	Training Time
Chance level	0.25	-	0.25	-
Control	0.47 (0.45, 0.50)	14.0 (13.8, 14.2)	0.38 (0.36, 0.40)	13.5 (13.4, 13.7)
ET	0.58* (0.55, 0.61)	11.8 (11.7, 11.9)	0.45* (0.43, 0.47)	13.1 (12.9, 13.3)
ET (shuffled)	0.53* (0.50, 0.56)	15.1 (14.8, 15.4)	0.39 (0.36, 0.42)	13.3 (13.2, 13.4)
DS	0.49 (0.47, 0.52)	13.9 (13.7, 14.1)	0.44* (0.42, 0.48)	11.5 (11.4, 11.6)
DS (shuffled)	0.58* (0.55, 0.60)	12.6 (12.5, 12.8)	0.45* (0.42, 0.48)	11.0 (10.9, 11.1)
L-WISE	**0.60* (0.58, 0.64)**	**11.1 (11.0, 11.2)**	**0.47* (0.44, 0.50)**	**10.5 (10.4, 10.5)**

Table 1: **Both image enhancement tapering (ET) and image difficulty selection (DS) contribute to ability of L-WISE to assist learners.** The benefits of image enhancement are dependent on easy-to-hard sequencing ("ET" outperforms "ET (shuffled)"), but the benefits of difficulty-based selection appear to stem from simply showing an easier distribution of images during training ("DS (shuffled)" is as good as or better than "DS"). Training times are in minutes. Parentheses show 95% bootstrap confidence intervals for the mean with 10,000 bootstrap replicates. * denotes a significant difference in accuracy from the control group ($p < 0.01$, $\chi^2(1)$ test).

Our Logit-Weighted Image Selection and Enhancement (L-WISE) algorithm operates on image trial sequences used for training human participants on image classification tasks. The performance of each participant is subsequently evaluated in a testing phase, which includes randomly-selected, unmodified images (unaffected by L-WISE). During the early portion of the training phase, L-WISE randomly selects images from below a certain difficulty percentile that linearly increases as the training phase progresses. Selected images are enhanced at each trial during this period, within an ℓ_2 pixel budget ϵ that decreases in a stepwise-exponential fashion (Fig. 3C-D).

We tested L-WISE's efficacy at improving test-time accuracy of human learners in three challenging image category learning tasks (Figs. 3-4). Participants were randomly assigned to an L-WISE group or a control group (randomly-selected, non-enhanced images). L-WISE increased the average test-time accuracy margin above chance levels by 57.6% on a 4-way moth classification task ($p < 0.001$ on chi-squared test), by 72.3% on a 4-way skin lesion dermoscopy task ($p < 0.001$), and by 33.1% on a binary colon polyp histology task ($p = 0.023$) (Fig. 4A). In all three tasks, participant accuracy in the L-WISE group increased initially and then declined to varying degrees as more and more difficult images were selected and the degree of enhancement simultaneously reduced. In addition to improving test-time accuracy, L-WISE decreased the mean time to learn the task (with a fixed number of training trials) by 20% for the moth task, 23% for the dermoscopy task, and 22% for the histology task (Fig. 4A).

4.3 Image enhancement and selection both contribute to efficacy of L-WISE

We tested several ablated versions of L-WISE to determine the relative contributions and potential additive benefits of its components, which include (A) image enhancement based on logit optimization, (B) image selection based on logit-estimated difficulty, and (3) easy-to-hard curriculum trends enabled by A and B. The results of our ablation study, which employed the moth classification and skin lesion dermoscopy tasks, are shown in Table 1. In "Enhancement Tapering," (ET) only the enhancement component of L-WISE is used, with no difficulty-based image selection. Conversely, in "Difficulty Selection" (DS), images are not enhanced. In "ET (shuffled)" and "DS (shuffled)," the order of training trials is randomly permuted after either the ET or DS intervention. We hypothesized that shuffling would abolish at least some of the potential accuracy gains attributable to ET and DS due to the possible usefulness of easy-to-hard ordering.

The results show that both ET and DS have strong benefits in isolation. ET increased the test-phase accuracy margin above chance by 46.8% for the moth task and 56.5% for the dermoscopy task, while DS increased the same margin by 8.1% (not significant) and 53.2% respectively. ET (shuffled) was less effective, increasing the margin above chance by 23.0% for the moth task and 11.2% (not significant) for the dermoscopy task. Surprisingly, DS (shuffle) outperformed DS without shuffling, increasing the margin above chance by 45.2% for the moth task and 58.2% for the dermoscopy task (the increase of DS (shuffle) relative to DS is statistically significant for the moth task only). Unablated L-WISE numerically outperformed all ablated conditions, increasing the margin above chance by 57.6% for the moth task and 72.3% in the dermoscopy task. However, additional paired com-

parisons indicated that the differences between these increases and those from ET or DS (shuffled) are not statistically significant for either task. ET and DS did demonstrate a statistically significant additive benefit in terms of learning efficiency, however. On the moth task, training time was 6% shorter for L-WISE than the next fastest group, which was ET. Similarly, on the dermoscopy task, training time was 5% shorter for L-WISE than the next fastest group, DS (shuffled).

5 DISCUSSION

In this study, we demonstrate that the leading models of the ventral visual stream, the robustified ANNs, can be used to both predict the empirical image recognition difficulty as reported by humans, and generate "perturbed", enhanced, versions of it that are easier to categorize as the ground truth. We harness these properties to design a model-based curriculum design for human image category learning. We show that a combination of selecting images within a certain difficulty range and perturbing those images to enhance the perception of the ground truth category leads to substantial improvements in human test-time classification performance under realistic conditions.

The results of our ablation study show that at least a portion of these improvements can be achieved with image enhancement alone: humans can learn from distorted images and subsequently achieve superior generalization to unseen examples (Table 1). There are several possible explanations for this effect. Image enhancements might draw the learner's attention to relevant features, such as the distinctive dot in the middle of each wing of the *Idaea biselata* moth in Fig. 3B (Hufnagel, 1767), or the multiple colors and irregular borders that appear to be enhanced in the melanoma image of Fig. 4 (Tsao et al., 2015). Enhancements might also attenuate features that distract from or contradict the ground truth: for example, in Appendix Fig. S8B, buffalo standing behind the ground truth "antelope" are variously blurred or nearly erased. Analogously, images with high ground truth logits (low predicted difficulty) might tend to have clear class-relevant features and few distracting or contradictory features. These parallel explanations for the effects of image enhancement and image selection could help explain why the "enhancement tapering" and "difficulty selection" strategies in isolation provide comparable accuracy gains to each other, and why combining both strategies did not lead to large additive improvements in accuracy (although full L-WISE did enable significantly shorter training time than either strategy alone).

5.1 LIMITATIONS

This work is a proof-of-concept demonstration that robustified models can be applied to augment image category learning in humans. We did not exhaustively search for optimal curriculum design strategies image enhancement hyperparameters, nor did we study the "dose-dependency" of image enhancement or selection. L-WISE applies a fixed schedule for all learners: human learning could plausibly be boosted further by adapting the degree of image enhancement or the difficulty of selected images to the learner's progress in real time (Lu & Dosher, 2022; Mettler & Kellman, 2014).

One caveat to our approach is that logit maximization can sometimes appear to have a homogenizing effect on image distributions. For example, "benign mole" dermoscopy images enhanced with high ϵ budgets tend to all resemble smooth and uniform blobs (see appendix Fig. S9H for an example): this clearly illustrates a task-relevant difference from melanoma (which tends to be asymmetric with irregular borders (Tsao et al., 2015)), but obscures much of the real-world heterogeneity among benign moles. A similar risk might apply to image selection: images with high ground truth logits might belong to limited regions of the overall class distribution. Biased perturbations or selections reflecting biases in the underlying datasets are another concerning possibility, particularly for dermoscopy (Daneshjou et al., 2022). These caveats must be thoroughly investigated before real-world educational applications of L-WISE, particularly in the medical domain.

6 METHODS

Predicting Image Difficulty. To predict the relative difficulty $d \in [0, 1]$ of each image, we extract the logit value corresponding to the image's ground truth class (L_{gt}) immediately upstream of the final softmax function. We sort the logits in descending order such that $L_{(1)} \geq L_{(2)} \geq ... \geq L_{(n_{c_i,s})}$. $n_{c_i,s}$ is the number of images for a given class c_i and training/validation/testing category designation s. We calculate difficulty percentile d_j for image j using the equation $d_j = \text{rank}(L_{(j)})/n_{c_i,s}$.

Generating Perturbations to Enhance Images. To enhance an image using a pretrained ANN, we maximize L_{gt} through projected gradient ascent, onto a hypersphere of radius ϵ, in pixel-space (See

Appendix Section S1 for model training and projection details). In some cases, we also explicitly minimize the logit of competing classes. We generate perturbed image x' via the optimization:

$$x' = x + \underset{\|\delta\|<\epsilon}{\arg\max}\, L_{\text{gt}}(x+\delta) - \frac{\alpha}{c-1} \sum_{c_i \in C, c_i \neq \text{gt}} L_{c_i}(x+\delta) \tag{1}$$

In equation 1, δ is a perturbation tensor of the same dimensionality as x and an ℓ_2 norm less than pixel budget ϵ. L_{gt} is the ANN's logit score associated with the ground truth class, and L_{c_i} is the logit associated with class c_i (c is the total number of classes, C is the set of competing classes). α determines the extent to which logits for competing classes are minimized.We set $\alpha = 0$ for the ImageNet animal classification experiments and $\alpha = 1$ for the learning experiments.

Image classification and learning experiments with human participants. We recruited 521 human subjects using the online platform Prolific. We allowed subjects to participate in multiple experiments, but only once for each of the three learning tasks. We used the JsPsych library (De Leeuw, 2015) with the JsPsychPsychophsics plugin (Kuroki, 2021) for all experiments. All experiments included 10-12.5% attention check trials where the participant is asked to classify an image of a circle or triangle. We analyzed data from participants with $\geq 90\%$ attention check accuracy.

To measure the effects of enhancement on a presumably already-learned task, we tested the accuracy of human subjects at classifying 16 basic types of animals (frog, bird, dog, etc., see Appendix section S3). Images were shown for 17 milliseconds each, after which the participant was given 15 seconds to respond. All images were drawn from the validation set of ImageNet (Deng et al., 2009). Appendix Fig. S2 shows screenshots of the task as it appeared to participants. Each subject viewed interspersed images from 9 conditions: original images, images enhanced by maximizing L_{gt} with $\epsilon = 5, 10, 15$, and 20, images disrupted by minimizing L_{gt} with $\epsilon = 10$, and images enhanced with three baseline enhancement algorithms (see Fig. 1). Participants were notified after each incorrect response and given a small monetary bonus for each correct response (except for disrupted images). 62 participants viewed 18 images for each of 16 classes, 2 from each condition, in a shuffled ordering for a total of 288 trials per participant and 17,856 overall. These main trials followed a screening phase with 32 trials (200ms presentation times, 24 or more correct with at least one correct per class required to proceed, multiple screening attempts allowed) and a 32-trial warm-up phase with 17ms image presentations. Screening and warm-up phases used original, unmodified images, and data from these phases was not included in any analyses.

The image category learning tasks consisted of either 4 (moths, dermoscopy) or 2 (histology) image classes. Participants were shown each image for up to 10 seconds and used the mouse (4-way tasks) or keyboard (binary task, "F" and "J" keys) to respond. After each trial, the participant was notified of the ground truth label and whether their response was correct. (screenshots in appendix Fig. S3). Each session consisted of 8 training blocks of 16 trials each (4 per class, or 2 per class for histology), and 2 testing blocks of 20 trials each during which no post-trial feedback was provided. Each block contained an equal number of images from each class in random order. Participants were informed upon recruitment that they could receive a progressively higher monetary bonus if their test-phase accuracy exceeded certain thresholds. Before participating in the main learning tasks, subjects had to first learn an easier binary classification task (leatherback vs. loggerhead turtles) and respond correctly to at least 7 of 8 test-phase trials. We randomly assigned ~30 participants per experimental condition, with ~60 participants for ablation study control conditions (min. 27, max. 68).

Assisting learners with the L-WISE algorithm. L-WISE consists of two strategies applied in parallel: Enhancement Tapering (ET) and Difficulty Selection (DS). Both strategies operate only on images in the first 6 of 8 trial blocks in the training phase. In ET, we enhanced images in the first block of training-phase trials with $\epsilon = 8$. ϵ is halved for each subsequent block until it is set to 0 after the 6th block. In DS, only images with $d < d_b$ were sampled for each block b. d_b was incremented by 0.15 at the end of each of the first six blocks, beginning at $d_1 = 0.1$ and reaching $d_7, d_8 = 1.0$. Determining an effective schedule of ϵ and d_{max} did not require extensive hyperparameter tuning. After a pilot experiment in which we decreased ϵ linearly starting from $\epsilon = 20$, (see appendix Fig. S9H-M), we switched to the ϵ schedule above changed no other hyperparameters for any of the three learning tasks/image domains. In the "shuffled" versions of DS and ET (Section 4.3), all trials in the first 6 blocks are selected or enhanced before a constrained shuffling procedure, where each image may switch positions with any other of the same class regardless of d or ϵ.

7 ETHICS STATEMENT

This study involved experiments with human participants conducted over the internet, using the Prolific platform for the main experiments and Amazon Mechanical Turk for pilot experiments. We followed a study protocol approved by the ((anonymized IRB information)). Participants provided informed consent before participating in any experiments. The experiments posed no greater than minimal risk to the participants. All participants were anonymous, and all data is de-identified. Participants were provided with our contact information, and that of the ((anonymized research office of our institution)), for any questions or concerns about the study. We calibrated the participant compensation amounts for each experiment to meet or exceed the equivalent of $15.00 USD per hour, including during screening tasks. Participants were recruited using the "Standard Sample" option in Prolific, and were diverse in gender, age, and race/ethnicity (please see Table S2 for a demographic breakdown).

We hope that our work will eventually lead to practical and beneficial applications in education - for example, in the training of doctors in certain specialties such as pathology, radiology and dermatology where visual perceptual learning is an important part of clinical training. We wish to emphasize, however, that more work is needed before our methods can be safely applied in sensitive or high-stakes settings. For example, we apply our approach to improve human performance on a dermoscopy skin lesion classification task derived from the HAM10000 dataset (Tschandl et al., 2018). This dataset is heavily skewed towards images of pale skin, likely a reflection of the lower incidence of skin cancers such as melanoma among people with darker skin tones (Cormier et al., 2006). Models trained on this dataset are known to perform poorly for patients with darker skin (Daneshjou et al., 2022), where melanoma tends to have a different appearance, unfamiliarity with which on the part of clinicians contributes to delayed diagnosis and increased mortality among such patients (Thompson et al., 2023). It is plausible that maximizing the melanoma-associated logit of a robustified model perturbs the images to look more like an average presentation of melanoma (i.e., on light skin), which would risk imparting this bias onto the learner. A similar risk might apply to image selection: images with the highest groundtruth logits (which L-WISE presents at the beginning of learning) might tend to belong to specific subclasses or limited regions of the overall class distribution. The possibility of biased perturbations or image selections must be thoroughly investigated in future work before any applications of our methods in this domain.

8 REPRODUCIBILITY STATEMENT

We list all hyperparameters for training robustified neural networks and using them to enhance images in the appendix. Upon publication, we will provide source code that reproduces all data processing steps, figure plotting, and experiments. This includes experiments with human participants: we have developed a flexible framework for automated deployment of web-based image category learning experiments to a suite of cloud-based services, including hosting the task as an interactive web page and mechanisms for random group assignment and data collection. The experiments with human participants can therefore be reproduced and readily extended or modified with only a modest amount of configuration required.

REFERENCES

Adobe Inc. Adobe photoshop lightroom classic, 2024. URL https://www.adobe.com/products/photoshop-lightroom-classic.html. Version 13.5.1.

Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. *Advances in neural information processing systems*, 34:20014–20027, 2021.

Saeed Anwar, Salman Khan, and Nick Barnes. A deep journey into super-resolution: A survey. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020.

Robert Baldock, Hartmut Maennel, and Behnam Neyshabur. Deep learning through the lens of example difficulty. *Advances in Neural Information Processing Systems*, 34:10876–10889, 2021.

Dongliang Chang, Kaiyue Pang, Ruoyi Du, Yujun Tong, Yi-Zhe Song, Zhanyu Ma, and Jun Guo. Making a bird ai expert work for you and me. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12068–12084, 2023.

Joseph A Chapman and John EC Flux. Introduction to the lagomorpha. In *Lagomorph biology: evolution, ecology, and conservation*, pp. 1–9. Springer, 2008.

Minkyu Choi, Kuan Han, Xiaokai Wang, Yizhen Zhang, and Zhongming Liu. A dual-stream neural network explains the functional segregation of dorsal and ventral visual pathways in human brains. *Advances in Neural Information Processing Systems*, 36:50408–50428, 2023.

Janice N Cormier, Yan Xing, Meichun Ding, Jeffrey E Lee, Paul F Mansfield, Jeffrey E Gershenwald, Merrick I Ross, and Xianglin L Du. Ethnic differences among patients with cutaneous melanoma. *Archives of internal medicine*, 166(17):1907–1914, 2006.

Francesco Croce and Matthias Hein. Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-free Attacks. *International Conference on Machine Learning*, 2020.

Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances*, 8(31): eabq6147, 2022.

Joshua R De Leeuw. jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, 47:1–12, 2015.

Edoardo Debenedetti, Vikash Sehwag, and Prateek Mittal. A light recipe to train robust vision transformers. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 225–253. IEEE, 2023.

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*, pp. 248–255. Ieee, 2009.

Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019. URL https://github.com/MadryLab/robustness.

Thomas Fel, Ivan F Rodriguez Rodriguez, Drew Linsley, and Thomas Serre. Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in neural information processing systems*, 35:9432–9446, 2022.

Iuri Frosio and Jan Kautz. The best defense is a good offense: adversarial augmentation against adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4067–4076, 2023.

Roy Ganz, Bahjat Kawar, and Michael Elad. Do perceptually aligned gradients imply robustness? In *International Conference on Machine Learning*, pp. 10628–10648. PMLR, 2023.

Guy Gaziv, Michael Lee, and James J DiCarlo. Strong and precise modulation of human percepts via robustified anns. *Advances in Neural Information Processing Systems*, 36, 2024.

Chong Guo, Michael J Lee, Guillaume Leclerc, Joel Dapello, Yug Rao, Aleksander Madry, and James J Dicarlo. Adversarially trained neural representations are already as robust as biological neural representations. *International Conference on Machine Learning*, 2022.

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*, pp. 770–778, 2016.

Johann Siegfried Hufnagel. Tabelle von den nachtvögeln. *Berlinisches Magazin*, 1(4):618, 1767. URL http://resolver.sub.uni-goettingen.de/purl?PPN484874233_0004.

Ziheng Jiang, Chiyuan Zhang, Kunal Talwar, and Michael C Mozer. Characterizing structural regularities of labeled data in overparameterized models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5034–5044. PMLR, 18–24 Jul 2021.

Daniel J Jobson, Zia-ur Rahman, and Glenn A Woodell. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image processing*, 6(7):965–976, 1997.

Edward Johns, Oisin Mac Aodha, and Gabriel J Brostow. Becoming the expert-interactive multi-class machine teaching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2616–2624, 2015.

Juyeop Kim, Jun-Ho Choi, Soobeom Jang, and Jong-Seok Lee. Amicable aid: Perturbing images to improve classification performance. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Daiichiro Kuroki. A new jspsych plugin for psychophysics, providing accurate display duration and stimulus onset asynchrony. *Behavior Research Methods*, 53:301–310, 2021.

Weiyang Liu, Bo Dai, Ahmad Humayun, Charlene Tay, Chen Yu, Linda B Smith, James M Rehg, and Le Song. Iterative machine teaching. In *International Conference on Machine Learning*, pp. 2149–2158. PMLR, 2017.

Zhong-Lin Lu and Barbara Anne Dosher. Current directions in visual perceptual learning. *Nature reviews psychology*, 1(11):654–668, 2022.

Oisin Mac Aodha, Shihan Su, Yuxin Chen, Pietro Perona, and Yisong Yue. Teaching categories to human learners with visual explanations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3820–3828, 2018.

Aleksander Mądry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2018.

David Mayo, Jesse Cummings, Xinyu Lin, Dan Gutfreund, Boris Katz, and Andrei Barbu. How hard are computer vision datasets? calibrating dataset difficulty to viewing time. *Advances in Neural Information Processing Systems*, 36:11008–11036, 2023.

Everett Mettler and Philip J Kellman. Adaptive response-time-based category sequencing in perceptual learning. *Vision research*, 99:111–123, 2014.

Aleksander Mądry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, (6), 2018.

Ana Belén Petro, Catalina Sbert, and Jean-Michel Morel. Multiscale retinex. *Image processing on line*, pp. 71–88, 2014.

Yunliang Qi, Zhen Yang, Wenhao Sun, Meng Lou, Jing Lian, Wenwei Zhao, Xiangyu Deng, and Yide Ma. A comprehensive overview of image enhancement techniques. *Archives of Computational Methods in Engineering*, pp. 1–25, 2021.

Zeju Qiu, Weiyang Liu, Tim Z. Xiao, Zhen Liu, Umang Bhatt, Yucen Luo, Adrian Weller, and Bernhard Schölkopf. Iterative teaching by data hallucination. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 9892–9913. PMLR, 25–27 Apr 2023.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Hadi Salman, Andrew Ilyas, Logan Engstrom, Sai Vemprala, Aleksander Mądry, and Ashish Kapoor. Unadversarial examples: Designing objects for robust vision. *Advances in Neural Information Processing Systems*, 34:15270–15284, 2021.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, pp. 407007, 2018.

Adish Singla, Ilija Bogunovic, Gábor Bartók, Amin Karbasi, and Andreas Krause. Near-optimally teaching the crowd to classify. In *International Conference on Machine Learning*, pp. 154–162. PMLR, 2014.

Brandon Thompson, Toni Jenkins, John Paul Sánchez, Matthew Frederick, Alba Posligua-Alban, and Naiara Sbroggio Barbosa. Melanoma: Does it present differently in darker skin tones? *MedEdPORTAL*, 19:11311, 2023.

Hensin Tsao, Jeannette M Olazagasti, Kelly M Cordoro, Jerry D Brewer, Susan C Taylor, Jeremy S Bordeaux, Mary-Margaret Chren, Arthur J Sober, Connie Tegeler, Reva Bhushan, et al. Early detection of melanoma: reviewing the abcdes. *Journal of the American Academy of Dermatology*, 72(4):717–723, 2015.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

Jamalbek Tussupov, Kairat Kozhabai, Aigulim Bayegizova, Leila Kassenova, Zhanat Manbetova, Natalya Glazyrina, Mukhamedi Bersugir, and Miras Yeginbayev. Applying machine learning to improve a texture type image. *Eastern-European Journal of Enterprise Technologies*, 122(2), 2023.

Grant Van Horn and Oisin Mac Aodha. inat challenge 2021 - fgvc8, 2021. URL `https://kaggle.com/competitions/inaturalist-2021`.

Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12884–12893, 2021.

Pei Wang, Kabir Nagrecha, and Nuno Vasconcelos. Gradient-based algorithms for machine teaching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1387–1396, 2021a.

Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4555–4576, 2021b.

Jerry Wei, Arief Suriawinata, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Louis Vaickus, Charles Brown, Michael Baker, Naofumi Tomita, Lorenzo Torresani, et al. A petri dish for histopathology image analysis. In *Artificial Intelligence in Medicine: 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event, June 15–18, 2021, Proceedings*, pp. 11–24. Springer, 2021.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Xiaojin Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Karel Zuiderveld. Contrast limited adaptive histogram equalization. In *Graphics gems IV*, pp. 474–485. 1994.

Appendix

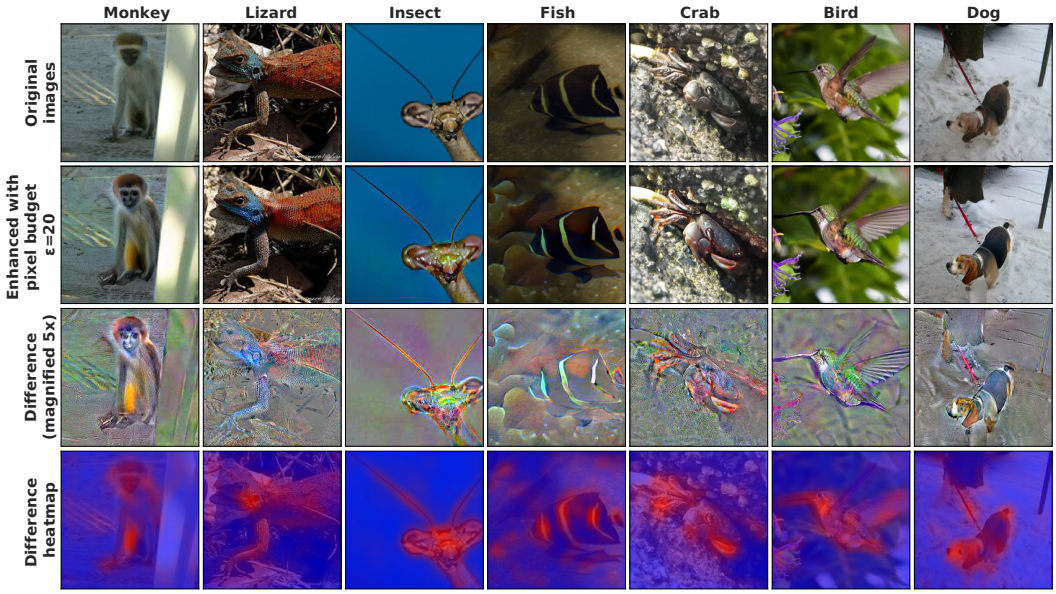

Figure S1: **Ground truth logit enhancement with robustified ANNs leads to semantically meaningful perturbations.** *The top row shows original ImageNet images, and the second row shows the same images after enhancement by robustified ResNet-50 (training ϵ=3) with a pixel budget of $\epsilon = 20$. The third row shows a 5x magnified version of the difference between the enhanced image and the original, and the bottom row shows a heat map where red regions correspond to larger changes and blue regions correspond to smaller changes.*

S1 Details on Training and Using Robustified Guide Models

We adversarially trained a ResNet-50 model on ImageNet (Deng et al., 2009), and another on iNaturalist 2021 (Van Horn & Mac Aodha, 2021), with the hyperparameters following Gaziv et al. (2024):

- Epochs = 200
- Base learning rate of 0.1, decreasing by a factor of 10 every 50 epochs
- Batch size = 256
- Weight decay = 0.0001
- Adversarial training $\epsilon = 3.0$ (ImageNet) or $\epsilon = 1.0$ (iNaturalist)
- 7 gradient steps for adversarial attacks
- Adversarial attack step size of 0.5 (ImageNet) or 0.3 (iNaturalist)

The ResNet-50 model adversarially trained on ImageNet was used directly to generate perturbations and difficulty rankings for the 16-way animal classification task, using the logits of the original, fine-grained ImageNet classes (i.e., not the 16 superclasses, "grasshopper" not "insect") for both enhancement and difficulty prediction. The same model was adversarially fine-tuned on the HAM10000 and MHIST datasets before their application (as part of L-WISE) to the dermoscopy and histology tasks respectively. Generally, we trained the models on all available classes in each dataset. For example, we fine-tuned on all 7 classes of the HAM10000 dermoscopy dataset (Tschandl et al., 2018), even though we only used 4 of them in the learning tasks. When enhancing the images, we include only classes that are part of the experimental tasks as competing classes to have their logits minimized (see main-text equation 1).

For the moth task, we adversarially fine-tuned the (adversarially) iNaturalist-pretrained model on the four moth classes to be used in the task (which are part of iNaturalist). We subjectively judged the perturbations from this fine-tuned model to be more compelling than those generated using the iNaturalist-pretrained model without fine-tuning, as the four classes of interest are among the 10,000 iNaturalist classes.

All fine-tuning used $\epsilon = 1.0$ for adversarial training. We tuned various hyperparameters (learning rate, data augmentation strategies, etc.) during fine-tuning to maximize validation set performance on the images to be used for the learning tasks. We fine-tuned the entire network end-to-end for each task.

Our choice of ϵ=3 for ImageNet pretraining follows Gaziv et al. (2024), who found this to be an optimal choice for perturbations that disrupt category perception (relative to ϵ=1 and $\epsilon = 10$). In practice, we found that the models were unable to learn finer-grained tasks with training-time adversarial perturbations as large as ϵ=3 (iNaturalist pretraining and fine-tuning on moth photos, dermoscopy images, and histology images) - therefore, we reverted to ϵ=1 for these settings.

To enhance images in a category-specific manner, we perform the optimization of Equation 1 (main text) in a series of steps using projected gradient descent (equation 2), where k denotes the optimization step, η the step size, and Proj_ϵ a projection onto a hypersphere of radius ϵ with x at its center. All enhancements with pixel budget ϵ use ceil(2ϵ) steps of $\eta = 0.5$ in a $224 \times 224 \times 3$ pixel space.

$$\delta_{k+1} = \text{Proj}_\epsilon\left(\delta_k - \eta\nabla_\delta\big(L_{\text{gt}}(x+\delta_k) - \frac{\alpha}{c-1}\sum_{c_i \in C, c_i \neq \text{gt}} L_{c_i}(x+\delta_k)\big)\right) \tag{2}$$

Fig. S1 shows several example images enhanced with $\epsilon = 20$ using this approach by $\epsilon = 3$ adversarially-pretrained ResNet-50, along with difference images and heat maps produced by the same method as Figs. 3-4 in the main text.

S2 Details on image preparation for experiments

All images presented in all experiments were of size $224 \times 224 \times 3$, matching the input dimensions of ResNet-50. Before presentation or any model-based enhancement or difficulty prediction, original images were resized such that the shortest dimension (width or height) was 224 pixels, and then center-cropped to 224×224. Any single-channel grayscale images were converted to RGB before further processing. The baseline enhancement algorithms Contrast-Limited Adaptive Histogram Equalization (CLAHE (Zuiderveld, 1994)), Multi-Scale Retinex with Color Restoration (MSRCR (Jobson et al., 1997; Petro et al., 2014)), and the "Auto" image tuning feature in Adobe Photoshop Lightroom (Auto-LR (Adobe Inc., 2024)) were applied before the resizing and center-cropping operations.

We generally used images from the validation sets of each dataset for the image category learning experiments, reasoning that the robustified models would be overfitted to training images which could potentially compromise the quality of perturbations and relative difficulty estimates. However, for the moth task, we were limited to 10 validation images per class in the iNaturalist dataset (Van Horn & Mac Aodha, 2021). In this case we used training set images during the training period of the human image category learning experiment and validation images during the test phase. We show that image enhancements are still effective for training set images in Fig. S8A.

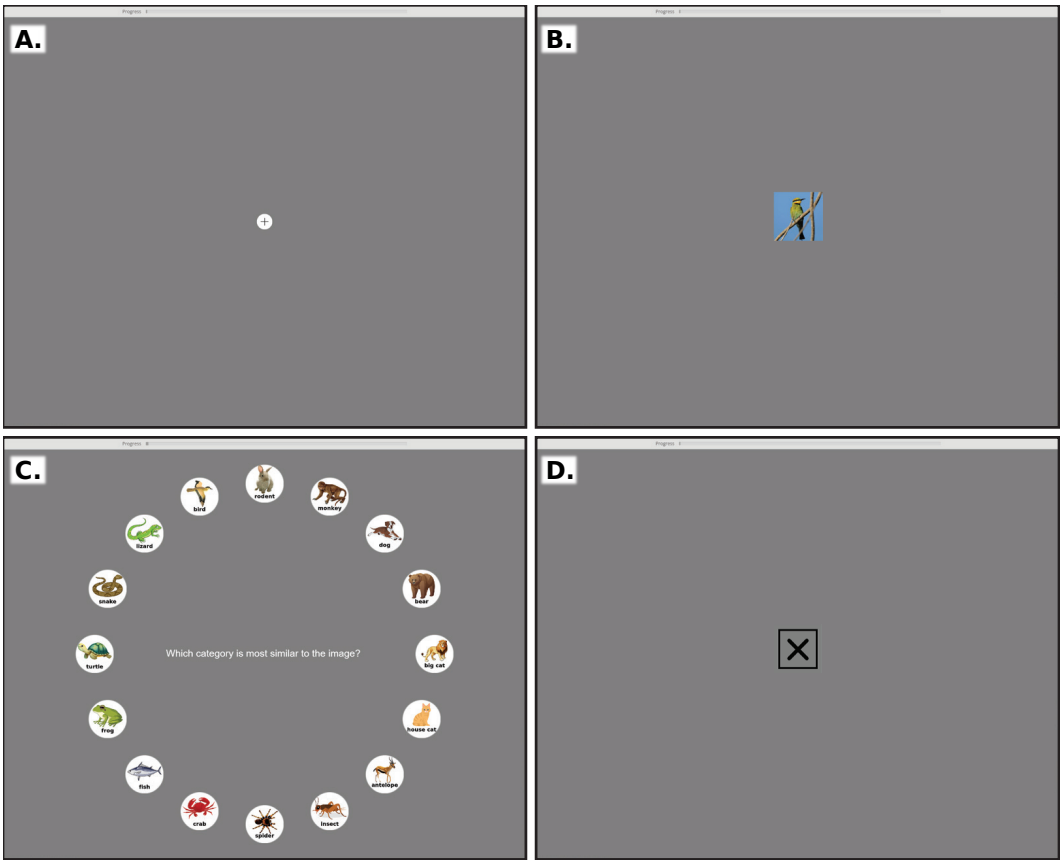

Figure S2: **Task interface for ImageNet animal classification with human participants.** *Subjects classified images among 16 categories. During each trial, the subject clicks the fixation cross (panel **A**) and the image is flashed for 17 milliseconds (panel **B**) with a 200ms presentation of a blank screen immediately before and after. The mouse cursor is hidden during the image presentation. Images are presented such that they subtend approximately 6 degrees of visual angle, with calibration for each participant using a blind-spot calibration procedure. After viewing the image, the participant clicks one of 16 buttons shown in panel **C**, which are randomly rotated in position every trial, within 15 seconds. For incorrect responses (except on images that are disrupted via ground truth logit minimization), or if 15 seconds elapses without a response, the participant is shown the black X for one second (panel **D**). Otherwise, no explicit feedback is given and the next trial begins immediately. Attention check trials featuring an image of a circle or triangle (see Fig. S3D for an example image) were interspersed with the main trials. For the attention check trials, two of the animal icons in panel **C** were randomly selected to be replaced with circle and triangle icons.*

S3 Details on 16-way ImageNet animal categorization experiments

We curated 16 sets of ImageNet classes corresponding to 16 basic animal superclasses for our basic animal classification experiments (Figs. 1-S2), adapting and expanding the Restricted ImageNet dataset defined in the Robustness library (Engstrom et al., 2019). The assignment of specific animal classes to each superclass is listed below:

- Dog: classes 151–268
- House Cat: classes 281–285
- Frog: classes 30–32
- Turtle: classes 33–37
- Bird: classes 80–100 and 127–146

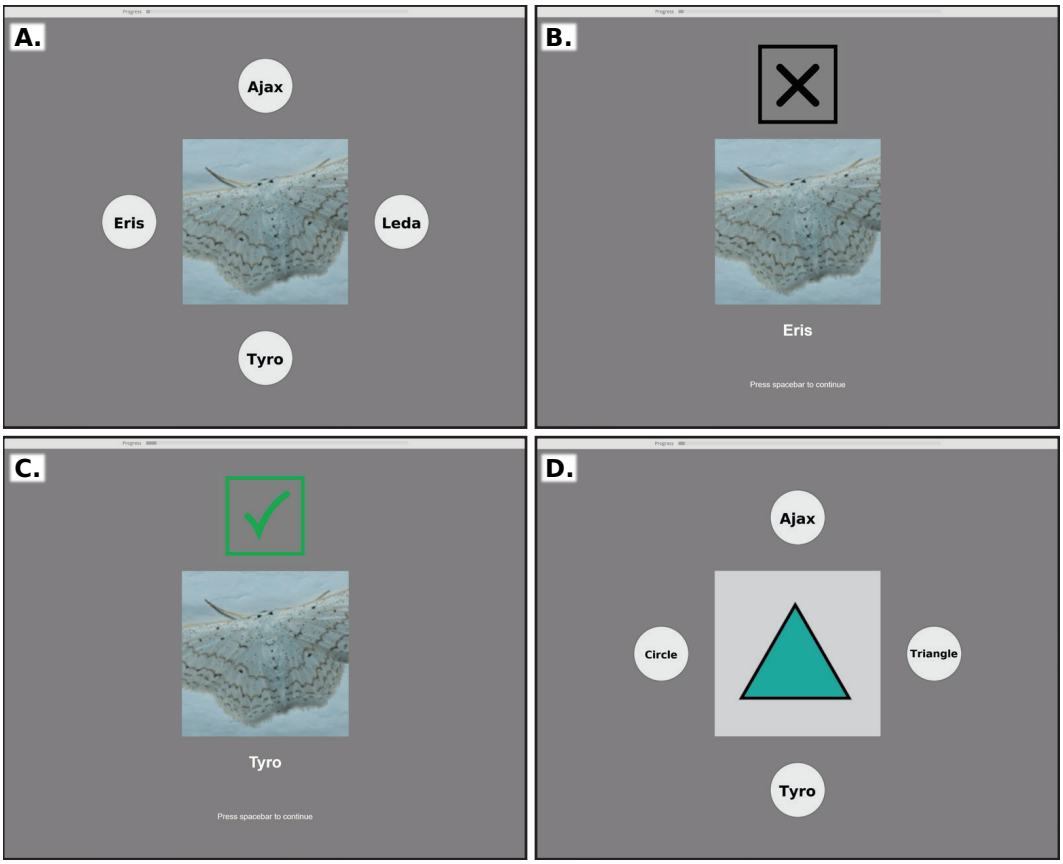

Figure S3: **Task interface for image category learning experiments.** *In the 4-way image category learning experiments (moths, dermoscopy), human subjects learned to classify four types of images that were represented by randomly assigned aliases "Ajax," "Eris," "Leda," and "Tyro." The image was shown for up to 10 seconds, during which the participant could click on one of the four buttons (panel **A**). Participants are shown a black X (1.5 seconds) immediately following an incorrect response or a >10s timeout (panel **B**), or a green check after a correct response (panel **C**). The alias corresponding to the correct class is also displayed on the feedback screen. Panel **D** shows an example of an attention check trial.*

- Monkey: classes 369–382
- Fish: classes 0, 1, 389, 391, 392, 393, 394, 395, 396, 397
- Crab: classes 118–121
- Insect: classes 300–320
- Lizard: classes 38–48
- Snake: classes 52–68
- Spider: classes 72–77
- Big Cat: classes 286–293
- Bear: classes 294–297
- Rodent: classes 330, 331, 332, 333, 335, 336, 338
- Antelope: classes 351–353

We excluded certain classes on a case-by-case basis in an attempt to minimize errors due to misunderstanding the animal categories as opposed to errors of visual perception. For example, we did not include porcupines, hedgehogs, or beavers in the "rodent" class (as many people may not recognize these as rodents), and we do not include eels in the "fish" class due to the possibility of

confusion with snakes. Our classification of rabbits and hares as "rodents" is technically incorrect, as they were reclassified to the order Lagomorpha in 1912 Chapman & Flux (2008) (we thank the participant who notified us of this).

We mistakenly classified rabbits and hares as "rodents," although they were reclassified to the order Lagomorpha in 1912.

S4 Details on image category learning experiments

Figs. S2 and S3 show the task interfaces for the 16-way animal classification and 4-way image category learning experiments with human participants, respectively. The positions of the four buttons used to indicate responses are randomly permuted for each participant.

To minimize biases stemming from prior declarative knowledge of the tasks or classes, for each participant in the 4-category learning tasks we randomly assign a four-letter, two-syllable name from the set "Ajax," "Eris," "Leda," and "Tyro" drawn from Greek mythology, each having four letters, two syllables, two consonants, and two phonetic vowels. We found no evidence that associating certain categories with certain aliases consistently affected test-phase accuracy (see Figs. S12 and S13).

We used a different approach for the binary histology task that employed the MHIST dataset (Wei et al., 2021), giving benign hyperplostic polyp the alias "benign" and sessile serrated adenoma the alias "malignant" (although sessile serrated adenoma is really a pre-cancerous legion). The histology task has an interface very similar in appearance to that of the 4-way tasks (as shown in Fig. S3), except that the "benign" and "malignant" buttons always appear either on side of the presented image (in a random order for each participant), and the participant responds by pressing the F key for the left-hand category or J for the right instead of clicking one of the buttons.

S5 Predicting image difficulty using ground truth logit of a robust model, compared with prior state-of-the-art approaches

In the L-WISE algorithm, we predict the difficulty of each image using its ground truth logit representation (L_{gt}) from a robustified ANN such as ResNet-50. We conducted an additional experiment to compare the ground truth logit with prior state-of-the-art predictors of image difficulty for humans established by Mayo et al. (2023). We apply logistic regression to predict correct v.s. incorrect responses to each image across all participants in our 16-way ImageNet animal classification experiment, using (1) c-score (approximated by the epoch during training at which an image is first correctly predicted (Jiang et al., 2021)), (2) prediction depth (earliest layer upon which a linear probe makes the same prediction as the final output (Baldock et al., 2021)), (3) image-level adversarial robustness (minimum magnitude of image perturbation required to change the network's prediction), and (4) ground truth logit from both (A) vanilla and (B) robustified ResNet-50 models (see Fig. S4). C-score, prediction depth, and adversarial robustness are implemented following Mayo et al. (2023). The results show that the ground truth logit from robust ResNet-50 (L_{gt}), the metric we use in L-WISE, significantly outperforms all other predictors of image difficulty, including all other metrics combined into one model ("Combined w/o L_{gt}" in Fig. S4). Furthermore, combining all other metrics with L_{gt} does not improve performance beyond L_{gt} alone. We also find that using a robustified model greatly improves the predictivity of L_{gt} and (marginally) adversarial robustness, but not of the c-score or prediction depth.

S6 Comparison of different ANN guide models for difficulty prediction and image enhancement

Our main results use robustified ResNet-50 as a guide model for perturbations. To evaluate the importance of the choice of guide model, we compared the accuracy of difficulty prediction (Fig. S5) and the effects of enhancement with $eps = 20$ (Fig. S6) using 6 different guide models in the 16-way animal classification task. The results show that setting ϵ to a value of 3 during adversarial

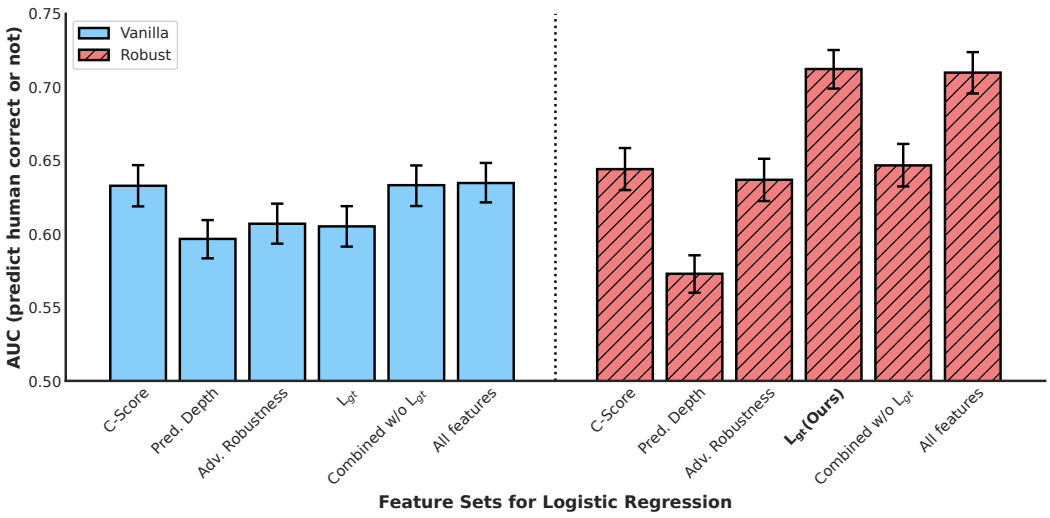

Figure S4: **Robustified ground truth logit is a state-of-the-art predictor of image difficulty for humans, outperforming the c-score, prediction depth, and adversarial epsilon of both vanilla and robustified models.** *AUC estimates are based on fitted logistic regression models using one or more features listed under each bar. Error bars are 95% confidence intervals for the mean from 10,000 bootstrap replicates. The chance level is AUC=0.5.*

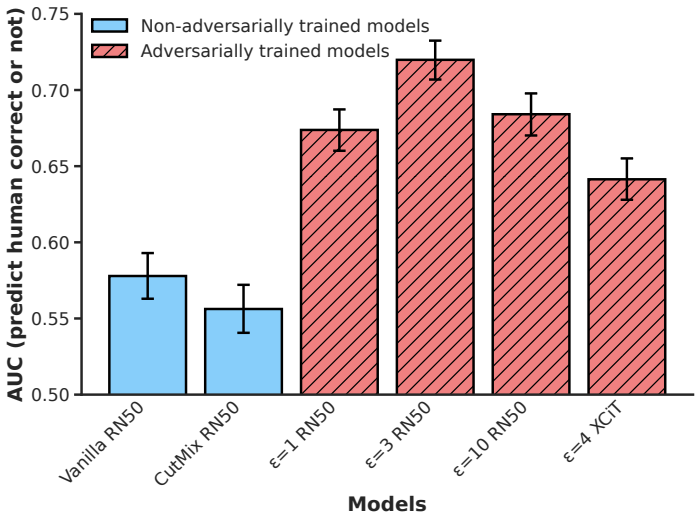

Figure S5: **Accuracy of image difficulty prediction using ground truth logits from different model types.** *AUC estimates and 95% confidence interval error bars are generated by the same procedure as in Fig. S4 - however, results are not directly comparable between the two figures as different ANN training runs were used for consistency within each experiment. RN50 = ResNet-50, and ϵ values in the labels for each bar show the magnitude of the adversarial perturbations during adversarial training. The chance level is AUC=0.5.*

training of ResNet-50 yields more accurate difficulty predictions and more effective perturbations than $\epsilon = 1$ or $\epsilon = 10$ training (consistent with disruption modulation results in Gaziv et al. (2024)), while perturbations guided by a non-adversarially-trained "vanilla" model have negligible effects. Although training ResNet-50 with CutMix improves its robustness to adversarial perturbations (Yun et al., 2019), CutMix-ResNet-50 does not outperform vanilla ResNet-50 in image difficulty prediction and perturbations using it as a guide model do not significantly increase accuracy beyond that on original images. In addition to ResNet-50, we tested the difficulty prediction and image

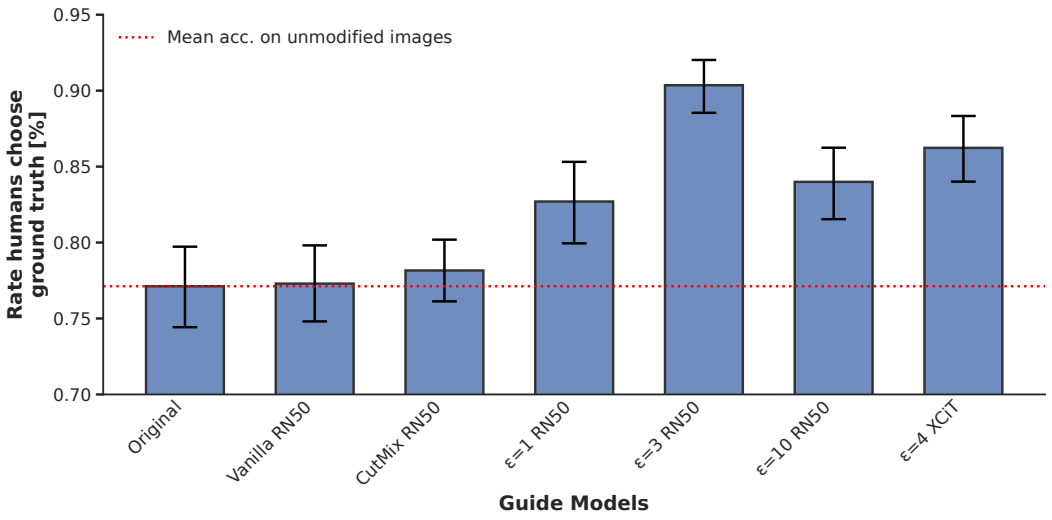

Figure S6: **Effectiveness of image category enhancement across different guide model types.** *Each bar shows the mean and 95% confidence interval (by bootstrap) of the rate at which humans choose the original ground truth label, in a 16-way basic animal classification task using ImageNet images. The "Original" bar shows accuracy for unmodified images, and other bars show accuracy of the same participants on images enhanced ($\epsilon = 20$) using gradients from the corresponding guide model. RN50 = ResNet-50, and ϵ values in the labels for each bar show the magnitude of the adversarial perturbations during adversarial training.*

enhancement capabilities of an adversarially trained vision transformer model, XCiT (Ali et al., 2021). Debenedetti et al. (2023) showed that the XCiT architecture is more suitable for adversarial training than the original vision transformer. XCiT generates reasonably accurate image difficulty predictions (on par with the previous state-of-the-art) and generates image perturbations that increase human categorization accuracy by a comparable degree to robustified ResNet-50. For the experiments in Figs. S5 and S6, we used pretrained guide models provided by Gaziv et al. (2024) (Vanilla, $\epsilon = 1$, $\epsilon = 3$, and $\epsilon = 10$ ResNet-50 models), Yun et al. (2019) (CutMix ResNet-50), and Debenedetti et al. (2023) ($\epsilon = 4$ XCiT). Examples of images enhanced by each of these guide models with $\epsilon = 20$ are displayed in Fig. S7.

S7 Ablation study on logit maximization approach to enhancement

To conduct a limited ablation study on our approach to image enhancement, we conducted an additional 16-way ImageNet animal classification experiment with 20 human participants. This experiment was mostly identical to the 16-way animal classification experiment described in the main text, except there were 6 image conditions instead of 9. Half of the trials used images from the ImageNet validation set (as in the main experiment), and the other half from the training set. Within each training/validation split, one third of the trials were original, unmodified images, one-third were enhanced by maximizing the ground truth logit with ℓ_2 pixel budget $\epsilon = 10$, and one-third were enhanced by minimizing the cross-entropy loss with $\epsilon = 10$. Results of this experiment are summarized in Fig. S8A. We hypothesized that logit-based enhancement would provide superior results, particularly for images that started off with low cross-entropy loss. We further hypothesized that enhancements would be less effective for training images due to overfitting of the guide model on them. The results show that logit maximization is effective on both training and validation images, and induces a higher increase in accuracy for a given pixel budget ϵ than cross-entropy minimization. Indeed, cross-entropy minimization significantly increased accuracy only for validation images and not for training images. Unexpectedly, participants were more accurate on original, unmodified training set images than on original, unmodified validation set images. According to Russakovsky et al. (2015), the ImageNet ILSVRC 2012 validation set was collected using the same methodology as the training set, but at a later time. It is therefore plausible that the images and labels in the

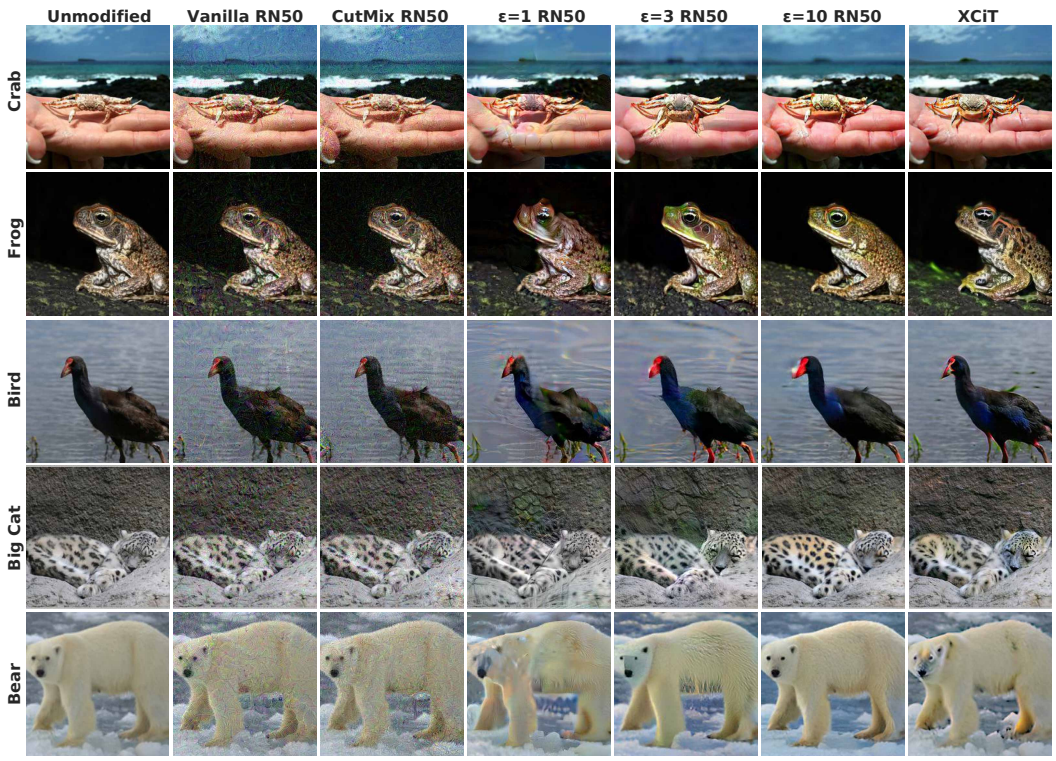

Figure S7: **Meaningful perturbations require robust models, and are possible with CNN and vision transformer architectures.** *Each row shows an image from ImageNet (original on the far left) enhanced with $\epsilon = 20$ by different guide models. A quantitative comparison of different models' perturbation efficacies with regards to improving human classification accuracy can be found in Appendix Fig. S6.*

validation set are drawn from a slightly different distribution than those in the training set, resulting in this accuracy discrepancy.

S8 Additional results from image category learning experiments

Fig. 3 in the main text shows learning curves (mean accuracy by condition as a function of trial number), and schedules for image difficulty selection and enhancement ϵ, for the moth photograph task: similar plots are shown for the dermoscopy task in Fig. S9 and the histology task in Fig. S10. Panels H-M of Fig. S9 show the results of an early pilot experiment that used image enhancement in isolation (no difficulty selection), in which we suspect the perturbation magnitude ϵ was set too high causing participants to learn exaggerated features and fail to generalize to natural images with subtler features. This prompted us to switch to the ϵ schedule we used for our main learning experiments, which starts at $\epsilon = 8$ instead of $\epsilon = 20$. Panel C of Fig. S10 shows the relationship between the ground truth logit from robust ResNet-50 model and how many of the 7 expert annotators of the MHIST histology dataset Wei et al. (2021) agreed on the same category label: on average, the model is more "confident" in its predictions on images where experts agree to a larger extent.

In addition to the agreement of expert MHIST annotators, the ground truth logit successfully predicts the proportion of human participants who select the correct ground truth label across all tasks we tested. Difficulty prediction results from the 16-way ImageNet task are shown in Fig. 1 in the main text, and from the moth photograph, dermoscopy, and histology tasks in Fig. S11 (Panels A1, B1, and C). For the non-ImageNet tasks, we rely on test-phase data from control group participants who had just learned the tasks in question.

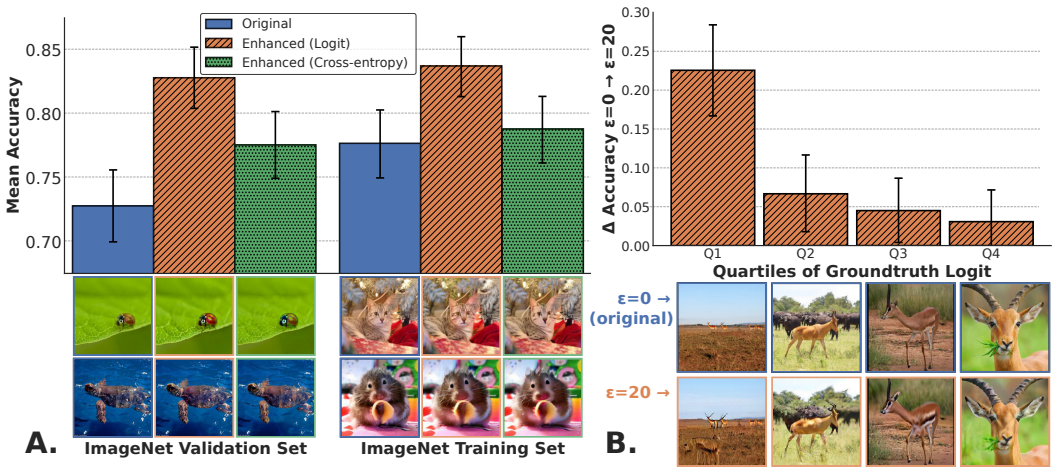

Figure S8: **Ablation results for image enhancement with ImageNet images.** *Logit maximization enhancement is effective for images used to train the robustified CNN used as a guide model, and also for held-out validation images (panel **A**). Logit-max enhancement is more efficient at increasing human accuracy within a given pixel budget ($\epsilon = 10$) than enhancement by cross entropy minimization (panel **A**). The efficacy of logit-max enhancement depends on the difficulty of the original image as estimated by the starting ground truth logit (panel **B**). In the bar plot of panel **B**, images were assigned to 4 quadrants based on their ground truth logit values, and for each quadrant the mean difference in accuracy was calculated between original, unmodified images and images enhanced with $\epsilon = 10$ (using data from the main 16-way animal classification experiment). The images below each bar illustrate an example image from the category "antelope" drawn from the corresponding difficulty quadrant. All error bars are 95% confidence intervals for the mean from 10,000 bootstrap replicates.*

We can also attempt to measure the extent to which images with higher levels of enhancement are easier for novice participants to recognize during the learning tasks (Fig. S11A2,B2). This analysis is limited to the first training trial blocks in the "ET (shuffled)" participant group in the ablation study (main-text Table 1), the only group that viewed enhanced images without monotonically decreasing ϵ. Note that there were 6 discrete ϵ values (1 per block in the non-shuffled ET condition), and the analysis is complicated by the fact that participants were still learning the task when they made the responses underlying these plots. We are also unable to compare with $\epsilon = 0$ unmodified images because participants did not view new unmodified images in the corresponding training blocks. There are no results here for the histology task because the ablation study was conducted only for moth photos and dermoscopy images. In the dermoscopy task, participants respond with the original, correct category label statistically significantly more often when viewing images enhanced with greater ϵ (Fig. S11B2). A similar trend was not statistically significant for the moth photograph task (Fig. S11B1), perhaps due to the limitations of this particular analysis outlined above. We can observe strong effects of image enhancement in both the moth photograph and dermoscopy tasks by examining the enhancement only ("Enhancement Taper/ET") arm of the ablation study (Table 1 in the main text): for both tasks, participants in the ET condition had higher test-phase accuracy than participants in the control conditions, and these differences were statistically significant.

To evaluate whether L-WISE has differential effects on human image category learning depending on the image class, we record test-phase precision and recall for each class among L-WISE and control groups in Table S1. The same data are visualized in Fig. 4B. Our experiments are statistically underpowered to detect class-specific differences in performance (as opposed to aggregated performance) - however, we can observe in a coarse sense that the sample means of precision and recall are numerically higher in the L-WISE group across all classes in all tasks. This suggests that overall accuracy improvements attributed to L-WISE are distributed among the various image classes, rather than being the result of isolated improvements in the detection of certain classes.

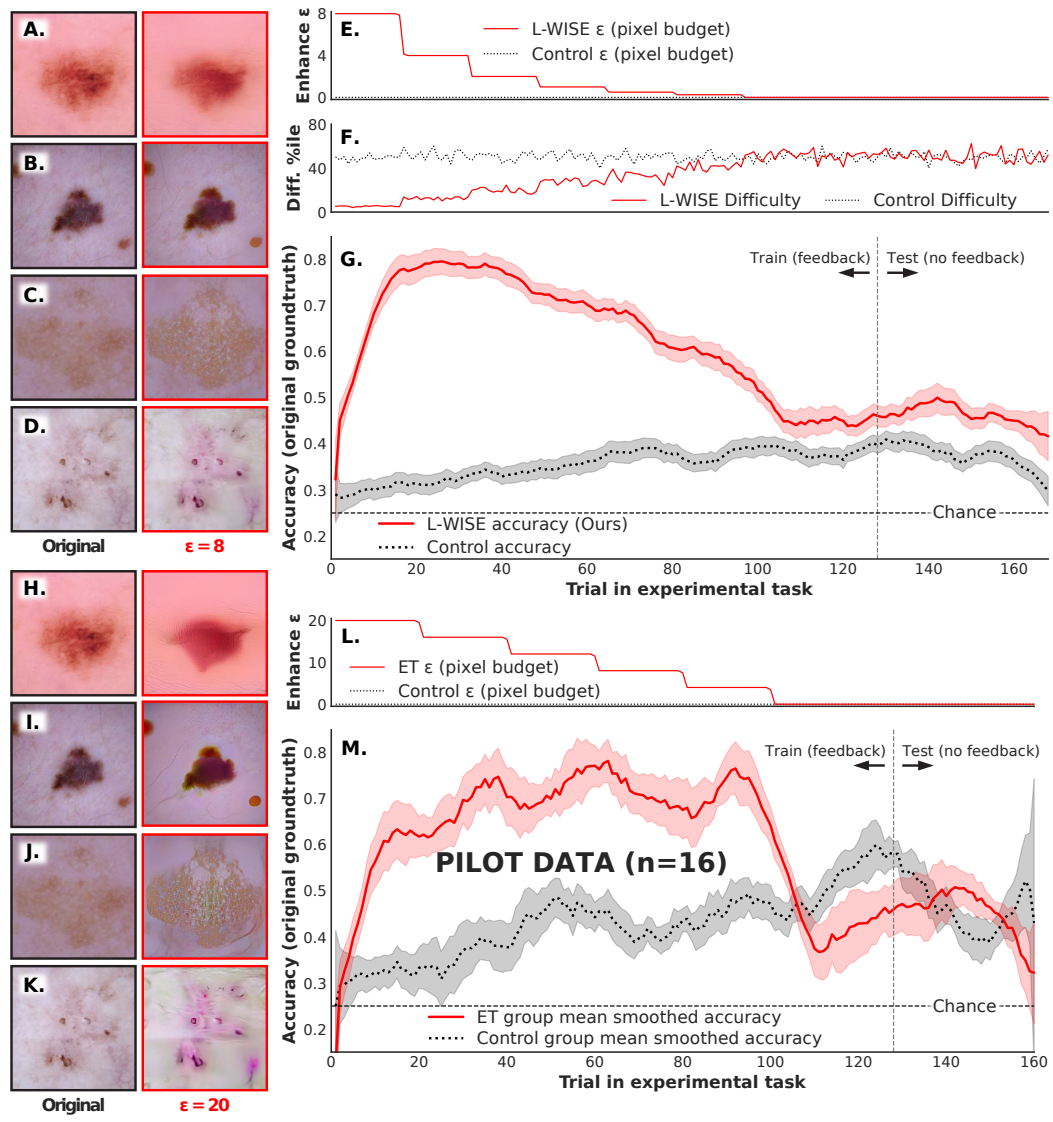

Figure S9: **Plots showing the accuracy trajectory of human participants throughout training/testing in the main dermoscopy learning experiment (panels A-G) and a preceding pilot experiment after which the epsilon tapering schedule was adjusted (panels H-M).** *All conventions are identical to main-text Fig. 3. There is a statistically significant difference between the test-phase performance of the L-WISE participants and that of the control participants (chi-square test, $p < 0.001$) in panel **G** but not for the pilot experiment in panel **M**. Notably, the last portion of the training phase does not feature any image enhancements (see Fig. 2F): we suspect that this is the reason for the sudden decline in accuracy in the enhancement group of the pilot experiment (**M**)*

S9 Participant dropout rates are lower when L-WISE assistance is provided

On the Prolific platform where we ran our experiments, participants can choose to withdraw from studies partway through if they no longer wish to participate (this is called "returning" a study in the Prolific interface). For the moth photograph and dermoscopy image category learning tasks, participants who received L-WISE assistance in full or partially ablated form (see Table 1) were less likely to withdraw.

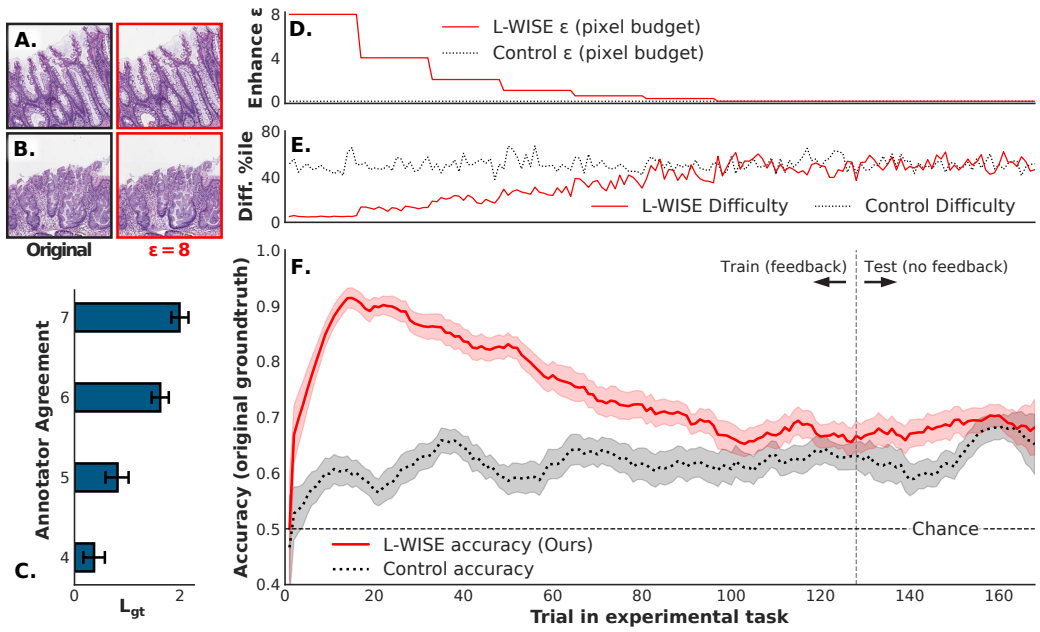

Figure S10: **Plot showing the accuracy trajectory throughout training/testing of human participants in the histology learning experiment.** *All conventions follow Fig. 3. Panel* ***C*** *shows the association between agreement among the 7 expert pathologist annotators of the MHIST dataset (Wei et al., 2021) and the ground truth logit score of each image from a robustified ResNet-50. Possible values of annotator agreement are 4, 5, 6, or 7 of the annotators agreeing with each other (3 and below switches the "ground truth" category). Error bars are 95% confidence intervals for the mean from 10,000 bootstrap samples.*

	Precision		Recall	
	Control	L-WISE	Control	L-WISE
Moth photos				
seriata	0.46 (0.39–0.52)	**0.52** (0.45–0.59)	0.43 (0.37–0.50)	**0.63** (0.55–0.71)
tacturata	0.45 (0.40–0.50)	**0.63** (0.55–0.71)	0.54 (0.47–0.61)	**0.68** (0.60–0.76)
biselata	0.35 (0.30–0.41)	**0.41** (0.32–0.50)	0.36 (0.31–0.42)	**0.42** (0.35–0.50)
aversata	0.50 (0.44–0.56)	**0.58** (0.50–0.66)	0.57 (0.50–0.63)	**0.68** (0.59–0.77)
Dermoscopy				
Benign mole	0.38 (0.33–0.43)	**0.42** (0.36–0.48)	0.43 (0.38–0.48)	**0.47** (0.40–0.54)
Melanoma	0.33 (0.29–0.38)	**0.39** (0.35–0.44)	0.37 (0.31–0.42)	**0.42** (0.37–0.47)
BCC	0.41 (0.36–0.46)	**0.54** (0.46–0.61)	0.41 (0.36–0.47)	**0.56** (0.48–0.63)
Benign keratosis	0.26 (0.22–0.30)	**0.39** (0.34–0.43)	0.30 (0.25–0.35)	**0.43** (0.36–0.49)
Histology				
SSL (malignant)	0.58 (0.54–0.62)	**0.60** (0.56–0.64)	0.66 (0.61–0.71)	**0.70** (0.66–0.75)
HP (benign)	0.59 (0.55–0.64)	**0.61** (0.56–0.67)	0.61 (0.56–0.66)	**0.66** (0.62–0.70)

Table S1: **L-WISE improves test-phase precision and recall across all image classes in three image category learning tasks.** *BCC=basal cell carcinoma, SSL=sessile serrated adenoma, and HP=hyperplastic polyp. In parentheses are 95% confidence intervals for the mean from 10,000 bootstrap replicates, resampling from participant-wise precision and recall values.*

Nine participants withdrew from the moth photograph category learning experiment. Among them, six had been assigned to the control group, one to the "enhancement taper" group, one to the "difficulty selection" group, and one to the full L-WISE group. We can calculate the probability of $d = 6$ or more participants among the $n = 9$ who withdrew being from the control group, under the null hypothesis that the probability of withdrawal is independent of group assignment, using the bino-

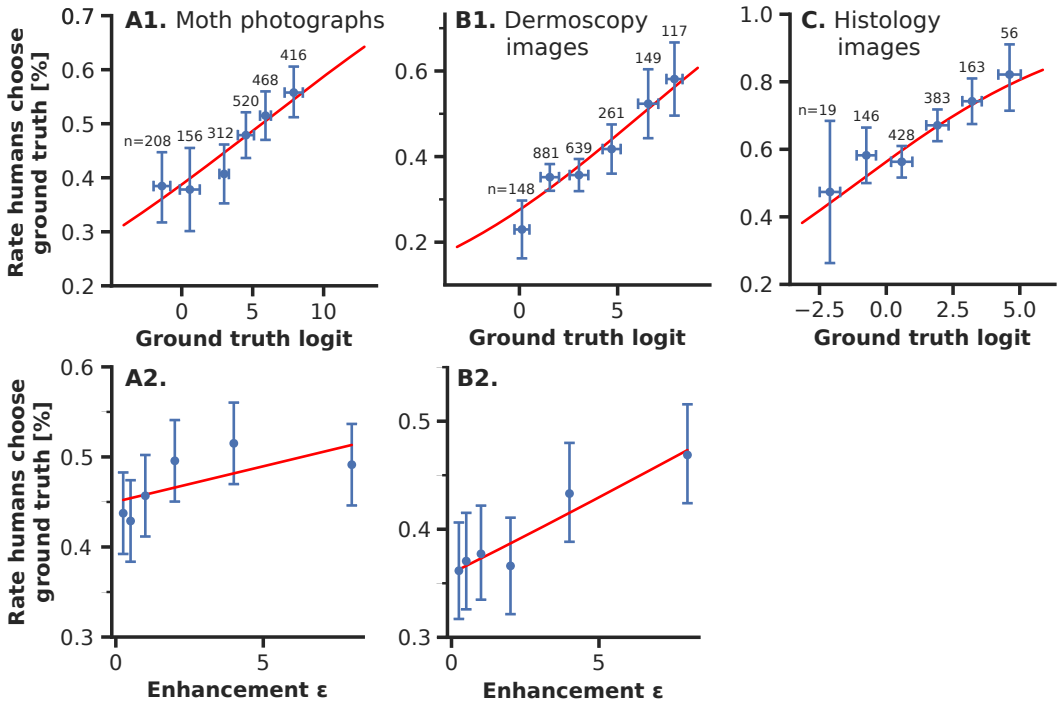

Figure S11: **Difficulty prediction and image enhancemement are effective across image domains.** *Panels **A1**, **B2**, and **C** show the relationship between the ground truth logit from a fine-tuned robustufied ResNet-50 model and the rate at which human participants (from the control group of the ablation study, see Table 1 in the main text) choose the ground truth label during the test phase following a training phase in which they had just attempted to learn the task. Images are binned by ground truth logit to produce the scatter plots, with the number of total trials listed for each bin. Vertical error bars are 95% confidence intervals by bootstrap, and horizontal error bars show the standard deviation. Red lines illustrate fitted logistic regression models. All logistic regression models had statistically significant coefficients for ground truth logit ($p < 0.01$). Panels **A2** and **B2** show the relationship between enhancement ϵ and the rate at which humans choose the ground truth category. This analysis is limited to the first training trial blocks in the "ET (shuffled)" participant group in the ablation study (main-text Table 1), the only group that viewed enhanced images without monotonically decreasing ϵ. The logistic regression coefficient for ϵ was statistically significant for dermoscopy images (**B2**, $p = 0.003$) but not for moth photographs (**A2**, $p = 0.14$).*

mial distribution via Equation 3 below (where p is the probability of being assigned to the control group). Equation 3 evaluates here to a probability of 0.02, indicating that participants who withdrew were significantly more likely to have been assigned to the control group than would be expected if L-WISE assistance had no impact on the probability of withdrawal.

$$P(X \geq d) = 1 - P(X \leq d-1) = 1 - \sum_{k=0}^{d-1} \binom{n}{k} p^k (1-p)^{n-k} \tag{3}$$

Similarly, in the dermoscopy category learning experiment, 13 participants withdrew, of whom 6 were from the control group. In this case, Equation 3 evaluates to a probability of 0.041, again indicating that participants in the control group withdrew at a significantly higher-than-expected rate. Furthermore, 4 more of the 13 withdrawals were from the "Enhancement Taper (shuffled)" group, which had test-phase accuracy indistinguishable from the control group (see Table 1). None of the withdrawals from the dermoscopy experiment were from the full L-WISE group.

Overall, these results show that participants were more likely to withdraw from the study when they did not receive assistance from L-WISE, perhaps reflecting the difficult nature of the moth

photograph and dermoscopy image tasks at baseline. None of the participants withdrew from the histology image experiment, precluding a similar analysis.

S10 NOTES ON "HALLUCINATIONS" IN ENHANCED IMAGES

To support our approach to assisting human learners, we demonstrate the ability to enhance category perceptions in response to images using low-norm perturbations. Previous work by Gaziv et al. (2024) showed that an image from one category can be perturbed in a targeted way such that a human perceives it to belong to a different category. Features introduced by these perturbations could be described as "hallucinations:" perceptions (by the model and the human viewer) of objects that are typically not present in the camera's view. Our image enhancement approach is a special case in the wider realm of categorically targeted image modulation, in which maximization of the ground truth logit perturbs the image such that it becomes a stronger and/or less ambiguous example of its class according to the model's judgement. Do these perturbations accentuate features that are already present such that they are easier for humans to perceive under challenging conditions, or do they improve human accuracy by hallucinating new features associated with the target category? Subjectively, both phenomena seem to occur: panels **C** and **D** in Fig. 1 appear to show bolder contrasts and (in panel **C**) even the appearance of better camera focus in class-relevant regions of the perturbed images. Panel **B** in appendix figure Fig. S8 shows a clear example of hallucination, where a semblance of an entire additional "antelope" appears in the foreground of the image. This distinction may be important for education-oriented applications of our enhancement approach, as hallucinations could plausibly impart potentially misleading information to the learner. On the other hand, it is possible that hallucinated features can impart useful and therefore desirable representations of the ground truth class despite departures from a natural image distribution.

S11 PARTICIPANT RECRUITMENT AND DEMOGRAPHICS

We recruited a grand total of 521 participants via the online platform Prolific. All participants live in the United States and are fluent in English (as determined by Prolific). Each participant was eligible to complete each learning experiment only once, to avoid collecting data from participants already familiar with the task.

Our decision regarding the number of participants to recruit for each learning task experimental group (targeting 30 on average) was intended to exceed the requirements of a simple power analysis we conducted following pilot experiments. Pilot experiments showed differences in test-time accuracy between control and either enhancement taper (equivalent to ET in main-text Table 1) or difficulty selection (equivalent to DS in Table 1) participants to be roughly 10%, with a standard deviation of roughly 10% in each group.

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

Given:

$$\delta = 0.1 \text{ (estimated mean difference)}$$
$$\sigma = 0.1 \text{ (estimated standard deviation)}$$
$$\alpha = 0.05 \text{ (significance level)}$$
$$1 - \beta = 0.8 \text{ (power)}$$

$$\text{Estimated effect size } d = \frac{\delta}{\sigma} = 1.0$$

$$\begin{aligned} \text{Required sample size per group: } n &= 2(z_{1-\alpha/2} + z_{1-\beta})^2/d^2 \\ &= 2(1.96 + 0.84)^2/1^2 \\ &\approx 16 \text{ subjects per group at minimum} \end{aligned}$$

We provide a demographic breakdown of the participants in our study, aggregated across experiments, in Table S2. Some participants took part in more than one of the experiments, but are only counted once in the table.

Total participants	521
Pts. w/ demographic data	519 (99.6%)
Age	
Mean (SD)	36.6 (11.9) years
Range	18-83 years
Sex	
Female	289 (55.7%)
Male	227 (43.7%)
Not specified	3 (0.6%)
Ethnicity	
White	338 (65.1%)
Black	54 (10.4%)
Asian	50 (9.6%)
Mixed	44 (8.5%)
Other	23 (4.4%)
Not specified	10 (1.9%)

Table S2: Demographic characteristics of study participants, aggregated across all experiments.

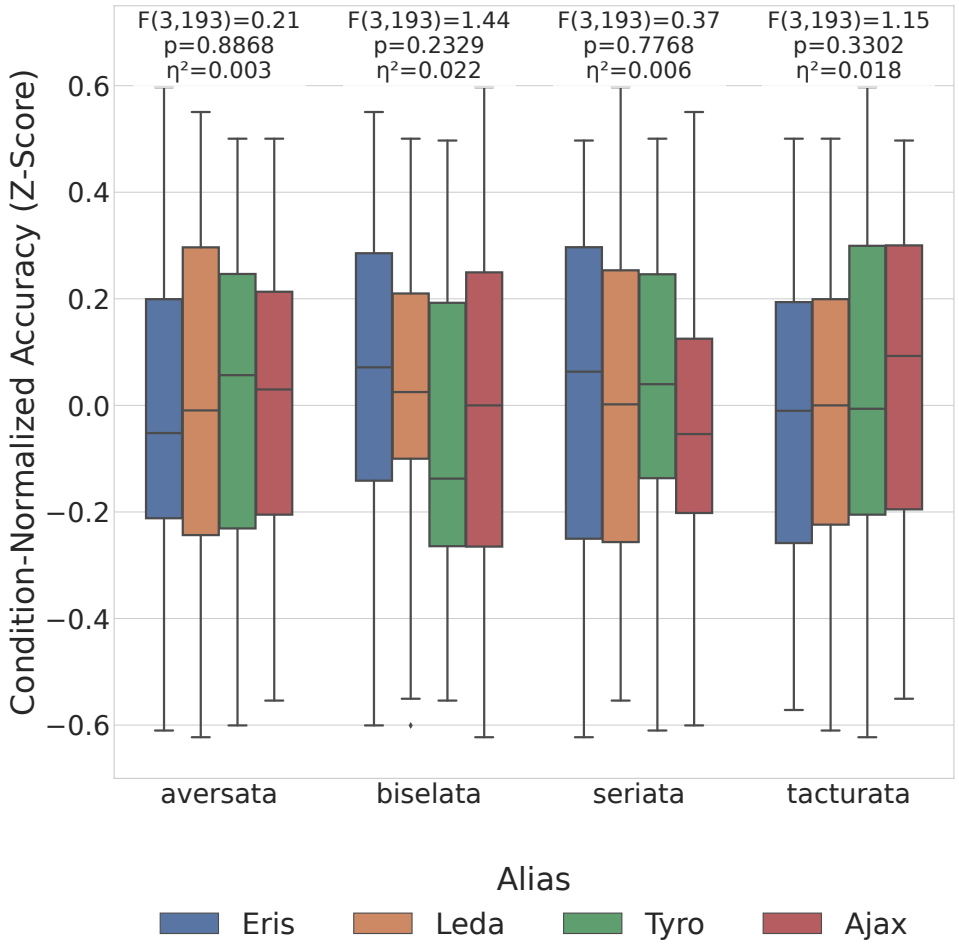

Figure S12: **Randomized assignment of aliases "Leda," "Ajax," "Eris," and "Tyro" had minimal impact on test-phase accuracy.** *Each group of four boxplots shows the relative effects of assigning each alias to a specific class from the moth classification experiment. Each individual boxplot indicates the distribution of participant-wise test-phase accuracy z-scores (normalized with mean and standard deviation within each condition separately) among participants with mapping of a specific alias onto a specific class - for example, the left-most boxplot within the right-most group describes the accuracy of participants who saw benign mole images labelled as "Leda." There is no evidence from one-way ANOVA that the random assignment of aliases to classes influences test-phase performance.*

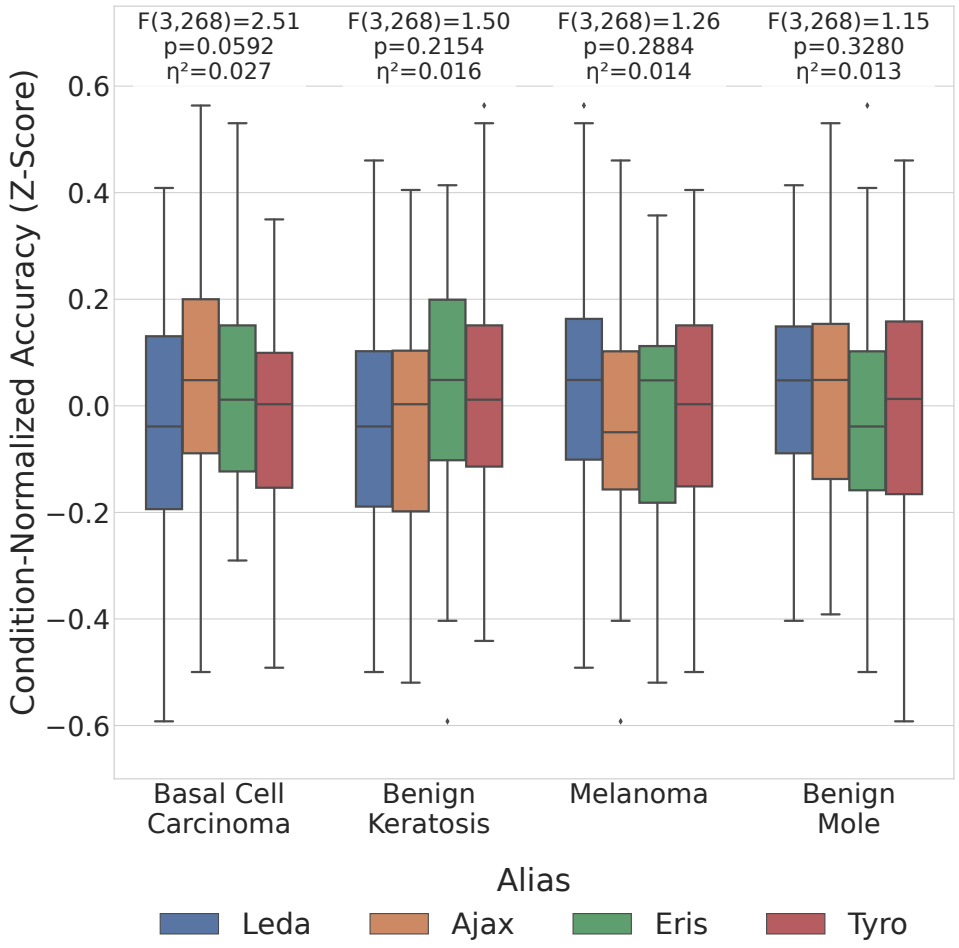

Figure S13: **Randomized assignment of aliases "Leda," "Ajax," "Eris," and "Tyro" had minimal impact on test-phase accuracy in the dermoscopy task.** *After correcting for multiple comparisons, there is no evidence that the random assignment of aliases to classes affects task performance. See also Fig. S12.*