Emulating and Enhancing Human Visual Perception and Learning with Image-Computable Models

by

Morgan B. Talbot

Submitted to the Harvard-MIT Program in Health Sciences and Technology
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN MEDICAL ENGINEERING AND MEDICAL PHYSICS

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2026

Authored by: Morgan B. Talbot
Harvard-MIT Program in Health Sciences and Technology
January 30, 2026

Certified by: Gabriel Kreiman, PhD
Professor of Ophthalmology, Thesis Supervisor

Accepted by: Collin M. Stultz, MD, PhD
Director, Harvard-MIT Program in Health Sciences and Technology
Nina T. and Robert H. Rubin Professor in Medical Engineering and Science
Professor of Electrical Engineering and Computer Science

THESIS COMMITTEE

Thesis Supervisor

Gabriel Kreiman, PhD
Professor of Ophthalmology
Boston Children's Hospital and Harvard Medical School

Thesis Readers

James J. DiCarlo, MD, PhD
Peter de Florez Professor of Neuroscience
Brain and Cognitive Sciences
Director, MIT Quest for Intelligence

Richard N. Mitchell, MD, PhD
Professor of Pathology
Brigham and Women's Hospital and Harvard Medical School
Associate Director, Harvard-MIT Health Sciences and Technology

Emulating and Enhancing Human Visual Perception and Learning with Image-Computable Models

by

Morgan B. Talbot

Submitted to the Harvard-MIT Program in Health Sciences and Technology
on January 30, 2026 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN MEDICAL ENGINEERING AND MEDICAL PHYSICS

ABSTRACT

The convergence of artificial intelligence (AI) with the neural and behavioral sciences has produced powerful image-computable models: artificial neural networks (ANNs) that encode the same raw stimuli as the human visual system. Optimizing these networks for behaviorally relevant tasks, such as visual categorization, enables them to generate neural representations of images that predict human brain activity as well as outputs aligned with human responses. Recent work has begun to explore whether predictive ANNs can be inverted to *modulate* human neural activity and behavior: what is the ideal stimulus to provide to a model, and ultimately to a human, in order to elicit a desired outcome? For example, can we use models to generate images that induce more accurate human judgments, or organize stimuli into sequences that catalyze more effective learning? A major challenge with this approach is that, despite their impressive capabilities, standard ANNs diverge from human perception, cognition, and learning in fundamental ways. Consequently, stimuli produced by inverting ANNs are often unintelligible and therefore ineffective for humans. This thesis develops and validates ANNs that are functionally aligned with human visual perception and learning, and applies these models to augment human performance in these domains. I first address the divergence between biological and artificial memory, engineering a brain-inspired algorithm to mitigate the "catastrophic forgetting" that afflicts standard ANNs. This establishes a more realistic model of human-like continuous learning. Next, I show that ANNs aligned with the perceptual robustness of human vision are strong predictors of stimulus difficulty, and can be inverted to produce image perturbations that augment human visual perception. I demonstrate that these predictions and perturbations can be leveraged to accelerate human learning of challenging visual categorization tasks. I validate the translational application of this approach in a medical education setting to address a key bottleneck in residency-level pathology training. Finally, I close the loop between perception and learning by casting ANNs as "surrogate learners" that simulate human learning trajectories. By effectively inverting simulations of the learning process to find stimulus sequences that maximize surrogate performance, I optimize instructional curricula that significantly enhance human visual learning in a controlled experimental setting. These findings collectively suggest that the fidelity of a model's alignment with human cognition is the key determinant of its utility as an educational tool, validating the translational potential of human-inspired AI and establishing a new paradigm of model-driven learning optimization.

Thesis supervisor: Gabriel Kreiman, PhD
Title: Professor of Ophthalmology

Acknowledgments

Acknowledgments will be added in the final version.

Biographical Sketch

A short biosketch will be added in the final version.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Background

There is a common observation in academia that it is often easier to learn an unfamiliar and challenging subject, such as organic chemistry or calculus, from a teaching assistant who learned the material last year than from an accomplished professor who has shaped their field for decades. For the professor, performing a complex mathematical derivation may be as cognitively effortless as crossing a footbridge. Their expertise is characterized by a high degree of procedural automaticity: the mental stepping stones that a beginner must ponderously traverse have long since been subsumed by unconscious reflexes, which must be actively deconstructed in order to teach effectively. The teaching assistant, however, still inhabits a cognitive framework similar to that of a student. In recalling a topic well enough to teach it, they are forced to simulate the learner's internal state by necessity, allowing them to intuitively guide each conceptual leap.

This phenomenon reflects a fundamental principle of pedagogy: effective teaching requires a shared cognitive architecture. To bridge gaps between confusion and comprehension, the teacher must be able to model the learning process of the student. This is reflected in evidence that reciprocal peer teaching is a highly effective educational practice [1], and I suspect it is one reason why the same word is used to mean both "teach" and "learn" in languages such as Welsh (*dysgu*), Māori (*ako*), and Norwegian (*lære*). To teach is to run a simulation of another person's mind within your own.

Current applications of AI in education largely ignore this principle, ranging from recasting large language models as interactive personal tutors [2,3], to multimodal tracking of learner characteristics and the state of a learner's knowledge [4]. Very few approaches to AI-augmented education have used AI systems as simulators of learning itself. Some existing work in this domain has focused on teaching humans logical reasoning tasks (e.g., solving linear equations [5] or understanding the merge sort algorithm [6]) that are solvable with symbolic methods. Sen et al. [7] used a custom-built artificial neural network (ANN) to simulate humans learning to recognize molecular structures across different visual formats. They demonstrated that curricula optimized using this model improved the performance of human students beyond curricula designed by domain experts. This strongly validates the idea that simulations of learning can be used as powerful pedagogical tools. However, the approach relied on low-dimensional, hand-coded feature vectors (e.g., counts of atoms or bonds) to represent visual stimuli, restricting the model to a feature space defined a priori by a human domain expert. It is an open question whether *image-computable* models, systems that can directly and flexibly encode the same kind of visual inputs as humans without manual feature engineering, can be leveraged to enhance human visual learning in a similar way.

Deploying image-computable models as simulators of human learning presents a formidable challenge: state-of-the-art ANNs do not naturally see or learn the way humans do. One important difference is referred to as sample efficiency: in cases where domain-specific AI models have attained human-level or superhuman performance, this has typically required equally superhuman volumes of training. For example, the AlphaGo model that famously defeated world Go champion Lee Sedol in 2016 had been trained by playing 30 million games against itself [8], orders of magnitude more than a human Go player could expect to play in a lifetime. Additionally, ANNs in relatively realistic circumstances conducive to human learning, such as beginning to learn one task after mastering another, exhibit "catastrophic forgetting:" training on new tasks tends to rapidly degrade all of an ANN's previous knowledge [9]. Further discrepancies can be found at the level of perception: for example, ANNs tend to memorize high-frequency noise patterns that are idiosyncratic to the datasets they are trained on, rendering them vulnerable to "adversarial examples," tiny stimulus perturbations that can be undetectable to humans but cause catastrophic failures of ANN perceptual judgments [10].

There are a variety of approaches to aligning AI systems with humans to produce higher-fidelity models of human perception and cognition. A key strategy towards this goal is benchmarking a range of models with varying architectures, training configurations, and other properties using standardized, computable definitions of human alignment, thus determining which models best approach a given definition of "human-like" and guiding the development of increasingly human-like models. For example, the BrainScore project benchmarks models' ability to predict behavioral and neural activity patterns during vision and language tasks [11], and CogBench measures tendencies of large language models toward human-like cognitive biases and other characteristics [12]. One might expect models to inevitably become more human-like as a combination of engineering advances and increased architecture/data scaling brings AI performance closer to that of humans. There is some evidence for a process resembling convergent evolution, whereby similar optimization pressures in ANN development (both training and engineering) and human evolution cause ANNs to develop human-like (hypothetically approaching "universal" in some instances) representational spaces [13,14] and neural mechanisms such as predictive processing [15].

However, other evidence suggests no convergence or even divergence from humans in increasingly performant models. Although early deep neural networks that approached human performance levels in object classification showed a strong correlation between accuracy and neural predictivity of inferior temporal (IT) cortex in macaque monkeys (a widely accepted homolog of the human higher ventral stream) [13], more highly accurate and recently developed models exhibit spatial attention patterns that are increasingly dissimilar to those of humans [16]. On a behavioral level, models' ability to imitate humans in several cross-modal variations of the Turing test has strikingly little correlation to conventional performance measurements on the underlying tasks [17].

One strategy to develop more human-like AI systems is to simulate biological neural circuits from the ground up, or mechanistically modify existing architectures to be more biologically plausible. Promising work in this direction has convincingly emulated key circuitlevel properties of the human brain, notably efficient and durable memory encoding [18,19], a long-standing unsolved challenge in AI. These efforts are bottlenecked by the embryonic state of our mechanistic understanding of the brain: in terms of performing or learning tasks that are behaviorally relevant to humans, they currently lag far behind brute-force engineering approaches such as deep learning with large datasets.

Another promising strategy is to directly optimize ANNs to align with human behavioral or neural data [20]. For example, training ANN image classifiers on noisy distributions of human labels (rather than a single ground truth category for each image) not only confers more human-like uncertainty patterns, but can also improve out-of-distribution generalization and robustness to adversarial examples [20]. Similarly, introducing a loss function term that incentivizes spatial attention distributions resembling those compiled from crowd-sourced human annotations increases classification accuracy in addition to inducing more human-like attention patterns [16]. A prerequisite for this powerful approach is the collection of large-scale human behavioral data.

The present work employs rationale-driven strategies for aligning AI with humans, circumventing the need for massive behavioral datasets: first, identify a starkly observable behavioral discrepancy between humans and ANNs optimized for task performance, then rationally design a "patch" that attempts to resolve the discrepancy with additional mechanisms, training configurations, or other inductive biases, and finally attempt to harness any useful human-like properties that consequently emerge. I focus specifically on inducing human-like memory, visual perception, and learning dynamics, and ultimately on leveraging the resulting human-aligned models to enhance visual perception and learning in humans towards practical applications in education.

Chapter 2 of this thesis highlights the widely observed human-machine discrepancy of catastrophic forgetting, in which ANNs trained on new tasks show extremely rapid performance degradation on previously learned tasks, a departure from the human ability to sequentially learn a broad range of tasks with great flexibility. A novel algorithm built around modern ANNs, and loosely inspired by replay mechanisms in the mammalian brain, is shown to be capable of continuously learning new tasks. The same model is found to exhibit a human-like bias towards classifying objects based on their shapes, relative to standard ANNs that are biased more towards texture information.

Chapter 3 builds upon another key discrepancy between humans and machines: the surprising vulnerability of ANNs to adversarial examples, subtle perturbations to stimuli such as images that are barely detectable to humans yet profoundly disruptive to ANN processing. I show that applying established techniques that directly train ANNs to be "robust" to these perturbations yields models that can (i) strongly predict the difficulty of categorizing individual images for humans, and (ii) generate new image perturbations that make images easier for humans to recognize. I further demonstrate that both of these capabilities can be applied to enhance visual category learning in humans, allowing experiment participants to learn more rapidly and to higher levels of accuracy across multiple categorization tasks.

Chapter 4 presents a proof-of-concept study towards translational application of the ANN-based techniques introduced in Chapter 3. In a randomized controlled trial of 147 first-year pathology residents, curricula of breast and prostate histology images ordered by difficulty using ANNs, coupled with ANN-based enhancement of diagnostically relevant features, measurably improved learning outcomes relative to an active control condition without ANN assistance. This work ultimately aims to address a key bottleneck in pathology residency: less time for traditional histology instruction given expanding requirements in genomic medicine and molecular diagnostics [21,22], exacerbated by residents arriving with less histology experience due to the trend of medical schools replacing dedicated pathology courses with integrated curricula [23].

The learning enhancement approaches of Chapters 3 and 4 employ models of human visual *perception*, coupled with heuristics based on general observations from the cognitive science of learning, such as designing a curriculum sequence that begins with straightforward examples and progresses to more challenging ones. Chapter 5 introduces a novel approach leveraging human-like models of the process of *visual category learning* itself. Rather than relying on heuristics or rational curriculum design principles, I use ANN-based models as "surrogate learners" that learn visual tasks under realistic conditions for humans, and employ an evolutionary algorithm to directly optimize the curricula that they learn from to maximize surrogate performance. I demonstrate that a curriculum optimized using a simple, linear surrogate can enhance human visual category learning. This complements my collaborative work with Singh et al. [24] investigating how the performance of humans and machines is similarly influenced by the order in which a sequence of distinct tasks is learned. Finally, I show that curriculum properties depend on the surrogate learners used for optimization, and explore possible alternative approaches to developing surrogates that are increasingly aligned with human learning mechanisms and dynamics.

Ultimately, this work proposes a synergistic convergence of artificial and biological intelligence, implemented at the intersection of machine learning and education. By engineering AI systems that perceive and learn more like us, we do not merely build better machines; we build simulators of cognition that can help us understand, and ultimately enhance, our own capacity for learning.

Chapter 2

Tuned Compositional Feature Replays for Efficient Stream Learning

2.1 Abstract

Our brains extract durable, generalizable knowledge from transient experiences of the world. Artificial neural networks come nowhere close to this ability. When tasked with learning to classify objects by training on non-repeating video frames in temporal order (online stream learning), models that learn well from shuffled datasets catastrophically forget old knowledge upon learning new stimuli. We propose a new continual learning algorithm, Compositional Replay Using Memory Blocks (CRUMB), which mitigates forgetting by replaying feature maps reconstructed by combining generic parts. CRUMB concatenates trainable and re-usable "memory block" vectors to compositionally reconstruct feature map tensors in convolutional neural networks. Storing the indices of memory blocks used to reconstruct new stimuli enables memories of the stimuli to be replayed during later tasks. This reconstruction mechanism also primes the neural network to minimize catastrophic forgetting by biasing it towards attending to information about object shapes more than information about image textures, and stabilizes the network during stream learning by providing a shared feature-level basis for all training examples. These properties allow CRUMB to outperform an otherwise identical algorithm that stores and replays raw images, while occupying only 3.6% as much memory. We stress-tested CRUMB alongside 13 competing methods on 7 challenging datasets. To address the limited number of existing online stream learning datasets, we introduce 2 new benchmarks by adapting existing datasets for stream learning. With only 3.7-4.1% as much memory and 15-43% as much runtime, CRUMB mitigates catastrophic forgetting more effectively than the state-of-the-art.

Code available at: https://github.com/MorganBDT/crumb.git

© 2025 IEEE. This chapter is reprinted, with permission, from: M. B. Talbot*, R. Zawar*, R. Badkundri, M. Zhang†, and G. Kreiman†, "Tuned Compositional Feature Replays for Efficient Stream Learning." *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 2, pp. 3300-3314, Feb. 2025, doi: 10.1109/TNNLS.2023.3344085. *Equal contribution, †corresponding authors. Author contributions are detailed in Appendix A.

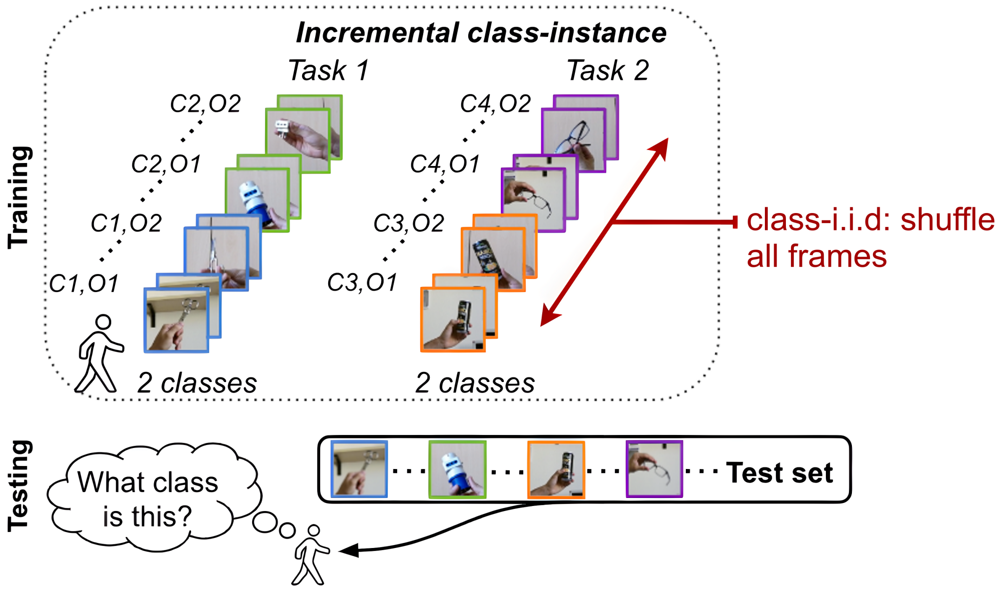


Figure 2.1: **Schematic of online stream learning protocols.** For each task, the model learns to classify a set of new classes (C1, C2, etc. in figure) while training on video clips of several objects from each class (O1, O2) for only one epoch. During testing, the model has to classify images from all seen classes without knowing task identity. In the class-instance training protocol, the order of video clips is shuffled but the order of frame images is preserved within each clip. In the class-i.i.d. training protocol, all images within each task are randomly shuffled. Class-i.i.d. is the only option for datasets such as ImageNet that consist of standalone images and not video clips.

2.2 Introduction

Humans adapt to new and changing environments by learning rapidly and continuously. Previously learned skills and experiences are retained even as they are transferred and applied to new tasks, which are learned from a stream of highly temporally correlated stimuli and without direct access to past experiences. In contrast, in standard class-incremental image classification settings in continual learning, neural networks are presented with images that are independently and identically distributed (i.i.d.), with multiple presentations of each image [25–27]. To better emulate a human learning environment, or that of an autonomous robot that must learn in real time, we focus on a challenging and realistic variant of class-incremental learning — *online stream learning*. Online stream learning has two key characteristics (Fig. 2.1): (a) the input is in the form of video streams with highly temporally correlated frames, and (b) each training example is presented only once: no repeated presentations of old data are allowed.

In online stream learning settings, current machine learning systems tend to fail to retain

In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of Massachusetts Institute of Technology's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink. If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

good performance on previously learned tasks, exhibiting catastrophic forgetting [28–30]. Catastrophic forgetting is a pervasive problem in continual learning settings for both deep neural networks [9] and other models such as linear regression [31] and self-organizing maps [32], and can also cause neural models to be biased towards more recently encountered training data [33]. One strategy for overcoming catastrophic forgetting is to store a copy of all or most encountered training examples for later replay, effectively converting to an offline learning paradigm [34]. This approach, however, often requires an impractically large amount of memory [35]. Moreover, much of the information in raw images is redundant, with many pixel values needed to represent each feature-level concept relevant to classification. Finally, storing old training data might also be undesirable from a data security or privacy standpoint, such as in hospitals and other healthcare settings [36].

To address both memory inefficiency and data privacy concerns while achieving state-of-the-art online stream learning performance, we propose a new continual learning approach, Compositional Replay Using Memory Blocks (CRUMB) (Fig. 2.2). In our method, each new image is processed by the early layers of a convolutional neural network (CNN) to produce a feature map tensor. The feature map is decomposed by slicing it into chunks, each of which is a vector of feature activations at a specific spatial location. Each chunk is then replaced by the most cosine-similar row ("memory block") of a trainable "codebook matrix." This mechanism encodes images as a composition of discrete feature-level concepts, some of which appear to have semantic interpretations. Storage of a complete training example for replay requires keeping only the indices of the memory blocks needed to reconstruct the original feature map, along with the class label, occupying only 3.6% of the memory footprint of a raw image. During replay, feature maps reconstructed via stored indices are fed to the later layers of the CNN, such that these layers are trained on both stored and newly encountered images to learn new tasks while retaining previous knowledge.

Our key contributions are:

- **Trainable Compositional Replay.** We propose a new compositional feature-level replay algorithm, CRUMB, for online stream learning. The composition mechanism is end-to-end trainable and reusable. CRUMB's codebook of memory blocks captures the essential components needed for reconstructing feature maps. During the pretraining phase, the memory block mechanism primes the CNN for stream learning with high accuracy and induces a beneficial bias towards object shapes. Using memory blocks as a shared basis for new and recalled examples helps stabilize the network during stream learning.

- **Reduced forgetting.** We tested CRUMB on 7 continual learning datasets alongside 13 competing methods, showing that CRUMB typically outperforms state-of-the-art approaches by large margins.

- **Superior Efficiency.** Storing n compositional feature maps for replay prevents catastrophic forgetting substantially more effectively than storing n raw images, while only requiring about 3.6% as much memory. Additionally, compared with the next most accurate method (REMIND [28]), CRUMB requires only about 15-43% as much training runtime, and occupies only 3.7-4.1% of REMIND's peak memory footprint.

- **New Benchmarks.** We adapted 2 datasets, Toybox [37] and iLab [38], to introduce new online stream learning benchmarks. All benchmark details along with source code, results, and data are available at https://github.com/MorganBDT/crumb.git.

2.3 Related work

2.3.1 Weight regularization

Weight regularization methods typically store weights trained on previous tasks and impose constraints on subsequent weight updates to minimize catastrophic forgetting [35,39–44]. However, storing the importance of the millions of parameters required by state-of-the-art recognition models across all previous tasks is costly [35,45]. Moreover, empirical comparisons suggest that weight regularization methods typically do not mitigate catastrophic forgetting as effectively as architecture adaptation and replay methods [46].

2.3.2 Architecture adaptation

Architecture adaptation methods expand or re-organize the structure of their neural networks to accommodate new tasks to be learned. Approaches include adding groups of new neurons (which does not always scale well) [35,39,41–43,47], isolating parts of a larger neural network for each task [18,48–51], compressing parameters in a consolidation phase [52], and pruning neurons or weights for later re-use. Pruning approaches include L1 regularization and activity threshold-based sparsification of neurons [53], and combining pruning of weights with parameter importance-based regularization [54]. Neuron pruning/re-use can also be combined with the addition of new neurons to improve performance and enable increased flexibility [55]. All of these approaches add significant complexity, and some require explicit labelling of task identities, which is not feasible in many online learning applications.

2.3.3 Image and feature replay

In replay methods, images or features from previous tasks are stored and later retrieved or re-generated to be shown to the model to prevent forgetting [40,56–61]. Replay can be highly effective, but comes with some caveats. Relying on replaying limited sets of stored examples can lead to overfitting. Storing a large number of raw images for replay is memory-intensive. To limit memory requirements, generative replay systems combine data from new tasks with synthetic data produced by generative models to resemble previously encountered stimuli [62–68]. However, the generative models needed to create adequate synthetic data remain large, memory-intensive, and difficult to train [45].

Other replay methods save memory by storing raw or compressed feature maps from intermediate CNN layers [28,69], or generate synthetic examples by sampling from simple feature-level probability distributions for each class [70]. REMIND [28] achieves high performance in online stream learning by compressing feature maps using a product quantizer [71]. However, the product quantizer is trained by performing k-means clustering on a large subset of training data stored in memory, a process that scales poorly for increasingly large datasets. In contrast, CRUMB's differentiable codebook is trained by backpropagation alongside other network parameters, dramatically reducing memory requirements for codebook initialization.

CRUMB's feature-based replay mechanism is inspired by biological replay observed in the hippocampus and other brain areas [72–74], and by complementary learning systems theory [75]. Recent work has explored modeling the hippocampus and neocortex as separate neural networks that interact via distillation losses and other mechanisms [76]. In contrast to storing knowledge implicitly in a short-term memory network, CRUMB's memory blocks and replay buffer store short-term memories that represent individual training examples, and interact with the CNN via replay to facilitate longer-term memory storage and consolidation.

2.4 Methods

2.4.1 Online stream learning benchmarks

Training protocols

We consider two online class-incremental settings: class-instance and class-i.i.d. [28] (Fig. 2.1).

Class-instance. Each task contains short video clips of different objects from two or more classes, and the video clips are presented one after another in random order within each task without repetition. An ideal learning algorithm in this setting would be stable enough to remember prior tasks while being sufficiently plastic to learn generalizable class boundaries for new classification tasks, despite encountering many highly correlated images of each object before moving on to the next.

Class-i.i.d.. Images/video frames are randomly shuffled within each task but not interspersed among tasks, and are shown only once like in class-instance. This is a less challenging protocol, and should not be considered stream learning in the strictest sense because the shuffling of images in each task destroys any temporal structure among images.

In both settings, our model and all competing baseline models are allowed to train for many epochs on the first task, but are restricted to viewing each image from subsequent tasks only once. This emulates real-time acquisition of training data that cannot be stored except in a limited-capacity replay buffer.

Stream learning benchmark datasets

We evaluated our model on five video datasets (class-instance and class-i.i.d. protocols), and two image datasets (class-i.i.d. only). For all datasets, we used different task and example orderings across training runs. A global holdout test set of images/frame sequences was used for all runs. To help address the limited number of online stream learning benchmarks, we adapted two datasets designed for studying object transformations, Toybox [37] and iLab [38], for online stream learning.

The **CORe50 video dataset [77]** contains images of 50 objects in 10 classes. Each object has 11 instances, which are 15 second video clips of the object under particular conditions and poses. We followed [28] for the training and testing data split, and sampled each video at 1 frame per second (FPS).

The **Toybox video dataset [37]** contains videos of toy objects from 12 classes. We used a subset of the dataset containing 348 toy objects, each of which has 10 instances containing different patterns of object motion. We sampled each instance at 1 FPS, resulting in 15 images per instance per object. We chose 3 of the 10 instances for our test set, leaving 7 instances for training.

The **iLab (iLab-2M-Light) video dataset [38]** contains videos of toy vehicles from 14 classses. We used a subset of the dataset containing 392 vehicles, with 8 instances (backgrounds) per object and 15 images per instance. We chose 2 of the 8 instances for our test set.

The **iCub (iCubWorld Transformations) video dataset [78]** contains videos taken by the iCub robot of 20 classes of household objects undergoing viewpoint transformations. We used isolated rotation, scaling, and background transformations as our training set, and the provided "MIX" sequence (which combines all transformations) as our test set.

The **iLab+CORe50 video dataset** combines iLab and CORe50 to create a stream learning benchmark with 24 distinct classes. All iLab classes are learned before CORe50, introducing a mild domain shift. We uniformly subsample iLab to balance the number of images per class with CORe50.

To evaluate our model in long-range online class-incremental learning with many more classes than the video datasets described above, we also include results on two image datasets. The standard **Online-CIFAR100 image dataset [79]** is split into 20 tasks with 5 classes each, while the standard **Online-Imagenet image dataset [80]** is split into 10 tasks with 100 classes each. Class-instance training is not applicable to image datasets because they do not consist of videos.

Baseline algorithms for comparison

All baseline algorithms use the same training protocols as CRUMB. CRUMB and most baselines use a SqueezeNet CNN pretrained on ImageNet [81], but due to implementation constraints AAN[82], CoPE [83], GSS [58], LwF [39], RM [61], and Stable SGD [84] use non-pretrained ResNet models [85]. We re-implemented some methods due to varying code availability. CRUMB and all baselines are implemented using the PyTorch library [86].

Weight Regularization: We compared against Elastic Weight Consolidation (EWC) [35], Synaptic Intelligence (SI) [42], Memory Aware Synapses (MAS) [87], Learning without Forgetting (LwF) [39], and Stable SGD (stSGD) [84].

Memory Distillation and Replay: We compared against Gradient Episodic Memory (GEM) [60], Incremental Classifier and Representation Learner (iCARL) [57], Bias Correction (BiC) [56], Gradient Sample Selection (GSS) [58], Continual Prototype Evolution (CoPE) [83], Adaptive Aggregation Network (AAN) [82], REMIND [28], and Rainbow Memory (RM) [61].

The **Lower bound** is trained sequentially over all tasks without any measures to avoid catastrophic forgetting.

The **Upper bound** is trained offline on shuffled images from both the current and all previous tasks over multiple epochs.

Chance predicts class labels by randomly choosing 1 out of the total of C_t classes seen in or before current task t.

2.4.2 Proposed algorithm: CRUMB

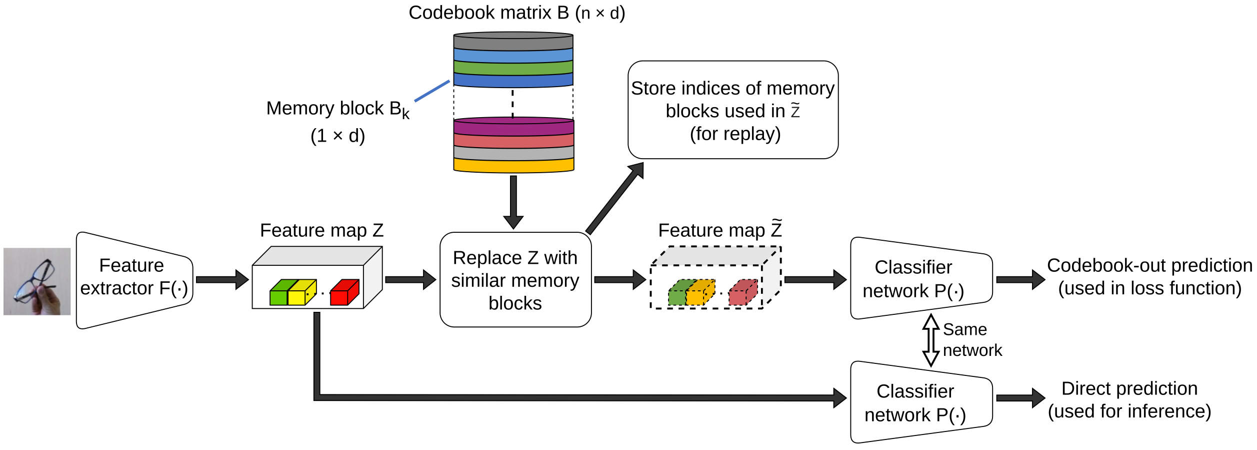

Codebook matrix B (n × d)
Memory block B$_k$
(1 × d)
Store indices of memory blocks used in Z̃ (for replay)
Feature extractor F(·)
Feature map Z
Replace Z with similar memory blocks
Feature map Z̃
Classifier network P(·)
Codebook-out prediction (used in loss function)
Same network
Classifier network P(·)
Direct prediction (used for inference)

Figure 2.2: **Schematic illustration of CRUMB, the proposed algorithm for online stream learning.** The model consists of a CNN ($\mathbf{F}(\cdot)$ for early layers and $\mathbf{P}(\cdot)$ for later layers) and a codebook matrix $\mathbf{B}$ used for compositional reconstruction of feature-level activation tensors (feature maps $\mathbf{Z}$). Each row $\mathbf{B_k}$ of $\mathbf{B}$ is a "memory block" vector. CRUMB uses the feature extractor $\mathbf{F}(\cdot)$ to produce an initial feature map, then determines which memory blocks to retrieve from $\mathbf{B}$ based on a cosine-similarity addressing mechanism. The feature maps reconstructed from the memory blocks ($\widetilde{\mathbf{Z}}$), and the original feature maps ($\mathbf{Z}$), are used to obtain separate classification losses from the same classifier network $\mathbf{P}(\cdot)$ ("**codebook-out loss**" and "**direct loss**", respectively). Only codebook-out loss is used for weight updates during stream learning, although the two losses are added in a weighted sum to calculate the total loss during pretraining. To avoid catastrophic forgetting, we store the row indices of retrieved memory blocks along with class labels for example images from each task. In later tasks, following each batch of new images, we "replay" a batch of old feature maps to $\mathbf{P}(\cdot)$ after reconstructing them using stored memory block indices.

We propose a new continual learning algorithm, Compositional Replay Using Memory Blocks (CRUMB). CRUMB consists of a 2-dimensional CNN augmented by an $n \times d$ codebook matrix B. A schematic of CRUMB is shown in Fig. 2.2, with further details described in algorithm 1. CRUMB extracts a feature map from each given image using the early layers of a pre-trained CNN, and stores a subset of the feature maps in a buffer. When CRUMB later encounters a new task, it avoids catastrophic forgetting of previous tasks by replaying stored feature maps of images from those tasks to the later layers of the network. To reduce memory requirements, CRUMB uses its codebook matrix B to reconstruct each feature map. Rows of B ("memory blocks") are concatenated to form tensors that approximate the original feature maps, and only the indices of activated memory blocks need to be stored to enable later reconstruction. All computations from memory block reconstruction forward are differentiable, allowing B to be learned alongside the CNN weights.

Feature extraction and classification

CRUMB's CNN backbone is split into two nested functions. The early layers of the network comprise $F(\cdot)$, a "feature extractor," while the remaining, later layers comprise $P(\cdot)$, a classifier.

Since early convolutional layers of CNNs are highly transferable [88], the parameters of $F(\cdot)$ are pretrained for image classification using ImageNet [80] and then fixed during stream learning. CRUMB passes each training image through feature extractor $F(\cdot)$ to obtain feature map Z, of size $s \times w \times h$ (number of features, width, height). Z is reconstructed using B to form $\widetilde{Z}$, and a class prediction output can then obtained as $P(\widetilde{Z})$. The parameters of $P(\cdot)$ are initially pretrained alongside $F(\cdot)$ on ImageNet using standard methods, but also undergo additional ImageNet pretraining with an objective incorporating predictions on both Z and $\widetilde{Z}$. Prior to stream learning, only the final layer of $P(\cdot)$ is randomly reinitialized to reflect the number of classes to be learned during streaming.

Reconstructing feature maps from memory

CRUMB produces reconstructed feature map $\widetilde{Z}$ using only Z and the contents of its $n \times d$ codebook matrix B, where each of the n rows B_k is a "memory block" vector. Hyperparameters n and d are determined empirically (see sections 2.5.4 and 2.5.4). Z is first partitioned evenly along its feature dimension into s/d tensors, with each tensor Z_f of size $d \times w \times h$. Each tensor Z_f is further partitioned by spatial location into $w \cdot h$ vectors, denoted $Z_{f,x,y} \in \mathrm{I\!R}^{\mathrm{d}}$, where d is also the dimension of each row B_k in the matrix B. For each vector $Z_{f,x,y}$ in Z, a similarity score γ_k is calculated between it and each memory block B_k as follows:

$$\gamma_{f,x,y,k} = \langle Z_{f,x,y}, \frac{B_k}{\|B_k\|_2} \rangle \tag{2.1}$$

Where $\langle \cdot, \cdot \rangle$ is the dot product, and $\| \cdot \|_2$ is the L2-norm. Because B_k is normalized, $\gamma_{f,x,y,k}$ is highest for the memory block most similar in vector direction to the given $Z_{f,x,y}$. The memory block B_k with the highest γ similarity value replaces $Z_{f,x,y}$ at its corresponding location in $\widetilde{Z}$ as follows:

$$\widetilde{Z}_{f,x,y} \leftarrow B_{k_{f,x,y}} \text{where } k_{f,x,y} = \operatorname*{argmax}_k(\gamma_{f,x,y,k}) \tag{2.2}$$

Because $\widetilde{Z}$ is reconstructed entirely from memory blocks B_k, we can save all information needed to reconstruct $\widetilde{Z}$ again later by storing both B and the values of k at each f, x, y location in $\widetilde{Z}$. Thus, the feature map for the i^{th} training image can be stored as $m_i = (k_{1,1,1}, ..., k_{f,x,y}, ..., k_{s/d,w,h})$.

For example, in our main implementation, Z is a $512 \times 13 \times 13$ tensor. $d = 16$ so that Z is split into $32 \cdot 13 \cdot 13 = 5408$ vectors of length 16, which are each replaced in $\widetilde{Z}$ by a 16-dimensional memory block from a 256×16 matrix B. The memory blocks themselves occupy a near-negligible amount of memory: $256 \times 16 = 4096$ floating point values, compared to the 5408 integers required to store a single compressed training example in this implementation.

Training

During training, both Z and $\widetilde{Z}$ are passed separately through the classifier $P(\cdot)$ to obtain two classification probability vectors $p = P(Z)$ and $\widetilde{p} = P(\widetilde{Z})$, where the dimension of p_t and $\widetilde{p}_t$ is equal to the total number of classes C_t that have been seen in or before the current task

Algorithm 1 CRUMB at task t

```
Input: training images $I_t$ from new classes, stored codebook matrix $B$, replay buffer $X$ of stored memory block indices and their class labels (maximum number $n_X$ of total stored examples in $X$ varies by dataset).

Training:
for batch in $I_t$ do
    Reconstruct feature map $Z$ as $\widetilde{Z}$ for each image in batch by concatenating memory blocks (rows of $B$)
    Train $P(\cdot)$ using loss $L(P(Z), P(\widetilde{Z}), y_c)$, with $\alpha = 0$ in streaming
    Train memory blocks in $B$ that form part of $\widetilde{Z}$, using $L_{CE}(P(\widetilde{Z}), y_c)$
    if $t > 1$ then
        Randomly sample images $x$ out of $X$ to form a replay batch
        Reconstruct $\widetilde{Z}$ for each $x$ by concatenating memory blocks
        Train $P(\cdot)$ using loss $L_{CE}(P(\widetilde{Z}), y_c)$
        Train memory blocks in $B$ that are part of $\widetilde{Z}$ via backpropagation
    end if
end for
Store in X: memory block indices for reconstruction of every $j^{\text{th}}$ image
Testing:
for batch in testing images do
    Compute predictions $p = P(F(\cdot))$ on test images using $Z$ only
end for
```

t. The loss function L used for training is a weighted sum of the cross-entropy losses L_{CE} derived from p and $\widetilde{p}$. With y_c defined as the ground truth class label of a given image:

$$L(p, \widetilde{p}, y_c) = \alpha L_{CE}(p, y_c) + \beta L_{CE}(\widetilde{p}, y_c) \tag{2.3}$$

Larger values of α penalize "direct" prediction errors from $P(Z)$, while larger values of β penalize "codebook-out" prediction errors from $P(\widetilde{Z})$. Although our model generates class predictions based on both Z and $\widetilde{Z}$, we use the empirically more accurate predictions from Z during inference on the test set. Empirically, the best performance was achieved by including both direct and codebook-out predictions in the loss function for pretraining ($\alpha = 1, \beta = 1$), and then removing the direct loss for stream learning ($\alpha = 0, \beta = 1$) (see section 2.5.4) for analysis). Setting α to 0 makes the loss function for new batches of images more similar to that used for replay, where only $\widetilde{Z}$ is available. Replacement of Z by the reconstructed version $\widetilde{Z}$ can be viewed as both a means to efficiently mitigate catastrophic forgetting and a regularization technique to prevent overfitting and stabilize the CNN.

Although values in CRUMB's memory blocks play the role of activation values in their reconstruction of $\widetilde{Z}$, they are trainable parameters of the model. Backpropagation from $\widetilde{Z}$-based predictions generates gradients for the values in each memory block used for reconstruction, and stochastic gradient descent modifies the memory blocks towards minimizing the same training objective used for the network weights (cross-entropy loss).

Initializing the codebook matrix and CNN

CRUMB's performance benefits from targeted initialization and pretraining of its CNN and the memory blocks in its codebook matrix, especially in the class-instance setting. The values in the codebook matrix B directly replace those in "natural" feature maps derived from images during training - accordingly, B is initialized using a simple univariate distribution designed to match that of natural feature maps from a pretrained network. In early experiments, we tried initializing the codebook using k-means cluster centers from feature vectors of CIFAR100 images [79], but this did not improve performance.

Stream learning performance was substantially improved by pretraining CRUMB on ImageNet [80] classification with 1000 classes, as compared to applying CRUMB to stream learning with a CNN pretrained by standard methods. Pretraining tunes the values in the memory blocks, and also regularizes the CNN in preparation for stream learning by training it to make predictions using lossy reconstructions of feature maps.

Replay to mitigate catastrophic forgetting

In online stream learning (see section 2.4.1), the model is presented with images I_t from new classes c^{new} in task t, where c^{new} are drawn from the subset of classes in the dataset that the model has not seen in previous tasks.

Replay of examples from previous tasks is a proven strategy to mitigate catastrophic forgetting in class-incremental settings [40,56–58], and feature-level replay can be considerably more memory-efficient than storing raw images [28]. We store compressed representations of feature maps from images in each task, and then replay a batch of stored feature maps after each batch of new images during later tasks to mitigate forgetting.

CRUMB stores up to n_X pairs of labels and tensors (y_i, m_i), corresponding to images from old classes c^{old} of previous tasks. Depending on the number of seen classes C_{t-1}, the storage for each old class contains n_X/C_{t-1} pairs. n_X is chosen for each dataset depending roughly on the total number of classes. Some algorithms select representative image examples to store and replay based on different scoring functions [89–92]. However, random sampling uniformly across classes yields outstanding performance in continual learning tasks [45], and we adopt this strategy to select examples from the buffer for replay. To choose training examples for storage in the replay buffer, CRUMB keeps every j^{th} image in each batch, where j is calculated by dividing the number of training images in each task by the replay buffer capacity n_x, such that the buffer is filled near the end of the current task's training epoch. In the class-instance setting, this maximizes sample diversity by minimizing the number of frames sampled from within the same video clip, and further avoiding sampling frames from the same clip that are in close temporal proximity.

For replay-based baseline methods (iCARL, REMIND, etc), we limit the number of examples that can be stored in the buffer to fit within a memory budget that is fixed across all methods (see details in Appendix B, Section B.5). Aside from n_x and the training batch size (which is smaller for video datasets), CRUMB uses the same hyperparameters for all datasets, including n, d, α, β, learning rate, and initialization and pretraining protocols.

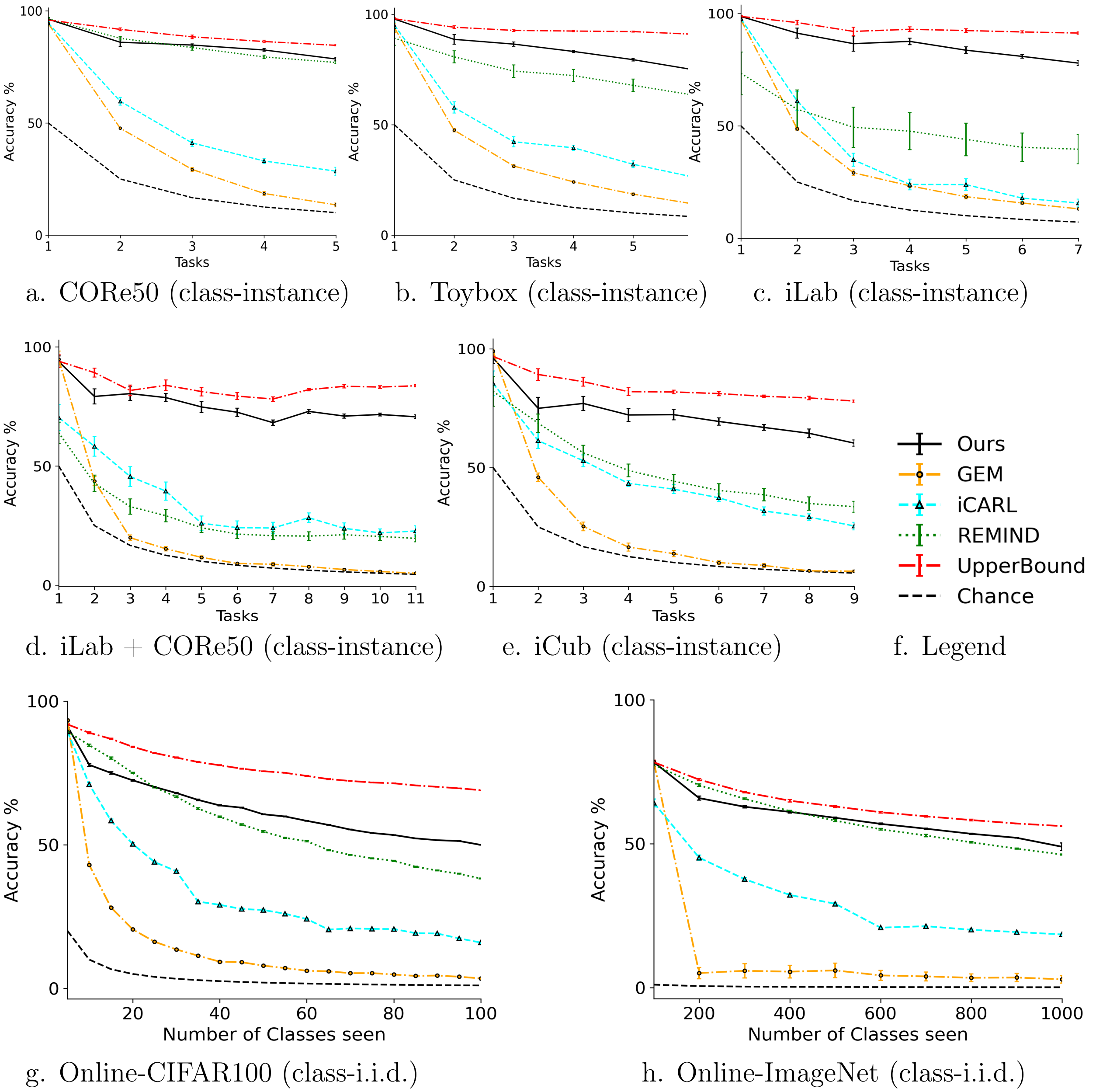
Table 2.1: **CRUMB outperforms state-of-the-art algorithms on most benchmarks.** Each number is the mean top-1 accuracy on all tasks/classes after the completion of stream learning. Values are averages from 10 (CORe50 [77], Toybox [37], iLab [38], iCub [78], iLab+CORe50, Online-CIFAR100 [79]) or 5 (Online-ImageNet [80]) independent runs. The highest accuracy in each column (excluding the offline upper bound) is in bold, while the second-highest is italicized. Algorithm name abbreviations can be found in section 2.4.1. Class-instance, in which video frames are presented in temporal order, is only applicable to video datasets CORe50, Toybox, iLab, iCub, and iLab+CORe50. Due to resource constraints, for Online-ImageNet, iCub, and CORe50+iLab, we tested a subset of baseline algorithms. Class-instance and class-i.i.d. settings are abbreviated as "inst." and "i.i.d." respectively. Results are grouped vertically with memory distillation and replay methods at the top (Ours - CoPe), weight regularization methods in the middle (EWC - LwF), and lower/upper bounds at the bottom. Data preparation methods are detailed in Appendix B, Section B.4.1.

	CORe50		Toybox		iLab		iCub		iLab+CORe50		CIFAR100	ImageNet
	inst.	i.i.d.	inst.	i.i.d.	inst.	i.i.d.	inst.	i.i.d.	inst.	i.i.d.	i.i.d.	i.i.d.
Ours	**78.5**	**81.2**	**74.9**	*75.7*	**77.9**	*79.5*	**60.0**	**65.8**	**66.0**	**74.4**	**49.9**	**48.9**
REMIND [28]	*77.0*	*76.0*	*66.2*	**84.1**	*48.1*	**81.0**	*33.2*	*58.5*	*22.9*	*67.2*	*38.2*	*46.2*
iCARL [57]	27.0	28.5	27.3	26.5	15.6	23.6	23.4	20.9	17.8	23.2	15.9	18.5
GEM [60]	11.9	13.5	14.3	15.7	13.0	12.8	5.2	6.1	4.6	4.5	3.5	2.9
RM [61]	12.0	12.4	9.8	20.8	18.2	9.3	-	-	-	-	4.2	-
AAN [82]	14.0	15.6	13.2	17.6	10.6	15.0	-	-	-	-	6.6	-
GSS [58]	15.0	15.6	14.7	15.0	13.0	12.8	-	-	-	-	3.2	-
BiC [56]	10.2	11.8	11.0	10.2	11.2	10.9	-	-	-	-	4.0	-
CoPE [83]	16.6	16.3	21.7	22.4	17.6	18.6	-	-	-	-	8.8	-
EWC [35]	12.2	12.4	14.3	15.7	13.5	13.0	7.3	6.4	11.7	12.0	3.9	0.1
MAS [87]	14.4	17.4	18.9	19.2	20.5	22.1	4.8	4.8	4.4	4.3	5.5	0.1
SI [42]	12.0	12.9	14.3	15.5	12.8	13.0	5.3	5.7	4.4	5.0	3.6	8.8
StSGD [84]	13.7	13.2	13.5	13.8	9.8	6.9	-	-	-	-	7.3	-
LwF [39]	12.5	12.4	21.9	20.9	10.5	11.9	-	-	-	-	4.2	-
Lower bound	12.1	12.8	15.5	16.9	12.8	16.4	5.8	6.0	4.5	4.6	3.5	3.0
Upper bound	85.3	84.6	91.0	92.0	91.3	91.4	76.9	78.0	83.1	81.7	69.0	56.1

2.5 Results

2.5.1 Stream learning on video datasets

A naive CNN trained on stream learning benchmarks learns each task effectively, but rapidly and catastrophically forgets all prior tasks in doing so. In contrast, a brute-force approach to overcoming catastrophic forgetting that achieves excellent performance in a stream learning setting is to store all encountered images and corresponding class labels, shuffle them, and exhaustively retrain on the resulting dataset in an offline, i.i.d. fashion. This renders the benchmark equivalent to offline class-incremental learning ("upper bound" in Fig. 2.3) [34]. By storing a subset of old examples and using a compositional strategy to compress these examples, CRUMB allows CNNs to approach the performance of a brute-force approach with roughly an order-of-magnitude reduction in training time and a tiny 0.013% fraction (on CORe50) of the memory footprint. Accordingly, given a fixed memory budget, CRUMB outperforms all competing models in all five tested video stream learning datasets in the class-instance setting, often by large margins. For example, as shown in Fig. 2.3a-e, CRUMB's top-1 accuracy on all tasks after class-instance stream learning exceeds that of REMIND by



a. CORe50 (class-instance)
b. Toybox (class-instance)
c. iLab (class-instance)
d. iLab + CORe50 (class-instance)
e. iCub (class-instance)
f. Legend
g. Online-CIFAR100 (class-i.i.d.)
h. Online-ImageNet (class-i.i.d.)

Figure 2.3: **CRUMB outperforms most baseline algorithms and approaches the upper bound on some datasets**. Line plots show top-1 accuracy in online stream learning on video datasets (a) CORe50 (b) Toybox (c) iLab (d) iLab + CORe50 (e) iCub in the class-instance training protocol (class-i.i.d. plots are in Appendix B, Fig. B.1), as well as image datasets (g) Online-CIFAR100 and (h) Online-ImageNet (class-i.i.d.). All models train on the first task for many epochs, but **view each image only once** on all subsequent tasks. Accuracy estimates are the mean from 10 runs (5 runs for ImageNet), where each run has different class and image/video clip orderings. Error bars show the root-mean-square error (RMSE) among runs. Results for all baselines are in Table 2.1.

1.5%, 8.7%, 29.8%, 26.8%, and 43.1%, iCARL by 51.5%, 47.6%, 62.3%, 36.6%, and 48.2%, and GEM by 66.6%, 60.6%, 64.9%, 54.8%, and 61.4% on CORe50, Toybox, iLab, iCub and iLab+CORe50 respectively. The class-instance performance of all models is shown in Table 2.1. CRUMB approaches the offline upper bound to within 6.8%, 16.1%, 13.4%, 16.9% and 17.1% on the same datasets, demonstrating strong mitigation of catastrophic forgetting.

The less challenging class-i.i.d. setting is similar to class-instance in that tasks are learned sequentially without revisiting previous tasks, and that each image is seen by the model only once. However, all images within each class-i.i.d. task are shuffled in an i.i.d. manner, removing the local temporal correlations introduced by sequential frames in video clips. As with class-instance, CRUMB achieves excellent class-i.i.d. performance: as shown in Appendix B, Fig. B.1a-e, CRUMB's top-1 accuracy on all tasks after class-i.i.d. learning exceeds that of iCARL by 52.7%, 49.2%, 55.9%, 44.9%, and 51.2% and of GEM by 67.7%, 60%, 66.7%, 59.7%, and 69.9% on CORe50, Toybox, iLab, iCub and iLab+CORe50 respectively. The class-i.i.d. performance of all models is shown in Table 2.1. CRUMB approaches the offline upper bound to within 3.4%, 16.3%, 11.9%, 12.2% and 7.3% on the same datasets. The performance of REMIND and CRUMB was comparable on class-i.i.d., with CRUMB's accuracy higher than REMIND's by 5.2%, 7.3% and 7.2% on CORe50, iCub and iLab+CORe50 respectively. However, REMIND's accuracy was higher than CRUMB's by 8.4% and 1.5% on Toybox and iLab respectively.

On all benchmarks, CRUMB's closest competitor by far was REMIND, with all other methods exhibiting much lower accuracy. In general, the regularization baselines greatly underperformed the replay-based methods. This is perhaps due to limited exposure to each task given that each image may be visited only once, and/or because of overfitting to temporally correlated data, especially in the class-instance setting: replay addresses both of these issues while regularization methods generally do not. Because we used a fixed memory budget for replay methods, CRUMB is able to store many more examples than replay methods based on raw images, such as iCARL and GEM. This increases the diversity of the replayed stimuli.

2.5.2 Stream learning on natural image datasets

Although stream learning of CORe50, Toybox, iLab, iCub, and iLab+CORe50 is highly challenging, these datasets have only 10-24 classes each. To demonstrate CRUMB's capacity for long-range stream learning of many classes, we employed standard image datasets Online-CIFAR100 and Online-ImageNet. CRUMB outperformed all baselines on both of these datasets (see Table 2.1). On Online-CIFAR100, CRUMB's mean top-1 accuracy after class-i.i.d. stream learning exceeds that of REMIND by 11.7%, iCARL by 34%, and GEM by 46.4%, performing within 19.1% of the offline upper bound. On Online-Imagenet, CRUMB outperforms REMIND by 2.7%, iCARL by 30.4%, and GEM by 46%, performing within 7.2% of the offline upper bound (see also Fig. 2.3g-h).

2.5.3 Memory and runtime efficiency

CRUMB's closest competitor in top-1 accuracy is REMIND [28]. Both models require specific pretraining procedures. REMIND trains a product quantizer using k-means clustering of feature vectors, which requires a large portion of training data to be held in memory simultaneously. In contrast, CRUMB's codebook matrix is trained by backpropagation in tandem with CNN parameter updates. This approach requires only 3.7-4.1% of the peak RAM usage of REMIND for large datasets such as Online-CIFAR100 and Online-Imagenet. CRUMB also has a runtime only about 15-43% as long as REMIND's (Table 2.2).

Table 2.2: CRUMB uses only 3.7-4.1% of REMIND's peak RAM usage, and its runtime is approximately 15-43% of REMIND's.

Dataset	Peak RAM (GB)		Runtime (hours)	
	Ours	REMIND	Ours	REMIND
CIFAR100	**0.036**	0.87	**0.29**	1.91
Imagenet	**1.66**	44.34	**15.50**	35.64

2.5.4 Model analysis

To elucidate the importance of CRUMB's various components, we performed a series of ablation studies and experiments with altered training procedures. Accuracy results on CORe50 in both class-instance and class-i.i.d. settings are included for each experiment in Tables 2.3 and 2.4, but throughout the text in this section we discuss class-instance results except where otherwise stated. Experiment names are in **bold** throughout this section. We conducted statistical significance tests for each experiment (see Appendix B, Section B.4.2).

Replay: n CRUMB feature maps beat n images

Feature-level replay is the main mechanism by which CRUMB prevents catastrophic forgetting. Removing replay dramatically reduces accuracy by 64.9%. However, CRUMB does not require storing a large number of feature maps to mitigate forgetting: reducing buffer size n_X from 200 (**Ours**) to 100 (**Half capacity**) and to 50 (**Quarter capacity**) had a relatively small impact, with 4.7% and 13.3% accuracy drops respectively.

The quality of stored replay examples is also important. **Ours**, which stores memory block indices to compositionally reconstruct up to n_X feature maps, had dramatically higher accuracy than storing *the same number* n_X *of entire raw images* and training the network without any feature map reconstruction (**Image replay**), even though CRUMB's reconstruction of feature maps inevitably discards information and uses only 3.6% as much memory. As shown in Table 2.3, **Ours** attains 10.4% higher accuracy than **Image replay** on CORe50 (buffer size $n_x = 200$), 4.2% higher on Toybox, 15.4% higher on iLab, 7.8% higher on iCub, and 2.8% higher on iLab+CORe50. This result appears to hold only for the five video streaming datasets, however: **Ours** attained accuracy 0.16% higher than **Image replay** on Online-CIFAR100 (not statistically significant), and 18.3% lower than **Image replay** on Online-ImageNet, in the class-i.i.d. setting. **Ours** uses only 3.6% as much memory as **Image replay**: when the amount of memory usage is held constant (e.g., in gigabytes) instead of the maximum number of replay buffer items n_x, CRUMB outperforms image replay methods such as iCARL by very large margins on all datasets (see Table 2.1).

Table 2.3 also shows that CRUMB continues to outperform image replay on CORe50 when memory resources for the replay buffer are not constrained. Even at $n_x = 6400$, where algorithms can store the entire CORe50 training set in the replay buffer, **Ours** outperforms **Image replay** by 9.8% on class-instance and 3.2% on class-i.i.d. The reduced performance of **Image replay** relative to **Ours** is partly rescued by adding CRUMB pretraining (**Ours p.t. + im. rep.**, 3.7% and 2.2% below **Ours** in class-instance and class-i.i.d. respectively), even though the memory blocks play no role in either of the two image replay conditions

Table 2.3: **CRUMB performs better on video stream learning with *n* feature map representations in its replay buffer (**Ours**) than with *n* raw images (**Im. replay**), even though the former uses only 3.6% as much memory**. This finding is demonstrated for both class-i.i.d. and class-instance, across a range of buffer sizes on CORe50, and across 5 video datasets, although it does not persist for image datasets Online-CIFAR100 and Online-ImageNet. For CORe50 in the "Dataset" column, the buffer size is indicated as n_x. The table shows mean final top-1 accuracy on all tasks, averaged across 5 independent runs that each begin with an independent pretraining run. Significant differences from **Ours** are marked with *.

Dataset	Experiment	Class-inst.	Class-i.i.d.
CORe50 ($n_x = 200$)	Ours	78.22	79.93
	Im. replay	67.80*	75.88*
CORe50 ($n_x = 400$)	Ours	79.08	81.37
	Im. replay	71.55*	78.36*
CORe50 ($n_x = 800$)	Ours	79.46	82.21
	Im. replay	68.73*	81.67
CORe50 ($n_x = 1600$)	Ours	79.30	82.60
	Im. replay	69.64*	79.62*
CORe50 ($n_x = 3200$)	Ours	81.39	80.83
	Im. replay	70.50*	79.72
CORe50 ($n_x = 6400$)	Ours	79.39	83.76
	Im. replay	69.60*	80.55*
Toybox	Ours	68.19	68.91
	Im. replay	64.03*	62.68*
iLab	Ours	67.80	71.96
	Im. replay	52.36*	63.38*
iCub	Ours	58.33	63.02
	Im. replay	50.50*	47.88*
iLab+CORe50	Ours	61.89	72.55
	Im. replay	59.05*	67.58*
Online-CIFAR100	Ours	-	47.97
	Im. replay	-	47.81
Online-ImageNet	Ours	-	23.99
	Im. replay	-	42.27*

during streaming (see Table 2.4).

Replaying high-level features also contributed to CRUMB's performance. Storing n_X low-level feature maps from layer 3 instead of layer 12 (**Early feature replay** vs. **Ours**) reduced performance by 16.6%. CRUMB effectively stores memories with a higher level of abstraction than both **Image replay** and **Early feature replay**, and comes with both accuracy and memory efficiency improvements. The Memory Recall method (MeRec, [70]) uses a further level of abstraction by storing only the element-wise mean and standard deviation of feature activations for each learned class, and generating examples for replay by sampling from a Gaussian distribution parameterized by these values. In **MeRec replay**, we implemented MeRec's replay mechanism as a drop-in replacement for CRUMB's memory block reconstruction at the same network layer, and observed accuracy 56.9% and 41.3%

below **Ours** for class-instance and class-i.i.d. respectively, but 8.1% and 28.1% above **No replay**. MeRec stores an amount of data equivalent to two complete feature maps (mean and standard deviation) per class. In the implementation of CRUMB used for **Ours**, this is equivalent to 16 compressed feature maps per class or $n_x = 160$ in total, which is fewer than **Ours** at $n_x = 200$ but more than **Half capacity** at $n_x = 100$.

CRUMB pretraining primes CNN weights for streaming

Our model's performance is maximized by using CRUMB to pretrain the CNN and memory blocks on ImageNet prior to stream learning, particularly in the class-instance condition. Using randomly initialized memory blocks and a CNN pretrained without CRUMB (**Vanilla pretrain**) instead of CRUMB pretraining (**Ours**) reduced performance by 25.5% and 3.2% on class-instance and class-i.i.d. respectively. Our results also indicate that the benefit of CRUMB pretraining is attributable primarily to changes in the CNN weights, rather than changes to the memory blocks. Starting stream learning with CRUMB-pretrained weights and randomly re-initialized memory blocks (**Pretrain weights**) performs only 1.1% and 0.5% worse than **Ours**, while starting with vanilla-pretrained weights and CRUMB-pretrained memory blocks (**Pretrain mem. blocks**) is 23.7% and 2.8% worse than **Ours**, a marginal improvement over vanilla pretraining. As explained in section 2.5.4, CRUMB pretraining also improves stream learning performance when raw images are used for replay.

CRUMB pretraining using the smaller CIFAR100 dataset (100 classes) instead of ImageNet (1000 classes) (**CIFAR100 pretrain**) decreases accuracy by 11.6%.

In **Freeze memory**, no updates to memory blocks were allowed after pretraining. This had no statistically significant effect on accuracy, indicating that fine-tuning the memory blocks was unnecessary for stream learning on CORe50.

CRUMB pretraining induces a shape bias in the CNN

We hypothesized that CRUMB's pretraining procedure induces a bias towards shape information over texture information by training the CNN to make class predictions using lossy feature representations with unperturbed spatial distributions. A bias towards shape information has been shown to help mitigate forgetting by flattening the local minima for each task in the loss landscape [94]. To gauge CRUMB's degree of reliance on shape and texture information, we evaluated its performance on test set examples with three different perturbations. "Spatial perturbation" and "feature perturbation" modify the $13 \times 13 \times 512$ feature map at the same level where it is reconstructed using memory blocks. In "spatial perturbation," the positions of all of the 512-dimensional feature vectors are randomly shuffled in the 13×13 spatial grid, destroying global shape information but leaving feature information intact. In "feature perturbation," for each image independently we set a random selection of 50% of the features (i.e., 256 out of 512 total features) to zero, perturbing feature information but leaving coarse shape information intact. Spatial and feature perturbations do not directly interact with CRUMB's reconstruction mechanism, because CRUMB uses the original, non-reconstructed feature map to make class predictions. The "feature perturbation" strategy assumes that feature information is more related to texture information than shape information. Therefore, we also include "style perturbation" for ImageNet by testing on the the Stylized-ImageNet

Table 2.4: **Ablation and other experiments demonstrate the importance of CRUMB's various components.** Top-1 accuracy on all tasks after stream learning is averaged over 5 runs for all experiments. * denotes significant difference from **Ours** ($p < 0.01$, paired-samples t-tests on batches of 100 images).

Category	**Experiment name**	**Class-inst. % avg. accuracy**	**Class-i.i.d. % avg. accuracy**
Unablated	**Ours**	**78.22**	**79.93**
Replay format	Image replay	67.80*	75.88*
	Ours p.t. + im. rep.	74.49*	77.72*
	Early feature replay	61.64*	64.28*
	MeRec replay [70]	21.35*	38.65*
Replay ablation	Half capacity	73.55*	75.14*
	Quarter capacity	64.90*	67.80*
	No replay	13.28*	10.58*
Pretraining ablation	Vanilla pretrain	52.70*	76.69*
	Pretrain weights	77.08*	79.40
	Pretrain mem. blocks	54.54*	77.18*
	CIFAR100 pretrain	66.64*	74.96*
Freeze memory	Freeze memory	78.06	80.44
Memory block init.	Normal init.	47.70*	74.28*
	Uniform init.	39.84*	64.82*
	Dense matched init.	77.24	78.84*
Number of memory blocks	1 block	9.60*	9.47*
	2 blocks	64.28*	71.23*
	4 blocks	70.10*	75.77*
	8 blocks	74.60*	79.96
	16 blocks	77.70	80.30
	...	...	...
	256 blocks (**Ours**)	**78.22**	**79.93**
	512 blocks	78.71	79.74
Memory block size	4-dim. blocks	74.21*	77.55*
	8-dim. blocks (**Ours**)	**78.22**	**79.93**
	16-dim. blocks	79.05*	81.02*
	32-dim. blocks	78.17	79.93
	16-dim. blocks adj.	79.69*	82.36*
	32-dim. blocks adj.	80.31*	81.64*
Loss functions	Ours - direct loss	73.82*	78.11*
	Ours + direct loss	65.40*	69.20*
	Direct loss	48.12*	50.07*
CNN Architecture	MobileNetV2 CNN	76.13*	79.27

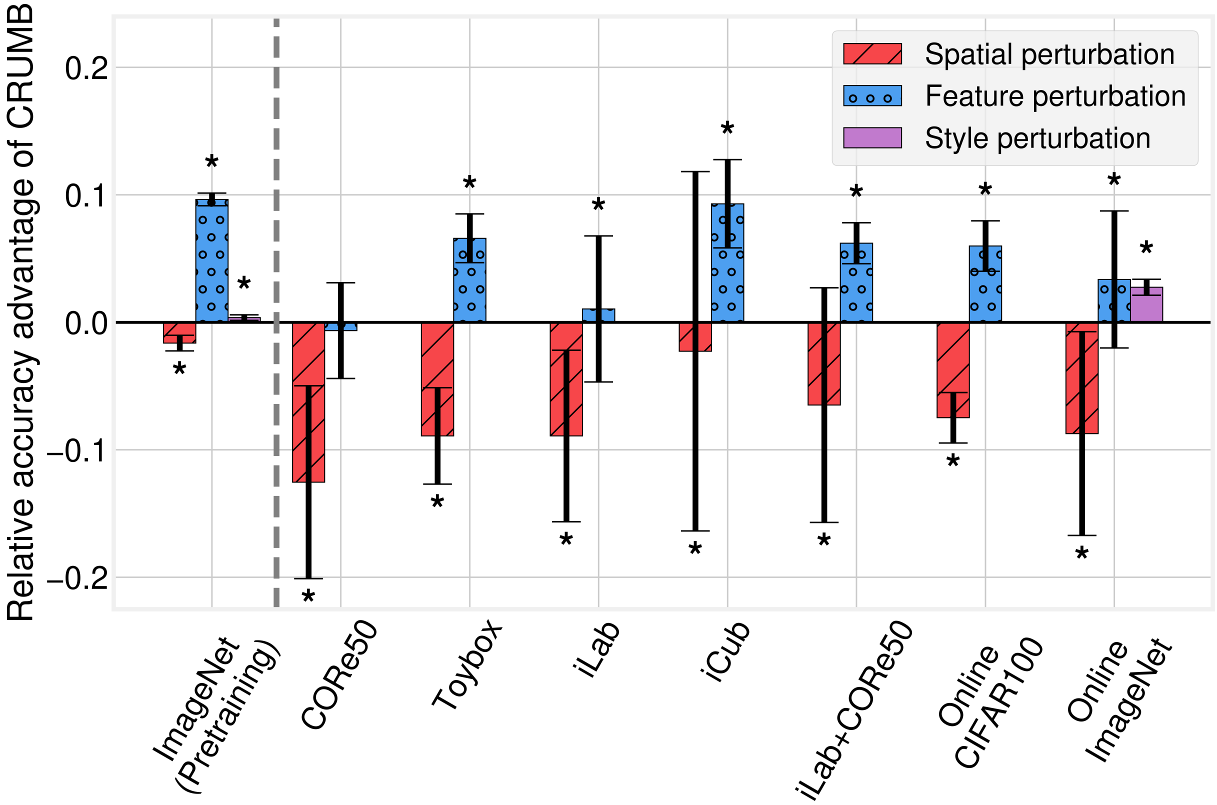


Figure 2.4: **CRUMB pretraining induces a bias towards shape information that often persists through stream learning**. The height of each bar shows how much smaller (or larger, if negative) CRUMB's drop in normalized test set accuracy under a perturbation is, in comparison to a control network (see section 2.5.4). "Spatial perturbation" shuffles the spatial positions of all feature vectors in an intermediate feature map (at the same layer where it is reconstructed by CRUMB), "feature perturbation" randomly sets half of the feature map's features to zero, and "style perturbation" uses images from Stylized-ImageNet [93]. Streaming results (to the right of grey dotted line) are in the class-instance setting for the video datasets and class-i.i.d. for CIFAR100 and ImageNet. Error bars are standard errors of the mean of relative accuracy advantage among 5 (CIFAR100 and ImageNet) or 10 (other datasets) independent runs. * denotes a statistically significant difference from 0, as determined by a Wilcoxon signed-rank test (see Appendix B, Section B.4.2).

dataset, which consists of images with heavily distorted local textures but largely intact global object shapes [93]. We would expect the performance of a relatively shape-biased network to be more severely reduced by spatial perturbation and less severely affected by feature and style perturbations than a control network. Fig. 2.4 shows CRUMB's "relative accuracy advantage" for each perturbation across several datasets. The relative accuracy advantage is calculated by dividing the accuracy drop (unperturbed accuracy minus perturbed accuracy) for each perturbation by the network's unperturbed accuracy, and subtracting this value for CRUMB from the corresponding value for a control network. The "ImageNet (pretraining)" condition in Fig. 2.4 shows CRUMB's shape bias on ImageNet following pretraining on ImageNet, with lower resilience against spatial perturbations (red bars with diagonal lines) and higher resilience against feature (blue with circles) and style (plain purple) perturbations. The control network is pretrained on ImageNet using the same procedure but without the CRUMB feature reconstruction step, thereby using only the "direct" loss for parameter updates. The other conditions in Fig. 2.4 correspond to CRUMB models evaluated on the test sets of these datasets after stream learning in class-instance (CORe50, Toybox, iLab, iCub, and iLab+CORe50) or class-i.i.d. (Online-CIFAR100 and Online-ImageNet) settings. Class-i.i.d. results for the video datasets are shown in Appendix B, Fig. B.3. Here, the control network performs stream learning with raw-image replay and "direct" loss instead of CRUMB's feature-level replay and "codebook-out" loss. Although shape bias testing results are noisy on some of the datasets, CRUMB most often retains a degree of bias towards shape information over texture/feature information after the completion of stream learning.

CRUMB can learn with very few memory blocks

CRUMB's performance did not change dramatically with changes to the number of memory blocks. Reducing the memory block count from **256 blocks** to as few as **16 blocks**, which effectively shrinks the library of feature combinations available to reconstruct feature maps, did not significantly decrease accuracy. Reducing the count further to **8 blocks** decreased accuracy by 3.6%, and reducing to **4** or **2 blocks** decreased accuracy by 8.1% and 13.9% respectively. Increasing to **512 blocks** did not significantly increase accuracy. This suggests a saturation effect, where a relatively small number of memory blocks is sufficient to reconstruct a wide variety of feature maps.

CRUMB is robust to different memory block sizes, and memory block size affects memory efficiency

CRUMB performs well with a range of memory block sizes. Decreasing the number of elements in each memory block from 8 to 4 (**4-dim. blocks**) results in a modest decrease in performance, 4.0% and 2.4% on class-instance and class-i.i.d. respectively. Increasing the number of elements from 8 to 16 or 32 (**16-dim. blocks**, **32-dim. blocks**), which arguably makes accurate reconstruction of feature maps more challenging because higher-dimensional vectors must be replaced by discrete choices of memory blocks, had negligible impact on performance (see Table 2.4).

The maximum number of examples stored in CRUMB's replay buffer (n_x) was held constant for the memory block size perturbations above. However, increasing the dimension of the memory blocks from 8 to 16 or 32 means that only half or one-quarter as many blocks respectively are needed to reconstruct each feature map, so only half/one-quarter as many indices need to be stored in the replay buffer per image. This allows double/quadruple the number of examples to be stored in the replay buffer within the same memory budget. When we allowed the maximum number of examples stored in the buffer to change accordingly ($2n_x$ for **16-dim. blocks adj.**, $4n_x$ for **32-dim. blocks adj.**), we observed accuracy improvements: **16-dim. blocks adj.** achieves 1.5% and 2.4% higher accuracy than **Ours** (n_x with 8-dimensional blocks) on class-instance and class-i.i.d. respectively, and **32-dim. blocks adj.** achieves 2.1% and 1.7% higher accuracy. During hyperparameter tuning for our main results, we observed that 16-dimensional memory blocks maximized testing accuracy.

Loss from reconstructed features is sufficient

CRUMB's performance is affected by the choice of components in its loss function. The loss function (equation 2.3) is the weighted sum of two terms, "direct loss" and "codebook-out loss." Our experiments show that the best performance is achieved when both direct loss and codebook-out loss are included in pretraining, but only codebook-out loss is included during stream learning. Removing direct loss from pretraining ("**Ours - direct loss**") results in a 4.4% drop in accuracy in the later stream learning tasks - learning from only reconstructed feature maps from start to finish, including during pretraining, is sufficient for decent performance. Including only codebook-out loss ("**Ours**") in stream learning yields a dramatic 30.1% gain in accuracy compared to using only direct loss ("**Direct loss**"), and a gain of 12.8% compared to using a weighted sum of direct loss and codebook-out loss ("**Ours + direct loss**"), despite the fact that only the direct, non-reconstructed feature map is used for inference on the test set.

Initialization of the memory blocks matters

CRUMB's performance is somewhat sensitive to the initialization of the values in the memory blocks. CRUMB trains its memory blocks in tandem with network weights after initialization, and concatenates them in different combinations to reconstruct feature maps produced by an intermediate network layer. We compared stream learning performance of four memory block initialization strategies, including initializing with values drawn from (1) a standard normal distribution (**Normal init.**), (2) a uniform distribution on the interval $[0, 1]$ (**Uniform init.**), (3) a distribution designed to match that of the non-zero values in the feature maps to be reconstructed, with 64% of all values reset to zero to approximately match the sparsity of typical feature maps (**Ours**), and (4) the same as (3), but with no values set to zero (**Dense matched init.**). Accuracy for **Normal init.** was 30.5% and 5.7% lower than **Ours** for class-instance and class-i.i.d. protocols respectively, accuracy for **Uniform init.** was 38.4% and 15.1% lower, and accuracy for **Dense matched init.** was 1.0% ($p = 0.045 > 0.01$) and 1.1% lower (see Table 2.4). It appears that drawing initial values for the memory blocks from a similar distribution to that of natural feature maps improves performance. When applying CRUMB to new network architectures, a simple alternative procedure to initialize the memory blocks would be to obtain feature maps from a batch of images, pool all values from all feature maps into one long vector, and initialize each memory block value by randomly sampling a value from this vector.

CRUMB is applicable across CNN architectures

In **MobileNetV2 CNN**, we implement CRUMB using the MobileNetV2 [95] CNN backbone instead of SqueezeNet [81]. Here, CRUMB reconstructs the $14 \times 14 \times 64$ input to MobileNetV2's sixth layer, using 256 8-dimensional memory blocks as in **Ours**. With the total memory usage of the replay buffer held constant, performance of **MobileNetV2 CNN** is comparable to **Ours** with only a 2.1% accuracy drop in class-instance and a 0.7% (not significant) drop in class-i.i.d.

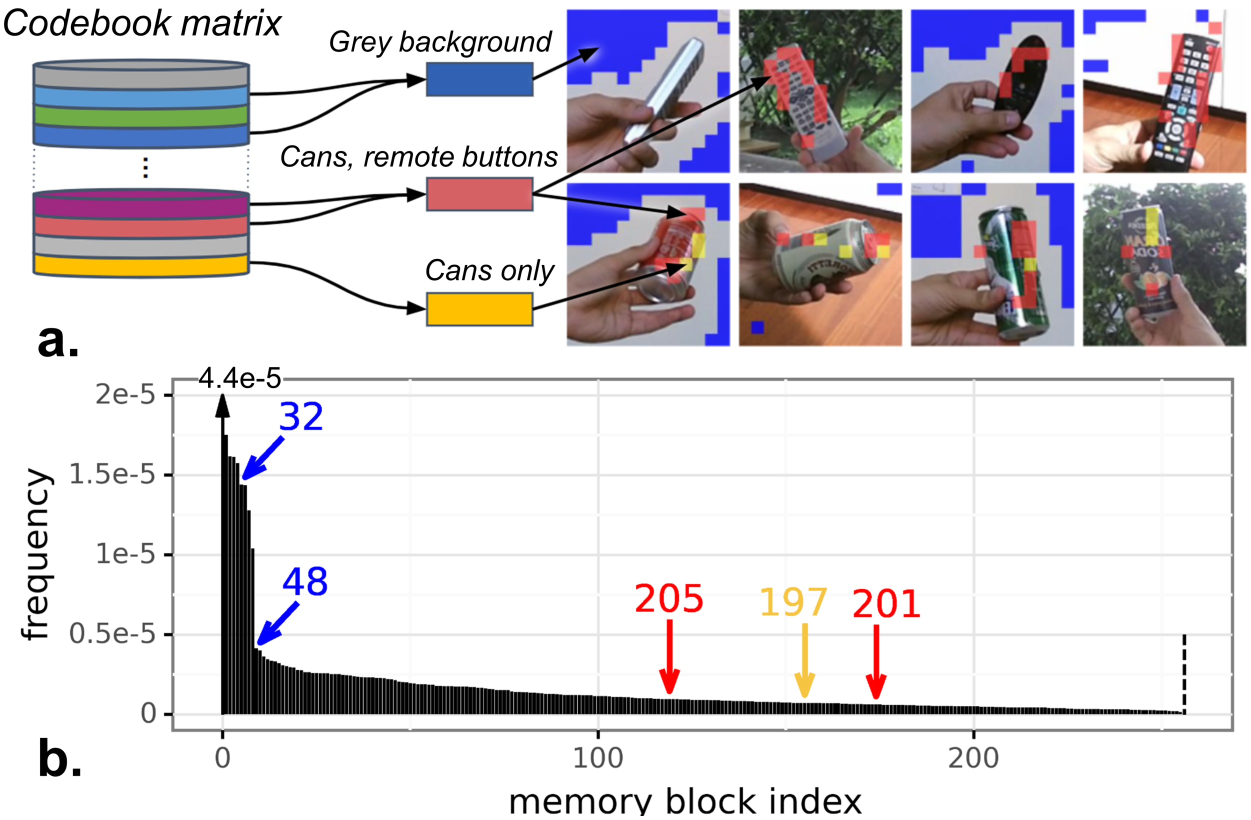


Figure 2.5: **Some memory blocks appear to have semantic interpretations**. Panel **a** shows images of "remote controls" and "cans" in the CORe50 test set, showing all-or-none activation of specific memory blocks at corresponding image locations. Of the 256 memory blocks in the codebook, blocks with indices 32 and 48 (blue squares) both similarly respond to greyish background regions, but not bright white or other backgrounds. Blocks 201 and 205 (red) both respond to buttons on remote controls and features of drink cans, while block 197 (yellow) responds only to can features. Similar blocks are aggregated by color (for blue and red) to produce a clearer visualization. Panel **b** shows the sorted usage frequencies in the CORe50 test set of each of the 256 memory blocks. Colored arrows show the blocks visualized in panel **a**. The upward black arrow shows the most-used block with frequency 4.4e-5.

Some memory blocks are coarsely interpretable

Visualizations of image locations where specific memory blocks are activated (Fig. 2.5) show that some memory blocks appear to be human-interpretable. Some blocks responded to features seen in images of one specific class or of a subset of classes, and others responded to features that are likely irrelevant to classification. In addition to the blocks visualized in Fig. 2.5, we found blocks that tend to respond to vertical lines, crosshatch patterns on balls and cups, pure white backgrounds, vegetation backgrounds, and wooden floor backgrounds, each of which can be interpreted as a semantic, compositional part of various test set images. Given the observations earlier in this section that either randomly re-initializing memory blocks prior to stream learning (**Pretrain weights**) or freezing memory blocks during stream learning (**Freeze memory**) has minimal effects on performance, it is not necessarily the case that these interpretable associations indicate learned representations within the memory blocks themselves. Another possibility is that a sufficient diversity of memory blocks allows useful associations between memory blocks and features or classes to be learned by the CNN through changes to the network weights.

The procedure for generating the visualizations in panel **a** of Fig. 2.5 can be understood as follows. For this analysis, we used a CRUMB model trained on CORe50 in the class-instance setting. CORe50 test set images are first passed through the early layers of the CNN to produce a feature map, which CRUMB then reconstructs by concatenating memory block vectors to produce an approximated version of the original feature map (see section 2.4.2). In this study, each feature map is of size $13 \times 13 \times 512$, meaning spatial dimensions of 13×13 with 512 features at each spatial location. Each memory block is one of 256 row vectors in the 256×8 codebook matrix used for this analysis. The memory blocks are 8-dimensional vectors, so each spatial location in the feature map's 13×13 grid is represented by a 512-dimensional vector formed by concatenating $512/8 = 64$ memory blocks end-to-end. Color coding in Fig. 2.5a shows at most one block per spatial location, the one activated by the first 8 features in the 512-dimensional feature vector, even though 64 memory blocks are activated at each location in total. We focus on the first 8 features for visualization purposes, because it is not necessarily the case that blocks activated by the first set of 8 features encode the same image features as they might when activated by the k^{th} set of 8 features (where $2 \leq k \leq 64$). To produce the images in Fig. 2.5a, each test set image is divided into a square 13×13 grid. Image grid locations are overlaid with colored squares, such that the color of each square depends on the memory block activated by the first 8 features at the corresponding spatial location in the feature map reconstructed by CRUMB. We only assigned colors to a handful of memory blocks with interesting properties, and we assigned the same color to sets of memory blocks that seemed to respond to very similar features. Fig. 2.5b shows the sorted distribution of the frequencies with which each of the 256 memory blocks were used to reconstruct feature maps from the CORe50 test set, with color and memory block index-coded arrows indicating the memory blocks visualized in Fig. 2.5a.

2.6 Discussion and conclusion

We developed a novel compositional replay strategy to tackle the problem of online stream learning, in which algorithms must learn tasks incrementally from non-repeating, temporally correlated inputs. Our algorithm, CRUMB, learns a set of "memory blocks" that are selected via cosine similarity and concatenated to reconstruct feature maps from an intermediate CNN layer. The indices of selected memory blocks are stored for a subset of training images, enabling memory-efficient replay of feature maps to mitigate catastrophic forgetting. CRUMB achieves state-of-the-art online stream learning accuracy across 7 datasets. Furthermore, CRUMB outperforms replay of an equal number of raw images by large accuracy margins across 5 video datasets, despite using only 3.6% as much memory as image replay.

Several factors seem to make important contributions to CRUMB's high performance. As shown in Fig. 2.4, pretraining with memory block reconstruction biases the CNN towards attending to object shapes rather than textures, an effect that typically endures throughout stream learning and which has been demonstrated to reduce catastrophic forgetting by flattening the loss minima for each task [94]. This could explain why CRUMB pretraining improves performance even if raw image replay is used during stream learning ("Ours p.t. + im. rep." in Table 2.4), in which case feature map reconstruction plays no role whatsoever during stream learning.

Backpropagation updates to memory blocks during pretraining and stream learning appear to be less important for performance than updates to the CNN weights. Randomly re-initializing the memory blocks after pretraining has a small negative effect on performance only in the class-instance setting, while keeping only the pretrained memory blocks but resetting to the original "vanilla" pretrained CNN weights before stream learning has a much larger negative impact ("pretrain weights" vs. "pretrain mem. blocks" in table 2.4). Furthermore, freezing the memory blocks during streaming has no discernible effect ("freeze memory" in table 2.4). However, "normal init." and "uniform init." in Table 2.4 show that the choice of probability distribution used to randomly initialize the memory blocks can have a dramatic impact on performance, with best performance attained by matching the univariate distribution of the memory blocks to that of natural feature maps. It seems to be important for stream learning that the later layers of the network receive feature maps with consistent univariate statistics, whether they are natural feature maps from the feature extractor layers or reconstructed feature maps from memory block concatenation.

Experiments in table 2.4 also show that the design of CRUMB's loss function is important. When training on new images, using only "codebook-out loss" from classification on reconstructed feature maps leads to much less forgetting than using "direct loss" from natural feature maps, either together with codebook-out loss ("ours + direct loss") or in isolation ("direct loss"). Only codebook-out loss is available when replaying feature maps reconstructed from memory blocks: using only codebook-out loss for new images means that only codebook-out loss is used throughout stream learning, rather than switching between direct loss for new examples and codebook-out loss for replayed examples. It appears that CRUMB's memory blocks form a shared, discretized basis for encoding training examples in feature space, which has a stabilizing effect on the CNN during stream learning. It is also notable in this context that, unlike during streaming, the inclusion of direct loss is important for gradient updates during pretraining ("Ours - direct loss" has lower accuracy than "Ours" in table 2.4). Combined with the observation that memory blocks should ideally be initialized from a univariate distribution approximating that of natural feature maps, this suggests that the pretraining process helps the CNN align its processing of natural and reconstructed feature maps, enabling the network to learn only from reconstructed feature maps during streaming even while continuing to make its most accurate predictions using natural feature maps instead.

The hypothesis that CRUMB's memory blocks provide a shared feature-level basis that stabilizes the CNN is consistent with the observation that, although CRUMB pretraining improves performance of raw image replay ("Ours p.t. + im. rep." in table 2.4), it still does not match CRUMB's performance with the replay buffer size n_x held constant: we speculate that redundant pixel-level information in raw images introduces additional noisy variation into the distribution of feature maps, which affects network stability. Another observation consistent with this hypothesis is that CRUMB only outperforms raw image replay on the 5 video datasets (with constant n_x, not constant memory usage), where network instability is likely to be more problematic due to temporal correlations within video clips and sudden transitions between them during training.

In both CRUMB and our raw-image replay ablation experiments, the early "feature extractor" layers of the CNN are frozen. When we apply CRUMB reconstruction to feature maps from an earlier layer (“early feature replay” in table 2.4) and correspondingly allow more layers after this point to have their parameters updated during stream learning, we observe performance worse than both CRUMB and image replay. One interpretation of this is that more unfrozen layers means that more network parameters are exposed to gradient updates and, consequently, to catastrophic forgetting. It is also possible that CRUMB's reconstruction mechanism is best suited to representing abstract, high-level features that are more likely to be found in later CNN layers, and that too much information is lost if CRUMB attempts to represent lower-level information that is interpreted by later layers in more granular ways.

Given the apparent necessity of preserving the information in feature maps during reconstruction, a surprisingly small codebook of memory blocks is sufficient. As few as 16 memory blocks are needed for optimal performance on CORe50, and CRUMB still performs remarkably well with only 2 memory blocks (see “number of memory blocks” experiments in Table 2.4). Although CRUMB is already highly memory-efficient with the memory blocks themselves occupying negligible space, reducing the number of memory blocks may enable further CPU memory usage optimizations (e.g., 4-bit integers as indices for 16 memory blocks) and also lowers GPU memory usage. Computational and memory efficiency is presumably critical in biological memory systems. Indeed, replay of neuronal activity patterns has been observed to help reinforce and consolidate memories in multiple brain areas across different mammalian species [72–74]. It is unlikely that neural circuits in the brain use replay mechanisms that preserve as much low-level information as pixel-level replay. Instead, it is interesting to speculate that one of the mechanisms by which brains avoid catastrophic forgetting is by replaying compositions of abstract, high-level features in a manner analogous to CRUMB's replay mechanism.

CRUMB's superior memory and runtime efficiency makes it ideally suited for settings with limited computational resources. Potential applications include edge computing in mobile devices, and autonomous robots that learn continuously from otherwise unmanageable amounts of incoming sensor data while they explore their surroundings. CRUMB could also be used in federated learning contexts, enabling highly effective replay of previously-seen data points via perhaps unrecognizably lossy representations, thereby minimizing both catastrophic forgetting and data security risks. The interpretable qualities of a subset of memory blocks, however, raises the possibility of identifying weak associations with certain generic features contained in a given training example, for a person with unauthorized access to CNN weights, memory blocks, encoded memories of interest, and a reference dataset to discover memory block interpretations. However, it would still be impossible for such a person to completely reconstruct CRUMB's memories in their original encodings (e.g., in pixels).

CRUMB is implemented here for SqueezeNet [81] and MobileNetV2 [95] CNNs, but could be used to mitigate forgetting across different neural network architectures in the future. For example, memory blocks could be used to efficiently reconstruct and replay vector outputs of self-attention heads at intermediate layers in transformer models [96,97].

Updating CRUMB's memory blocks using backpropagation in tandem with network weights is highly efficient, and also raises the possibility of tuning memory blocks for shifting domains on the fly. Although updates to the memory blocks beyond pretraining do not appear important for stream learning on CORe50, it is possible that fine-tuning may be necessary in tasks with substantial non-stationarity. Additionally, in this study, CRUMB does not adapt the early “feature extractor” layers of the CNN during stream learning. However, the early layers could theoretically be trained using the direct prediction loss while the late layers and memory blocks are trained using codebook-out loss or a combination of these two losses: this approach could enable additional flexibility for domain adaptation. Future studies could apply CRUMB to stream learning or reinforcement learning tasks with shifting domains, emulating humans or robots in continuously changing environments.

Chapter 3

L-WISE: Boosting Human Visual Category Learning Through Model-Based Image Selection and Enhancement

3.1 Abstract

The currently leading artificial neural network 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. We show that image perturbations generated by these models can enhance the ability of humans to accurately report the ground truth class. 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) applying image perturbations that aid recognition for novice learners. We find that combining these model-based strategies leads to categorization accuracy gains of 33-72% relative to control subjects without these interventions, on unmodified, randomly selected held-out test images. Beyond the accuracy gain, the training time for the augmented learning group was also shortened by 20-23%, despite both groups completing the same number of training trials. We demonstrate the efficacy of our approach in a fine-grained categorization task with natural images, as well as two tasks in clinically relevant image domains – histology and dermoscopy – where visual learning is notoriously challenging. To the best of our knowledge, our work is the first application of artificial neural networks to increase visual learning performance in humans by enhancing category-specific image features.

Project website/code: https://MorganBDT.github.io/L-WISE

This chapter is based on the following publication: M. B. Talbot, G. Kreiman, J. J. DiCarlo and G. Gaziv, "L-WISE: Boosting Human Visual Category Learning Through Model-Based Image Selection and Enhancement." *International Conference on Learning Representations (ICLR)*, 2025. Author contributions are detailed in Appendix A.

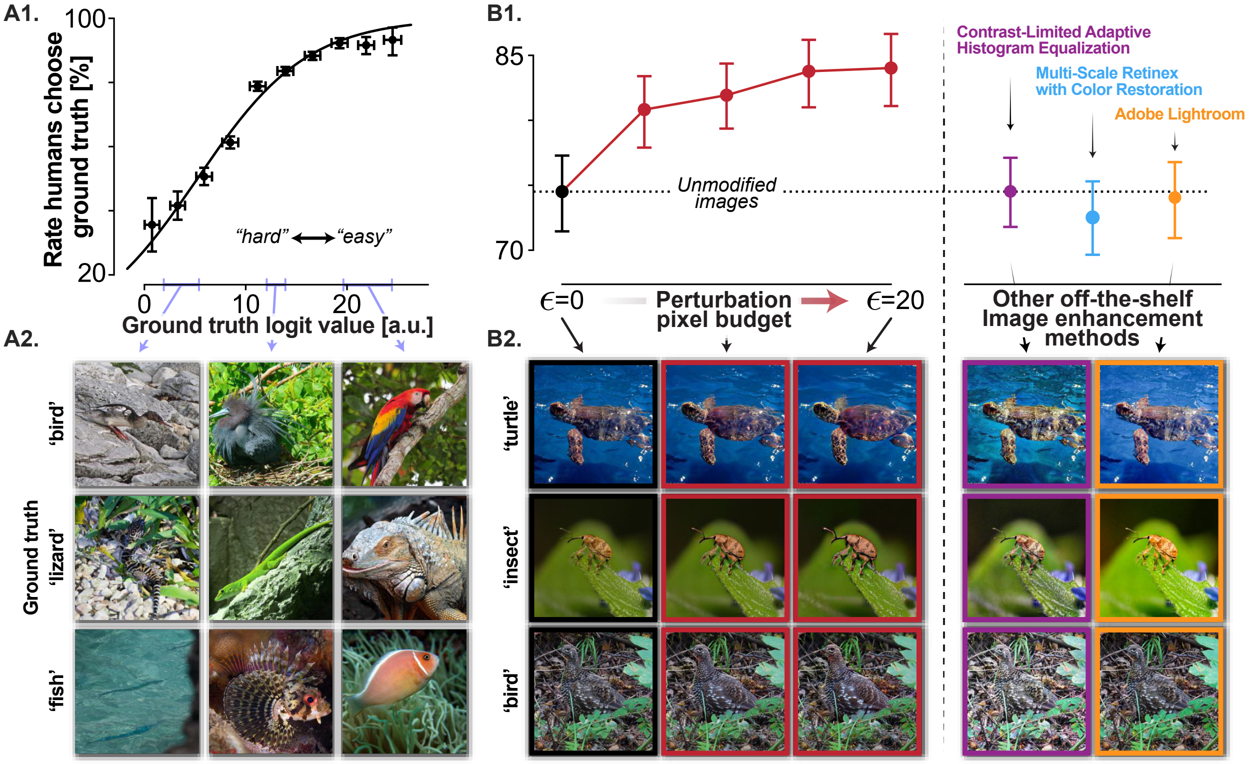


Figure 3.1: **Robustified ANNs can be used out-of-the-box as image recognition difficulty estimators and ground truth percept enhancers.** We consider a 16-way basic animal classification task. Panel **A1** shows the correspondence between human categorization accuracy and model-computed ground truth logit activation values. The curve denotes a logistic regression model predicting the probability of a correct response using only the logit value ($p < 0.001$ from the Wald statistic, AUC $= 0.72$ under 10-fold cross validation). **A2** shows example images with varying ground truth logit values (predicted difficulty). **B1** shows how perturbing images via ground truth logit maximization increases human recognition accuracy progressively with the ℓ_2-norm perturbation 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** shows example images: unmodified (left), enhanced by ground truth logit maximization with pixel budgets $\epsilon = 10$ and $\epsilon = 20$ (middle), and enhanced by baseline off-the-shelf methods (to the right of the dotted line). All vertical error bars are 95% confidence intervals by bootstrap. Horizontal error bars in panel A1 show the standard deviation among images within each logit value bin.

3.2 Introduction

Over the last decade, artificial neural network (ANN) models have demonstrated superior performance as image-computable emulators of neural processing along the human and monkey ventral visual stream. Iterative efforts have developed models that are increasingly aligned with primate vision, as measured by their ability to predict both neural activity and behavioral responses [98]. Beyond prediction, a promising class of these models – "robustified" deep ANNs [99] – has been shown to enable the generation of image perturbations that predictably control both ventral stream neural activity [100] and human object categorization reports [101,102]. In our work, we ask whether the prediction and control capabilities of robustified ANNs can be used to enhance human performance at visual categorization tasks.

Beyond categorization of familiar visual categories, one practically important task is

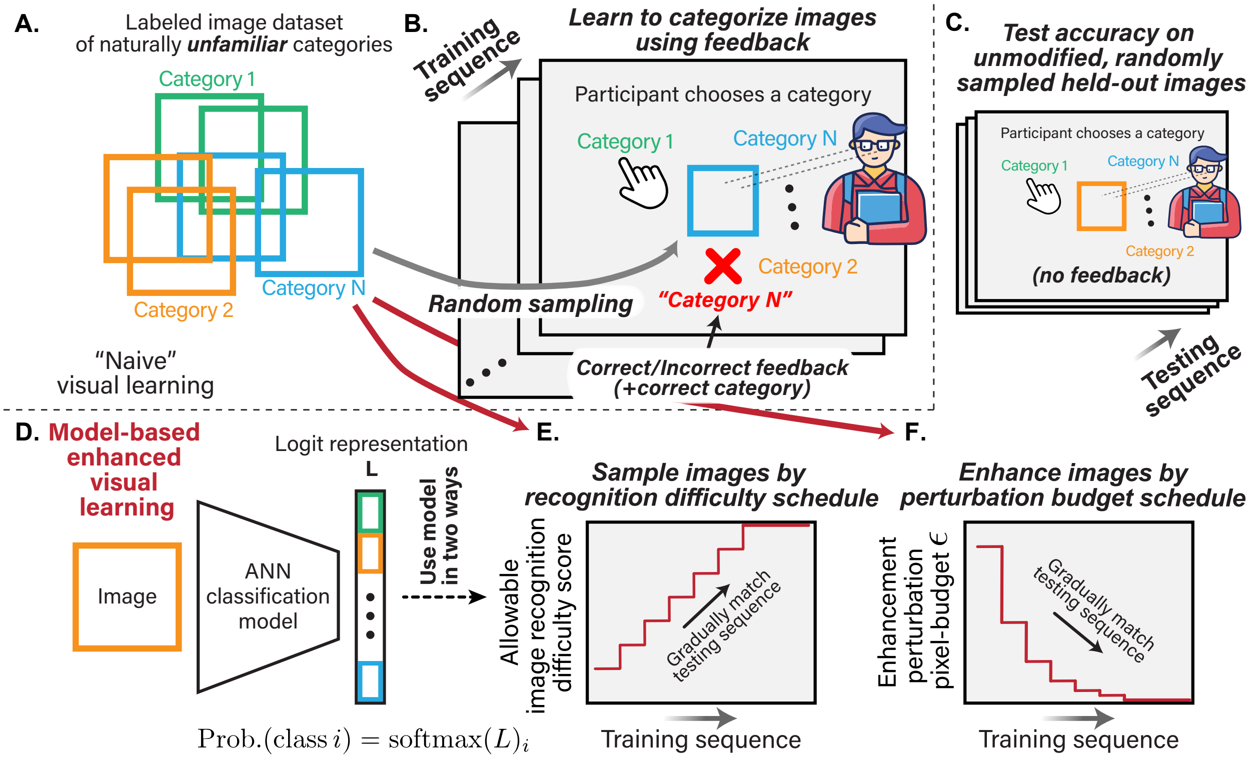


Figure 3.2: **Robustified ANNs can be used to boost image category learning in humans.** A novice human learner undertakes a challenging image categorization task, which consists of a training phase (**B**) and a test phase (**C**). Images for both phases are randomly drawn from a labeled image dataset of unfamiliar fine-grained categories (**A**). Feedback (correct/incorrect, with indication of the correct category) is delivered after each trial 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 curriculum by using the difficulty score to sample images based on a predefined increasing schedule of maximal difficulty per trial (**E**), and by enhancing images for easier recognition with an enhancement magnitude that decreases along a predefined schedule (**F**).

learning to recognize new, unfamiliar categories. Although humans can readily learn many new categories even from a single example [103], some consequential tasks require extensive training to reach high levels of performance. For example, medical specialists such as pathologists and radiologists devote numerous hours to mastering the diagnosis of various diseases from medical images. We identify a potential strategy to accelerate the visual learning process by extrapolating from the perceptual learning literature. Simple visual tasks, such as line orientation discrimination, reveal a curriculum effect whereby providing easy examples to a novice human learner, before gradually increasing the difficulty, promotes faster perceptual learning [104]. Motivated by these findings, we ask whether the human-aligned nature of robustified ANNs allows them to augment human learning in complex image domains, chiefly by reducing the initial task difficulty and then increasing the difficulty as learning progresses. A demonstration that these models can be used to enhance learning serves as an additional scientific test of the models' alignment with human perception, while also pioneering a potentially beneficial application of ANNs in education.

To test the viability of enhancing image category learning in humans, we first establish two key empirical observations, summarized in Fig. 3.1: (i) we find that the human error rate in an image categorization task is strongly predicted by the ground truth logit activation value of a robustified ANN, making it a valid image recognition difficulty score for humans; (ii) we find that this relationship also holds in reverse – pixel-level perturbations can be generated by the model to maximize the ground truth logit activation, producing an "enhanced" version of the image that is easier for humans to recognize as the ground truth category. While previous findings demonstrate that model-guided perturbations can modulate human perception *away* from the ground truth category [101], we conversely seek to *amplify* category-relevant features to facilitate correct classification. We observe that our model-based image enhancements yield significant increases in human accuracy on a classification task with basic animal categories, unlike conventional image enhancement techniques that adjust low-level image properties such as lighting and contrast. We thus propose an algorithm that combines model-based image enhancement and image difficulty prediction to generate optimized curricula for novice humans learning challenging image categorization tasks.

Our proposed method, "Logit-Weighted Image Selection and Enhancement" (L-WISE), is illustrated in Fig. 3.2. In our primary experimental setting, a human participant learns an unfamiliar image categorization task by viewing a series of examples, providing a category judgment for each image before receiving feedback indicating the correct category (Fig. 3.2A,B). Upon completion of the training phase, test accuracy is measured on held-out images in similar trials without feedback (Fig. 3.2C). L-WISE intervenes on this naive visual learning baseline by using a robustified ANN model in two ways: (i) it uses the ground truth logit difficulty score to sample training images based on a predefined, *increasing* schedule of maximal difficulty per trial (Fig. 3.2D,E); (ii) it enhances training images with a perturbation magnitude that *decreases* along a similarly predefined schedule (Fig. 3.2D,F).

Despite the human visual system being well-adapted for rapidly learning new visual categories, we find that L-WISE gives rise to substantial accuracy gains of 33-72% in test-time accuracy margins above chance relative to control participants. In addition to improved accuracy, L-WISE also significantly reduces the training phase duration by 20-23% with a constant number of trials. We demonstrate these effects across three varied image domains and category spaces: moth species in natural photographs, skin lesions in dermoscopy images, and pathologic findings in histology images.

Our main contributions are:

- We establish a new state-of-the-art in predicting image recognition difficulty for humans, using the robustified ANN ground truth logit activation value as a simple but effective difficulty metric.
- We show that robustified ANNs can guide image perturbations that enhance the ability of humans to accurately report the associated ground truth category label.
- We propose a novel model-based visual learning augmentation approach for humans that substantially increases test-time categorization accuracy and also reduces training time. To the best of our knowledge, this is the first application of image enhancement to augment human visual learning.
- We demonstrate the broad applicability of our proposed method in a variety of image domains, including clinically relevant dermoscopy and histology categorization tasks.

3.3 Related Work

We develop two important capabilities that form the foundation of our approach to assisting human learners: (1) state-of-the-art predictions of the recognition 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 training ANNs [105]. Leading approaches include the c-score learning speed proxy [106] and the prediction depth [107] calculated for each image. [108] applied both of these techniques to predict the recognition 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 a robustified ANN model's logit score associated with the ground truth class is a more accurate predictor of image difficulty for humans than prior methods.

Enhancing image quality has been the focus of many previous studies [109], ranging from correction of low-level properties such as lighting and contrast (e.g., [110,111]) to ANN models that "upsample" images to higher resolutions [112]. However, very little research has focused on enhancing images to more strongly represent a specific category. Previous works in this vein focused on making images easier for ANN models to correctly classify [113,114] or less vulnerable to subsequent adversarial attacks [115,116]. However, such methods do not strongly affect human responses due to perceptual misalignment between humans and naive ANNs [101].

Many studies have focused on model-human alignment. Brain-Score benchmarks models in terms of neural representations and behavior [98]. "Harmonization" methods drive alignment using an auxiliary objective on ANN-predicted feature importance maps and crowd-sourced human maps [16]. Other works introduce architecture components to account for additional aspects of human vision, such as the dorsal-stream pathway [117], or recurrent connections for contextual reasoning [118] and visual search [119].

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 [101,120]. In our present work, 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 augment human learning. The emerging field of machine teaching [121] employs machine learning to find or generate optimal "teaching sets" that can be used to train models or humans. While many such approaches have been successfully applied to training machine learning models (e.g., [122,123]), few studies have successfully enhanced image category learning in humans and most of these focus on teaching set selection. [124] 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. [125] extend a similar approach to select images in an online fashion by modeling the learner's progress. MaxGrad [126] 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 [127], 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 (i) a multi-class adaptation of STRICT, (ii) representativeness (mean feature-space distance to other images of the same class), and (iii) the estimated difficulty (entropy) of the CAM explanations. [128] 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 to learning augmentation 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. Our work is 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 [127] and [128] help learners by explicitly highlighting *where* learners should attend to in each image, we take a distinct and complementary approach by implicitly highlighting *what* learners should attend to in order to classify images correctly.

3.4 Overview of approach and experiments

Our approach to improving visual learning in humans is based on two key observations regarding human-aligned ANN models of the ventral visual stream: (i) they can accurately predict the recognition difficulty of specific images for humans (Fig. 3.1A), and (ii) they can be used to perturb images in a way that enhances the ability of humans to accurately report the original ground truth category (Fig. 3.1B). 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 design sequences of images that humans can use to more efficiently learn to recognize unfamiliar image categories. We test our approach in a variety of challenging image domains: natural photographs (moth species classification), dermoscopy images (skin lesion classification), and histology images (benign vs. pre-cancerous colon tissue classification).

3.4.1 Training task-specific robustified models

To obtain robustified models for task-specific category spaces, we adversarially trained ResNet-50 ANNs [85] on the ImageNet-1K [80] and iNaturalist 2021 [129] datasets (separately) using the same techniques as [99]. To adapt the resulting model to the three categorization tasks of interest, we conducted additional adversarial fine-tuning on the corresponding smaller datasets: a subset of moth species images from iNaturalist (after iNaturalist pretraining), the HAM10000 dermoscopy dataset [130] (ImageNet pretraining), and the MHIST histology dataset [131] (ImageNet pretraining).

3.4.2 Predicting category recognition difficulty

We propose a simple approach to predicting the human categorization error rate on specific images, in the form of a new image recognition difficulty score: the logit activation (pre-softmax) at the ground truth category output unit of a robustified ANN. The higher this logit value is, the lower the human categorization error rate. We establish this relationship through human participant responses during a natural image categorization task with 16 basic animal categories (Fig. 3.1A). We found this logit score to be the new state-of-the-art in predicting human error rates (see Appendix C, Fig. C.6).

3.4.3 Generating image perturbations to enhance category percepts

While the ground truth logit score predicts image difficulty at baseline, we show that we can also generate perturbations that maximize this value by backpropagating from the logit score to pixel space and running projected gradient ascent to update the pixel values, limiting the "pixel budget" ℓ_2 norm of the perturbation to a value ϵ. These perturbations make images easier for humans to recognize with respect to their ground truth labels (Fig. 3.1B). This approach is analogous to [101], but is designed to enhance the ground truth percept rather than guiding away from it.

3.4.4 Boosting image category learning in humans

Our approach to augmenting visual learning is summarized in Fig. 3.2. In the naive baseline scenario (Fig. 3.2A-C), the novice human learner is presented with a sequence of training trials through an online platform. In each training trial, a randomly selected image from one of N categories is presented, and the learner attempts to choose the correct category among N possible labels. The learner receives feedback after each training trial indicating the correct category label and whether or not their response was correct. In most of our experiments, to ensure that participants begin at the chance level and to avoid possible priors induced by the category names, we assign each category to a random Greek name unrelated to the task (with random assignments for each participant). After completion of the training phase, the experiment transitions into a test phase, in which the same task continues over a held-out set of images and no feedback is provided (Fig. 3.2C). During the test phase we measure the main visual learning outcome of interest, the test accuracy.

Harnessing key observations of robustified ANNs, our "L-WISE" approach uses one such 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. We adjust the "strength" of both mechanisms in a time-dependent manner: for the former, we set the maximum allowable difficulty score for an image to be shown in a given trial, and for the latter, we limit the enhancement perturbation magnitude to an ℓ_2 norm pixel-budget ϵ. The user of our approach can flexibly define arbitrary time-dependent profiles for image selection and enhancement (Fig. 3.2D-F). In this study, we used a step-wise linear ramp schedule for the allowable image difficulty at a given time, and exponential tapering of the enhancement ϵ. Intuitively, this should correspond to an easy-to-challenging trend during the training phase. Extensively optimizing these schedules was not a goal of our study, which serves primarily as a proof-of-concept. Applying image selection and enhancement led to significant gains in the accuracy of human participants on unmodified, randomly-selected test images, while also reducing the time needed to complete the training phase (which consists of a constant number of trials). This result was robustly observable across the varied image domains and category spaces we tested.

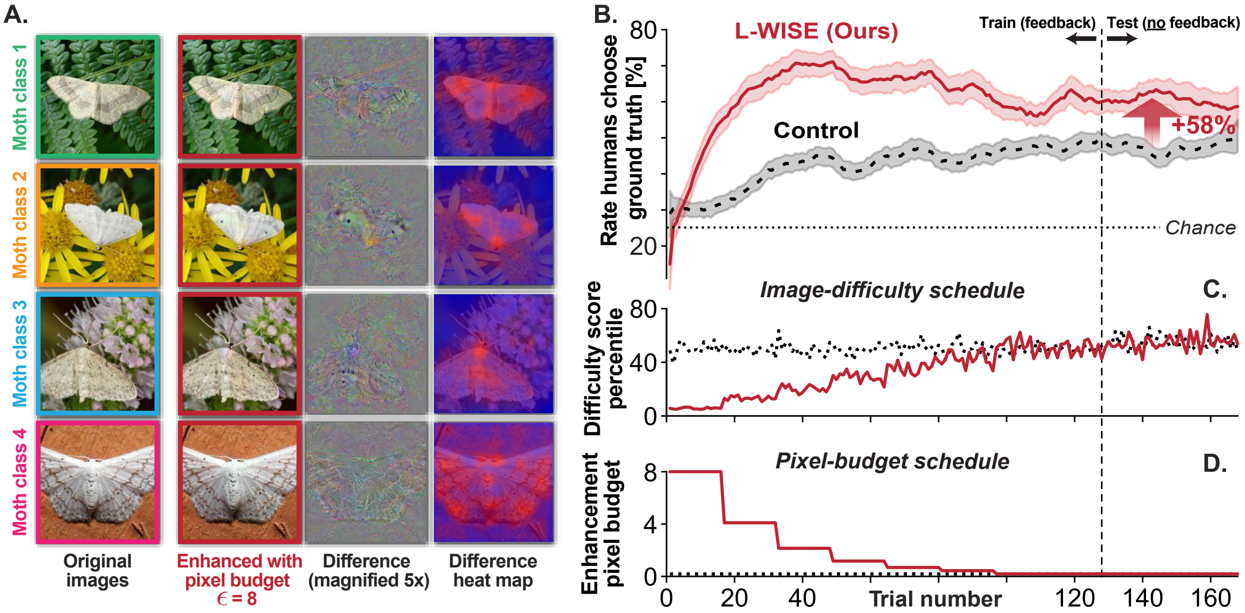


Figure 3.3: **Novice learners who had their curriculum augmented by our method showed improved test-time categorization accuracy for previously unfamiliar categories.** This figure shows empirical results from a 4-way fine-grained moth species classification task. Panel **A** shows examples of the 4 moth classes, side-by-side with their model-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 included are difference images showing the (5x magnified) difference between original and enhanced images, 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 the standard error of the mean. The test accuracy gain of the L-WISE group relative to the control group is statistically significant ($\chi^2(1)$ test, $p < 0.001$). **C, D** show the trial-dependent empirical profiles of the average image difficulty percentile of selected images, which (noisily) increases step-wise, and the perturbation pixel budget for enhancement (ϵ), which decreases step-wise. These profiles are uniform in the control group (black dotted lines), denoting randomly-chosen non-enhanced images.

3.5 Results

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

3.5.1 Robust models can both predict image recognition difficulty and reduce it

We tested the effects of a novel model-based image enhancement algorithm on human image categorization accuracy. We demonstrate that we can enhance images by maximizing the ground truth logit from a robustified ANN (ResNet-50) using gradient ascent in image pixel space. As the magnitude of the enhancement 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. 3.1B, chance = 1/16). While mean accuracy on the original,

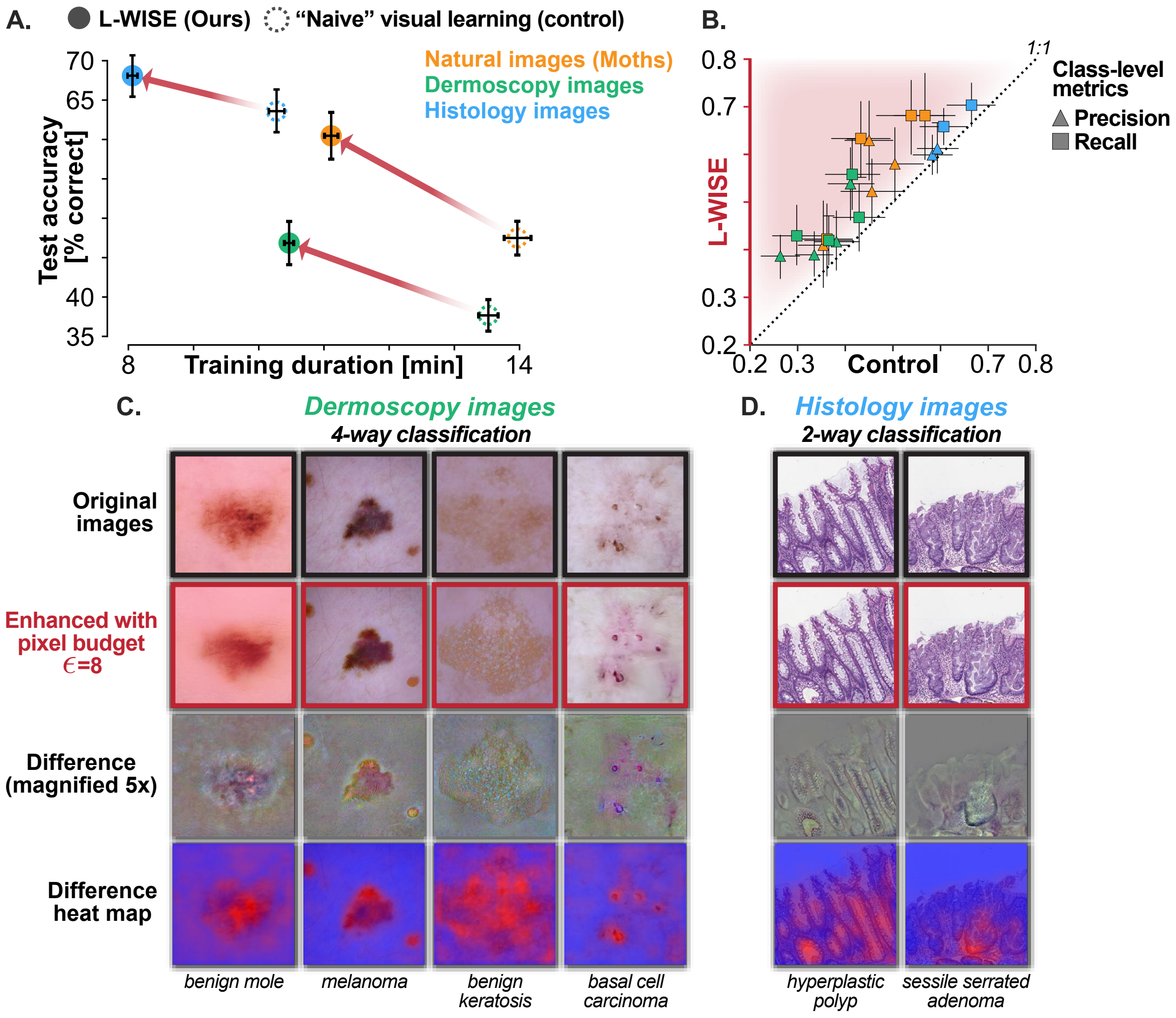


Figure 3.4: **Our approach can boost time efficiency and final accuracy of image category learning for humans across varied image domains, including in clinically relevant tasks.** Panel **A** compares the mean test-phase accuracy and training-phase duration of human participants who were randomized to L-WISE or control groups and learned a moth photo, dermoscopy, or histology classification task. All differences between L-WISE and the control group are statistically significant ($\chi^2(1)$ test, $p < 0.05$). Panel **B** shows 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. 3.3A.

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 approach 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 C, Fig. C.11: accuracy gains are larger for "difficult" images than for "easy" images. Baseline enhancement algorithms Contrast-Limited Adaptive Histogram Equalization (CLAHE, [110]), Multi-Scale Retinex with Color Restoration (MSRCR, [111,132]), and Adobe Photoshop Lightroom's "Auto" enhancement feature (LR, [133]) had no significant effect on human accuracy, 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 $L_{\mathrm{gt}}(x)$ 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. 3.1A). 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. We pooled responses to original images with those to modified control-group images that were not enhanced by robust models, recomputing L_{gt} for each image version (see also Appendix C, Fig. C.5). The logistic regression model (Fig. 3.1A) used L_{gt} to predict the binary correctness of the trial responses with Area Under the Receiver Operating Characteristic Curve (AUC) = 0.72 (10-fold cross-validation, $p < 0.001$ via Wald statistic). Notably, we find that this simple approach predicts the difficulty of individual images for humans more accurately than existing state-of-the-art metrics ([108], see Appendix C, Fig. C.6). In addition to ResNet-50, we demonstrate both difficulty prediction and image enhancement with XCiT vision transformers ([134], Appendix C, Figs. C.7-C.10).

3.5.2 L-WISE improves both test accuracy and learning speed for humans

We applied both image difficulty prediction and image enhancement as part of a novel method that designs image sequence curricula for novices learning challenging image classification tasks. Our proposed algorithm, "Logit-Weighted Image Selection and Enhancement" (L-WISE), operates on image trial sequences used to train human participants on image classification tasks through trial-by-trial feedback. The performance of each participant is evaluated in a subsequent testing phase without feedback, which includes only 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 stepwise-linearly increases as the training phase progresses. Selected images are enhanced at each trial during this period, via perturbations within an ℓ_2 pixel budget ϵ that decreases in a stepwise-exponential manner (Figs. 3.2E-F and 3.3C-D).

We tested L-WISE's efficacy at improving test-time accuracy and training duration of human learners on three challenging image category learning tasks (Figs. 3.3-3.4). Participants were randomly assigned to a control group with randomly-selected, non-enhanced images throughout the task, or to an L-WISE group. L-WISE increased the average test-time accuracy margin above chance levels by 57.6% on a 4-way moth species classification task ($p < 0.001$ on $\chi^2(1)$ test), by 72.3% on a 4-way skin lesion dermoscopy task ($p < 0.001$), and by 33.1% on a binary colon histology task ($p = 0.023$) (Fig. 3.4A). 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 was 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. 3.4A).

	Idaea moth photos		Skin lesion dermoscopy images	
	Mean acc.	Training duration	Mean acc.	Training duration
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 3.1: **Both image enhancement tapering (ET) and image difficulty selection (DS) contribute to the 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)" performs as well as or better than "DS"). Training durations are in minutes. Values in parentheses show 95% confidence intervals from 10,000 bootstrap replicates. * denotes significant differences in accuracy from the control group ($p < 0.01$, $\chi^2(1)$ test).

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

We tested several ablated versions of L-WISE (in the moth and dermoscopy tasks) to determine the relative contributions of its components: (A) image enhancement based on logit maximization, (B) selection of images according to logit-estimated difficulty, and (C) easy-to-hard curriculum trends enabled by A and B (Table 3.1). In "Enhancement Tapering" (ET), only the enhancement component of L-WISE is active, with random image selection as in the control group. Conversely, in "Difficulty Selection" (DS), images are selected based on difficulty but not enhanced. In "ET (shuffled)" and "DS (shuffled)," after applying ET or DS, the ordering of affected training trials is randomly permuted. Shuffling flattens easy-to-hard trends, isolating effects from (i) the mere presence of enhanced images in ET, and (ii) seeing easier images on average in DS (DS limits max. difficulty of early images, so DS/DS (shuffled) have easier training images than Control on average; see Fig. 3.3C).

The results show that both ET and DS have significant 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 (shuffled) 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 (shuffled) 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 comparisons 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 speed, however. On the moth task, training duration was 6% shorter for L-WISE than for the next fastest group, which was ET. Similarly, on the dermoscopy task, training duration was 5% shorter for L-WISE than for the next fastest group, DS (shuffled).

3.6 Discussion

In this study, we demonstrate that robustified ANNs can be used to both predict the empirical recognition difficulty of individual images for humans, and also generate enhanced versions of images that are easier for humans to correctly categorize. We harness these capabilities to develop a model-based curriculum design algorithm to augment 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 training speed and test-time classification accuracy on randomly-selected, unmodified images.

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 perturbed images and subsequently achieve superior generalization to unseen examples (Table 3.1). 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. 3.3A [135], or the irregular border and multiple colors that appear to be enhanced in the melanoma image in Fig. 3.4C [136]. Enhancements might also diminish features that distract from or contradict the ground truth: for example, in Appendix C, Fig. C.11B (second image from the left), 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 clearer class-relevant features and fewer distracting or contradictory features. These parallel explanations for the effects of image enhancement and image selection could help clarify 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 additive increases in learning speed were observed).

3.6.1 Limitations

This work is a proof-of-concept demonstration that robustified ANNs can be applied to augment image category learning in humans. We did not exhaustively search for optimal curriculum design strategies or image enhancement hyperparameters, nor did we study the "dose-dependency" of image enhancement or selection. L-WISE applies a fixed schedule of maximal image difficulty and image enhancement magnitude for all learners: human learning could plausibly be augmented more effectively by adapting the degree of image enhancement and the difficulty of selected images to the learner's progress in real time [104,137].

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 C, Fig. C.12H for an example). This clearly illustrates a task-relevant difference from melanoma (which tends to be asymmetric with irregular borders [136]), 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 [138]. These caveats must be thoroughly investigated before deployment of real-world educational applications of L-WISE, especially for clinical tasks with potential patient safety ramifications.

3.7 Methods

3.7.1 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 $L_{(1)} \geq L_{(2)} \geq ... \geq L_{(n_{c,s})}$. $n_{c,s}$ is the number of images for a given class c and training/validation/testing split designation s. We calculate class-specific difficulty percentile d_j for image j using the equation $d_j = \text{rank}(L_{(j)})/n_{c,s}$.

3.7.2 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 C, Section C.1 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:

$$x' = x + \underset{\|\delta\|<\epsilon}{\arg\max} \left(L_{\text{gt}}(x+\delta) - \frac{\alpha}{|C|-1} \sum_{c \in C : c \neq \text{gt}} L_c(x+\delta) \right) \tag{3.1}$$

In Equation 3.1, δ is a perturbation tensor of the same dimensionality as x and with 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 is the logit associated with class c (C is the set of all classes, with cardinality $|C|$). 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 three fine-grained image category learning experiments.

3.7.3 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 [139] with the jspsych-psychophysics plugin [140] for all experiments. All experiments included 10-12.5% attention check trials where the subject classifies an image of a circle or triangle (e.g., see Appendix C, Fig. C.4). We analyzed data from subjects 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 C, Section C.3). 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 [80]. Fig. C.3 (Appendix C) shows screenshots of the task as it appeared to participants. In our main experiment with this task (Fig. 3.1), each subject viewed interspersed images from 9 conditions: original images, images enhanced by maximizing L_{gt} with $\epsilon = 5, 10, 15$, and 20, images enhanced with one of three off-the-shelf baseline methods, and images disrupted by minimizing L_{gt} (internal control). 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 correct with ≥ 1 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 were 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 C, Fig. C.4). Each session consisted of 8 training blocks of 16 trials each (4 per class, or 8 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 a 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, except the ablation study control groups of ∼60 participants (overall min. 27, max. 68).

3.7.4 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_b did not require extensive hyperparameter tuning. After a pilot experiment in which we decreased ϵ linearly starting from $\epsilon = 20$, (see Appendix C, Fig. C.12H-M), we switched to the ϵ schedule above and changed no other hyperparameters for any of the three learning tasks/image domains. In the “shuffled” versions of DS and ET (Section 3.5.3), images in blocks 1-6 are selected or enhanced before a constrained shuffling procedure, whereby each image in blocks 1-6 may switch positions with any other of the same class regardless of d or ϵ.

3.8 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 Boston Children's Hospital's Institutional Review Board. 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 Office of Clinical Investigations at Boston Children's Hospital, 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 C.2 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 specialties such as pathology, radiology and dermatology where visual perceptual learning is particularly important. 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 [130]. 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 [141]. Models trained on this dataset are known to perform poorly for patients with darker skin [138], 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 [142]. 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 ground truth 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 applications of our method in this domain.

Chapter 4

Accelerating Histology Learning with Perceptually Aligned AI: A Randomized Study of First-Year Pathology Residents

4.1 Abstract

BACKGROUND

Pathology residency curricula are being squeezed at both ends: expanding requirements in molecular diagnostics and genomic medicine leave less time for traditional histologic training in surgical pathology, while the widespread transition to integrated medical school curricula without dedicated pathology courses often leaves graduates with limited basic histology skills. To address this educational bottleneck, we propose leveraging perceptually aligned artificial intelligence (AI) to accelerate visual perceptual learning for pathology residents.

METHODS

We developed a novel perceptual learning tool based on state-of-the-art artificial neural network (ANN) models of human visual recognition. We conducted a prospective randomized controlled study with 147 first-year pathology residents, comparing AI-assisted learning to unassisted practice. Residents learned to classify breast and prostate biopsies using multi-magnification image patches. ANNs assisted residents in two ways: (i) designing a progressively challenging curriculum by predicting case-level interpretation difficulty, and (ii) visually enhancing diagnostically relevant features. We further characterized the mechanisms of these two strategies in a four-arm mechanistic ablation study with 466 lay participants.

RESULTS

The proposed ANN-assisted learning approach yielded both improved accuracy and improved training times. These improvements depended on both the specific dataset/tissue type and the timing of post-tests. The ablation study indicated that ANN-sequenced curricula primarily improved accuracy on easier cases, while feature enhancement improved accuracy on difficult cases.

CONCLUSIONS

ANN-based assistance has the potential to accelerate histology learning and address a key bottleneck in pathology residency. Our results suggest that AI-designed curriculum sequences can rapidly build competence on straightforward cases, while a novel feature highlighting approach shows promise in preparing trainees for more challenging or borderline cases. To the best of our knowledge, our work is the first demonstration of AI-assisted perceptual learning in an applied educational setting, establishing a framework that could be extended to radiology, dermatology, cardiology (e.g., EKG), and other visually-intensive clinical specialties.

This chapter is based on a manuscript in preparation for journal submission, authored by Morgan B. Talbot, Anurag Vaidya, Kenneth Chian, Marina Muratova, Richard N. Mitchell, Guy Gaziv, and Gabriel Kreiman. Author contributions are detailed in Appendix A.

4.2 Introduction

Artificial intelligence is poised to transform pathology practice. The consensus among computational pathology experts, however, is that a range of barriers to AI adoption, such as domain shift [143], limited generalizability [144], and legal and ethical issues [145], will mandate expert human involvement in visual diagnosis for the foreseeable future [146]. While histology is therefore likely to retain its primacy in pathology training, it is increasingly de-emphasized in medical schools with the adoption of integrated curricula [23]. Meanwhile, rapid advances in genomic medicine and molecular diagnostics are expanding the scope of residency training and reducing the time available for traditional histologic training in surgical pathology [21,22]. Furthermore, shortening residency programs and increasing the number of training slots are among the proposed approaches to addressing global pathology workforce shortages [147]. New approaches are required to facilitate the rapid development of histology expertise within a shortened time frame.

Pathologists generally interpret histology images in multiple stages: locating diagnostically relevant regions-of-interest (ROIs) at low magnification, recognizing perceived features within ROIs, and making diagnostic decisions based on those features [148,149]. Among these stages, a 2023 study found that accurate histopathological feature recognition was by far the strongest predictor of overall diagnostic accuracy among residents and attending pathologists, controlling for experience level [149]. While ROI selection matures early in residency, feature recognition is the primary driver of diagnostic accuracy improvements throughout years of residency: its slow development creates a bottleneck in residency training [150]. We propose leveraging artificial neural networks (ANNs) trained in histopathology tasks to accelerate the development of visual pattern recognition skills in pathology residents.

We build upon the principle that short-duration, feedback-based practice promotes durable perceptual learning [151,152], employing a "flash card-style" intervention format shown to be strongly favored by pathology trainees [153]. Brunyé et al [149] called for enhancing this framework with established principles from learning science, particularly *fading*, which presents learners with easy or exaggerated examples before progressing to more difficult ones [154], and *feature highlighting*, which draws attention to diagnostically relevant features [155]. Our prior work provided the foundational proof of concept for automating these strategies in complex image domains, including histology [156]. We showed that the learned representations of perceptually-aligned "robustified" ANNs [99,101] can be leveraged to both predict image difficulty and enhance category-specific image features, both of which can be used for fading curricula [156].

The present study translates these techniques to an AI-driven visual category learning tool designed for anatomic pathology education, which we rigorously evaluated in a prospective randomized controlled longitudinal study with 147 first-year pathology residents across 53 institutions. We designed tasks with multi-magnification image sets from single patient cases, focusing on benign, malignant, and atypical presentations in H&E-stained breast and prostate histology. We employed fading curricula based on image difficulty predictions following previous work [156]. We adapted category-specific feature enhancement—previously used for fading—into a novel feature highlighting mechanism during trial-by-trial feedback, allowing residents to directly compare original and AI-"exaggerated" images to visualize salient features. The AI intervention measurably accelerated visual perception learning overall, though effects varied by task and time point. Intervention group participants achieved statistically significant gains in test-phase accuracy, response time, or training efficiency in both breast and prostate histology at different phases of the study. Our approach has the potential to meaningfully accelerate histology learning, addressing a key bottleneck in increasingly pressured residency curricula.

4.3 Methods

4.3.1 Study design and participants

We conducted a prospective randomized controlled study to assess the added value of the AI-based curriculum sequencing and feature highlighting within an educationally and clinically relevant visual perceptual learning activity. We recruited residents in their first weeks of training in anatomic pathology (AP), combined anatomic/clinical (AP/CP), and combined anatomic/neuropathology (AP/NP) programs from across the United States (via emails to Program Directors). Participants were randomized upon enrollment (July 2-31, 2025) to Control and Intervention groups in a double-blind design. All participants provided informed consent, and the study followed a protocol approved by the Institutional Review Board of Boston Children's Hospital. Full recruitment and compensation details are in Appendix D, Section D.1.

4.3.2 Design of histology tasks

Participants practiced classifying hematoxylin-and-eosin (H&E)–stained histology images, with breast and prostate images forming two separate task sequences (distinguishing benign vs. malignant for prostate, and benign, atypical, or malignant for breast). Images were sourced from the BRACS [158] and DiagSet [159] datasets. To ensure a high-quality "teaching set" for *both* Intervention and Control groups (Figure 4.1C), we eliminated blurry images using frequency spectrum thresholding, and removed images lacking epithelium (e.g., stroma, slide background) using a classifier trained on UNI-v2 foundation model embeddings (details in Appendix D, Section D.3).

Participants accessed the tasks using a custom web-based interface. Each trial presented four patches from a single patient tissue source: one at 5x magnification (top-left), one at 10x (top-right), and two at 20x (bottom). Participants clicked buttons on the screen to classify each case (prostate: "benign"/"malignant"; breast: "benign"/"atypical"/"malignant") within a 15-second time limit, followed by an untimed feedback screen (Figure 4.1B). Residents completed three sessions with flexible scheduling within pre-specified time windows (Figure 4.1A). Session 1 included "Training Phase 1" (all Training Phases are 96 trials with feedback) and "Test 1" (all Tests are 48 trials without feedback, identical across participants) for both tissue types, followed by a feedback questionnaire (Appendix D, Figure D.6). Session 2 (14-21 days later) included "Test 2a"; Intervention participants subsequently completed an additional "Training Phase 2" and "Test 2b." Session 3 (60-74 days after Session 1) included "Test 3" and a follow-up questionnaire (Appendix D, Figure D.7). Tissue type ordering (i.e., prostate first or breast first) was counterbalanced among participants but consistent within each

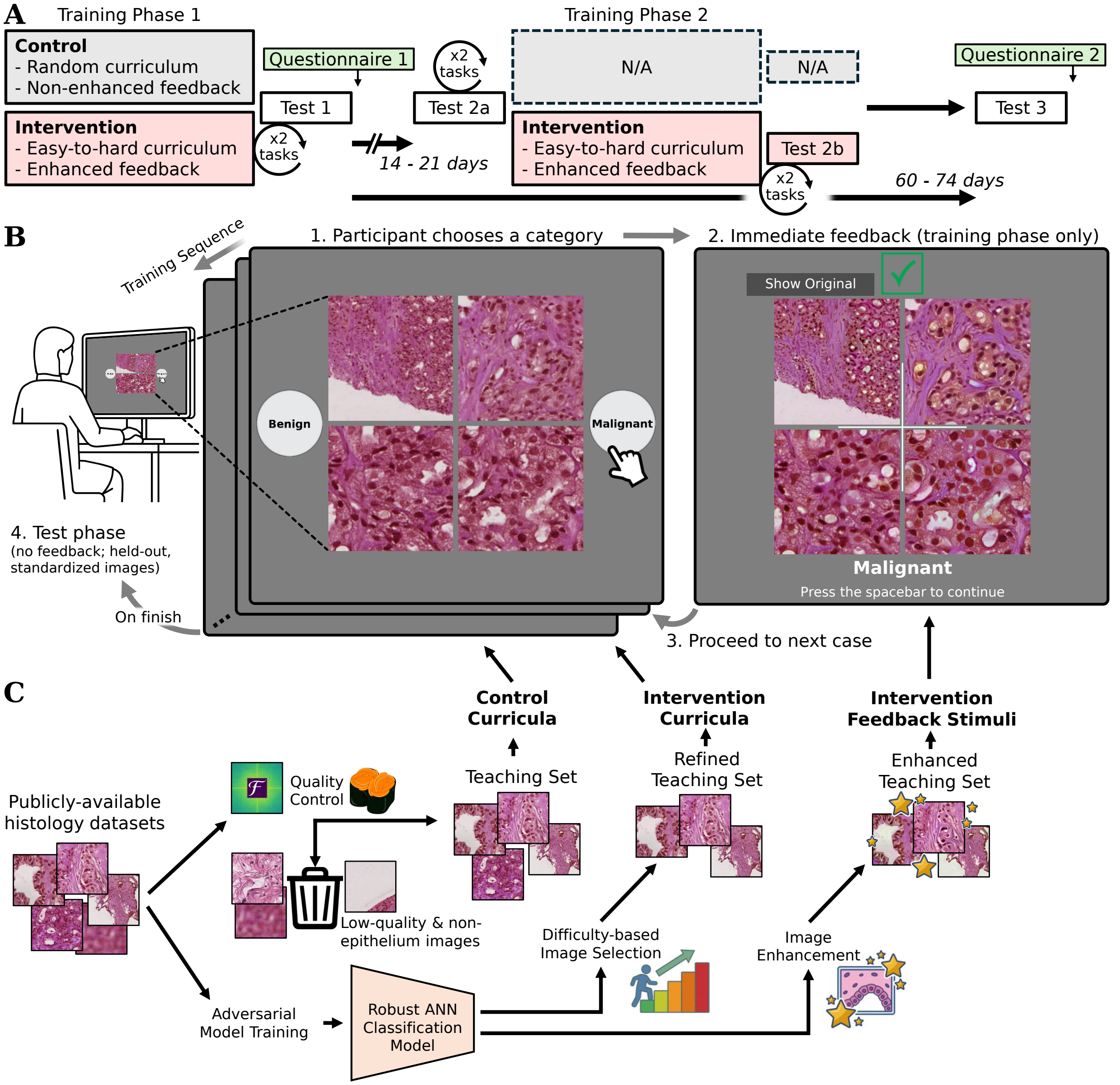


Figure 4.1: **Outline of longitudinal study and our proposed approach to AI-assisted histology learning.** Panel A illustrates the structure of each data collection session in the main resident study. Panel B shows the task interface seen by participants, with some elements enlarged for clarity (e.g., the "Show Original" button, which instantaneously toggles between original and enhanced image versions, and reads "Show Exaggerated" in its other state). The white "+" sign was visible within the 2×2 grid of images whenever enhanced images were shown. Panel C illustrates the image processing pipeline from source datasets to task trial sequences. Training sequences ("curricula") for both Intervention and Control groups are sampled from a quality-controlled teaching set, which excludes images that are blurry or do not include breast/prostate epithelium (e.g., stroma, slide background) via Fourier analysis and foundation model-based [157] approaches respectively. The raw datasets are used to adversarially train a ResNet-50 classifier, which is used for difficulty-based image selection to sample the Intervention group curricula, and for image enhancement to be shown to Intervention participants on the feedback screen.

participant.

4.3.3 Design of AI-based perceptual learning assistance

The learning intervention consisted of three pedagogical components: (i) difficulty-based image selection, (ii) difficulty-based curriculum sequencing, and (iii) category-specific image enhancement during the feedback phase of each trial. Both predictions of relative image difficulty (used for both image selection and sequencing) and category-specific feature enhancement were accomplished by repurposing ANNs that were trained to classify either breast (BRACS) or prostate (DiagSet) histology images (further ANN training/evaluation details in Appendix D, Section D.2).

Difficulty-based image selection and sequencing

For Intervention group participants, Training Phase images were selected and arranged in curriculum sequences based on difficulty percentile scores predicted by "perceptually aligned" ANN models. These models were trained to capture key aspects of human visual processing by learning invariances to subtle image perturbations that are imperceptible to humans yet profoundly disruptive to conventionally trained ANNs [99,101]. We previously demonstrated that the confidence of models trained in this manner is a strong predictor of human visual recognition difficulty [156]. Accordingly, our difficulty metric was defined as the model's pre-softmax logit score corresponding to the ground truth class, normalized within each class and each of the three magnification levels.

For the DiagSet dataset [159], ground truth labels were either "benign" (original class "N") or "malignant" (Gleason grades 3-5). The corresponding logit directly determined the difficulty percentile of each image patch. The BRACS model was trained on the original seven BRACS classes [158], which combine to form three superclasses in the learning task: "benign" (normal, pathological benign, usual ductal hyperplasia), "malignant" (ductal carcinoma in situ, invasive carcinoma), and "atypical" (flat epithelial atypia, atypical ductal hyperplasia). Within-superclass difficulty percentiles were based on a synthetic superclass logit, calculated by summing the probabilities (softmax-normalized scores) of the constituent subclasses and then converting back to a logit (log-odds) score.

Each Training Phase was split into eight class-balanced blocks (twelve trials each, where each trial has four image patches sourced from a single patient case) for all participants. In the Intervention group, image patches to form the patient cases in each block were sampled from Teaching Set subsets with capped difficulty percentiles. Blocks 1-2 included images up to the 10th percentile difficulty, blocks 3-6 gradually increased the difficulty (capped at the 20th, 30th, 40th, and 50th percentiles, respectively), and blocks 7-8 were capped at the 60th percentile. In the Control group, images for each block were randomly sampled regardless of estimated difficulty. Further details are presented in Appendix D, Section D.4.

Highlighting diagnostically relevant histology image features

Category-specific image enhancements generated by perceptually aligned ANNs have been shown to significantly increase the ability of human observers to identify images by category, and to augment visual learning when used in an easy-to-hard "fading" curriculum [156]. Here, we adapted this enhancement technique to a novel feature highlighting strategy, aiming to explicitly draw learners' attention to category-specific features. The feedback screen after each trial in our study re-displayed the images from that trial (Figure 4.1B). In the Intervention group only, instead of identical copies of the images, category-enhanced versions were displayed (see Figure 4.4G-H and Appendix D, Figures D.2-D.3 for examples). The interface included a button enabling participants to instantaneously switch between "enhanced" and "original" versions. This toggling approach has been shown to effectively highlight subtle differences between similar images [160], such as prior vs. current mammograms [161].

Our proposed image optimization approach uses a single, high-rate step of projected gradient ascent, limiting the ℓ_2 norm of the perturbation to a fixed pixel budget. Unlike iterative methods that can generate artificial "hallucinated" objects [101,156], this single-step method largely restricts the model to amplifying existing patterns (see Appendix D, Figure D.4 for a qualitative comparison of artifacts from the two approaches). Further details regarding image enhancement are in Appendix D, Section D.5.

4.3.4 Mechanistic ablation study

To deconstruct the specific effects of difficulty-based image selection and feature highlighting, we conducted a secondary randomized controlled experiment with lay participants recruited via Prolific. Participants were randomized to one of four groups: (i) Control, (ii) Intervention (combined image selection and feature highlighting), (iii) Select-Only (image selection without feature highlighting), and (iv) Enhance-Only (feature highlighting without image selection). The task structure mirrored Session 1 of the resident study for a single task (prostate histology). Full details are provided in Appendix D, Section D.10.

4.3.5 Data analysis

Given the finite total population of US first-year pathology residents ($n \approx 550$), sample size was determined by maximum feasible recruitment within the enrollment window. The primary modified intention-to-treat (mITT) analysis included all participants who completed Test 1 (see Figure 4.1A) for ≥ 1 tissue type (N=121). Accuracy outcomes were analyzed using logistic generalized linear mixed models (GLMMs) with a group × time point interaction (Test 1 vs. Test 2a). Models included random intercepts for participant and test case, adjusting for difficulty, sequence position, task order, and prior experience. Bonferroni correction was applied to P values and confidence intervals (CIs) across the four primary accuracy comparisons (Tests 1 and 2a for breast and prostate). Secondary outcomes (response time, training duration) were log-transformed and analyzed using linear mixed models and linear regression, respectively.

Missing task performance data due to participant attrition were primarily handled using a likelihood-based approach within mixed models. As a sensitivity analysis, we performed multiple imputation by chained equations (MICE, 100 imputations) with predictive mean matching on participant-level aggregated measures and pooled estimates using Rubin's rules [162]. Full details on data analyses are provided in Appendix D, Sections D.7, D.8.1, and D.10.

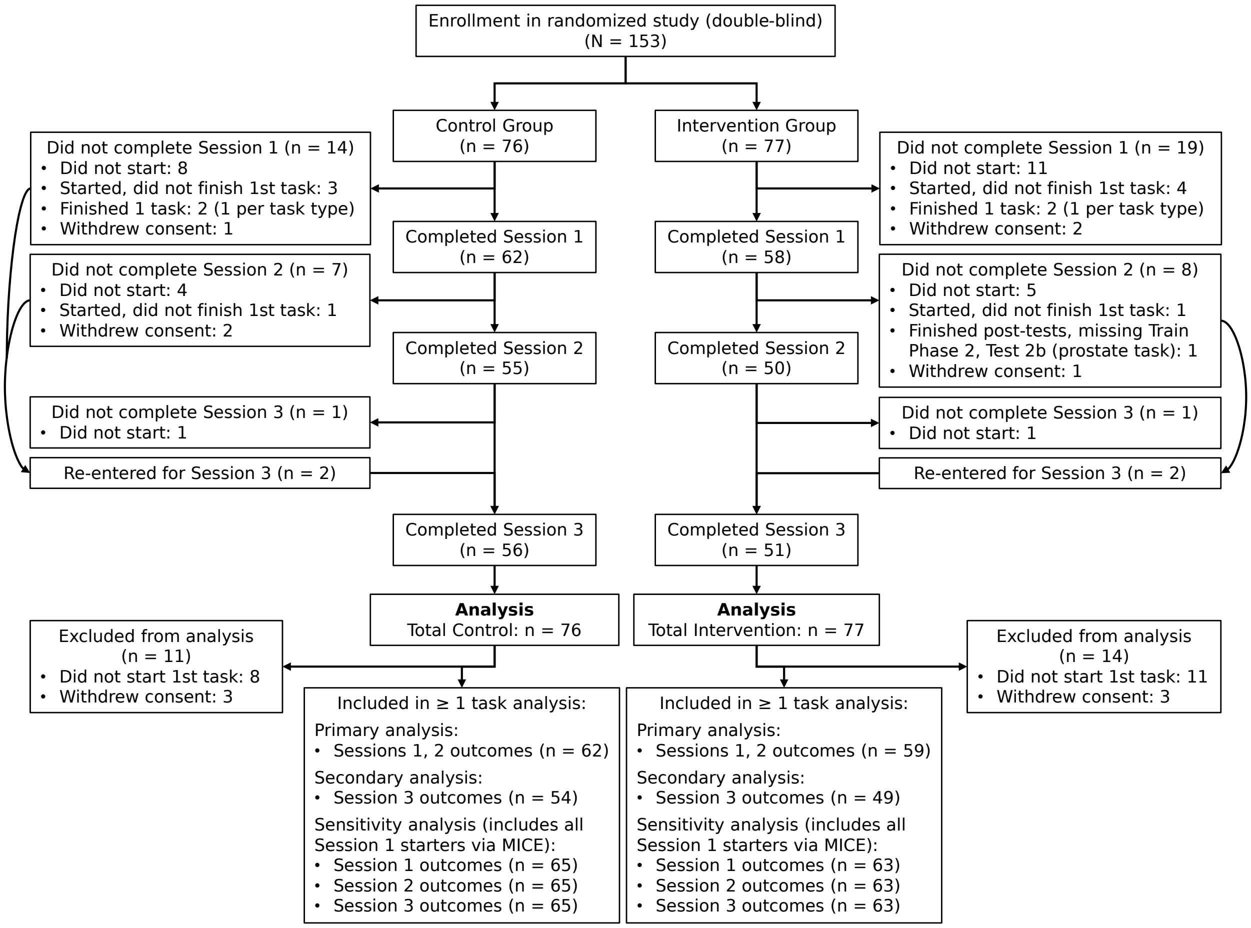

Enrollment in randomized study (double-blind)
(N = 153)
Control Group
(n = 76)
Intervention Group
(n = 77)
Did not complete Session 1 (n = 14)
• Did not start: 8
• Started, did not finish 1st task: 3
• Finished 1 task: 2 (1 per task type)
• Withdrew consent: 1
Did not complete Session 1 (n = 19)
• Did not start: 11
• Started, did not finish 1st task: 4
• Finished 1 task: 2 (1 per task type)
• Withdrew consent: 2
Completed Session 1
(n = 62)
Completed Session 1
(n = 58)
Did not complete Session 2 (n = 7)
• Did not start: 4
• Started, did not finish 1st task: 1
• Withdrew consent: 2
Did not complete Session 2 (n = 8)
• Did not start: 5
• Started, did not finish 1st task: 1
• Finished post-tests, missing Train Phase 2, Test 2b (prostate task): 1
• Withdrew consent: 1
Completed Session 2
(n = 55)
Completed Session 2
(n = 50)
Did not complete Session 3 (n = 1)
• Did not start: 1
Did not complete Session 3 (n = 1)
• Did not start: 1
Re-entered for Session 3 (n = 2)
Re-entered for Session 3 (n = 2)
Completed Session 3
(n = 56)
Completed Session 3
(n = 51)
Analysis
Total Control: n = 76
Analysis
Total Intervention: n = 77
Excluded from analysis
(n = 11)
• Did not start 1st task: 8
• Withdrew consent: 3
Excluded from analysis
(n = 14)
• Did not start 1st task: 11
• Withdrew consent: 3
Included in ≥ 1 task analysis:
Primary analysis:
• Sessions 1, 2 outcomes (n = 62)
Secondary analysis:
• Session 3 outcomes (n = 54)
Sensitivity analysis (includes all Session 1 starters via MICE):
• Session 1 outcomes (n = 65)
• Session 2 outcomes (n = 65)
• Session 3 outcomes (n = 65)
Included in ≥ 1 task analysis:
Primary analysis:
• Sessions 1, 2 outcomes (n = 59)
Secondary analysis:
• Session 3 outcomes (n = 49)
Sensitivity analysis (includes all Session 1 starters via MICE):
• Session 1 outcomes (n = 63)
• Session 2 outcomes (n = 63)
• Session 3 outcomes (n = 63)

Figure 4.2: **Participant recruitment and flow through the main study.** The largest source of attrition was participants who enrolled in the study but did not open the first task in Session 1. The primary analysis used a modified intention-to-treat population including all participants who completed Test 1 for the corresponding task (breast or prostate). A sensitivity analysis expanded this to include all participants who opened the first task in Session 1.

4.4 Results

4.4.1 Participants

A total of 153 residents were enrolled and randomized. Six participants withdrew consent and were excluded. The remaining 147 participants represented 62 institutions across 25 US states (demographics in Appendix D, Tables D.4-D.5). Nineteen participants (12.9%) did not initiate the first task and were excluded. The primary analysis included 121 participants who completed Test 1 for ≥ 1 tissue type, while a sensitivity analysis included 128 participants who started the first session. Figure 4.2 details participant flow.

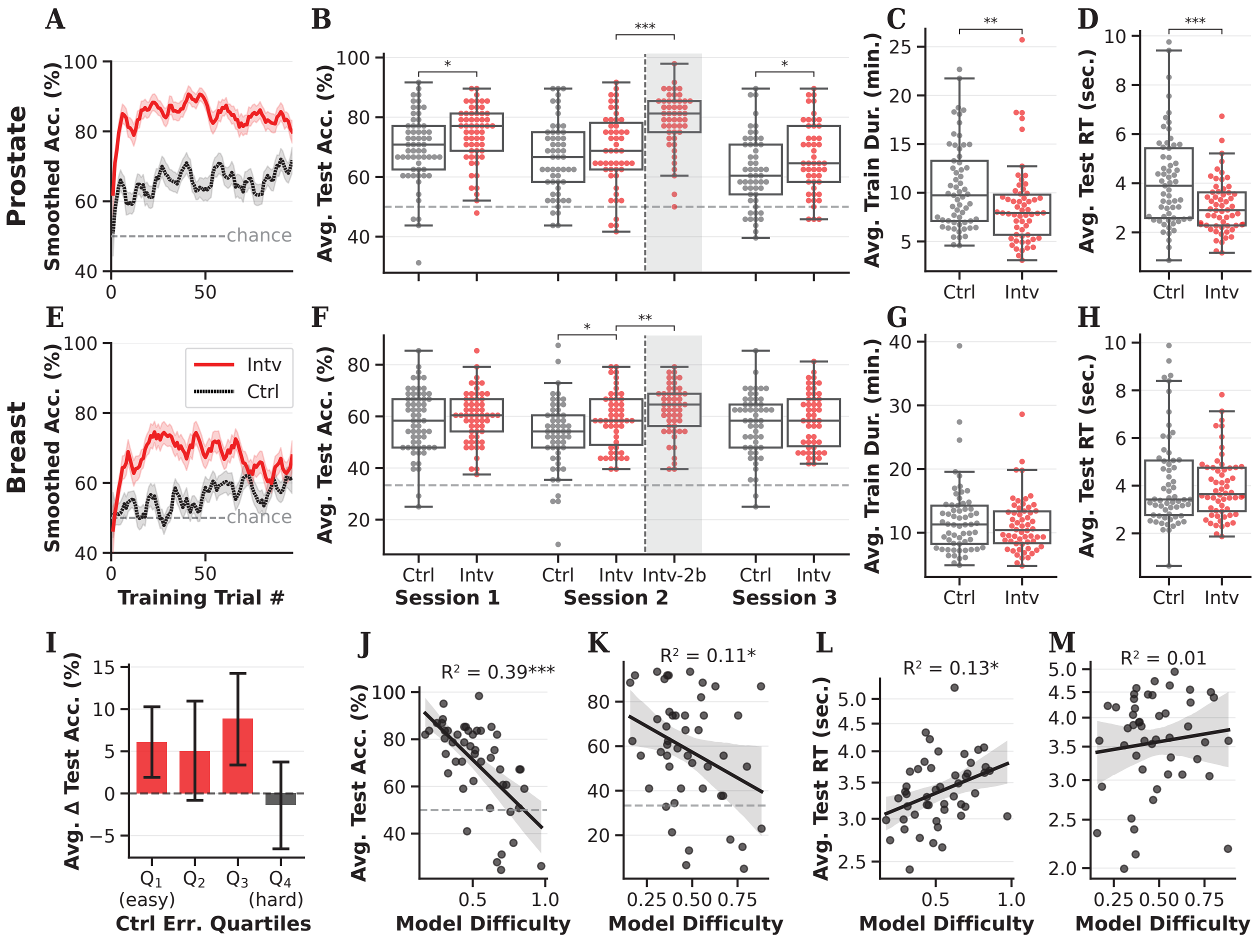


Figure 4.3: **Assistance from AI visual recognition models enhances histology learning in first-year pathology residents.** The top row of plots shows results for the prostate task, with breast results in row 2. Panels A (prostate) and E (breast) show learning trajectories across the 96 training trials, smoothed with a Savitzky-Golay filter and averaged across participants in Control (Ctrl, gray) and Intervention (Intv, red) groups. Shaded regions indicate the standard error of the mean. The dashed horizontal line denotes chance levels. Panels B, F are swarm plots showing test-phase accuracy across 3 study sessions. Each plotted point represents one study participant. Box plots indicate median and interquartile range (IQR), with whiskers at the farthest data points within 1.5×IQR of the box edges. "Intv-2b" refers to Test 2b, which follows a supplementary training session for Intv only (highlighted in shaded gray area). Horizontal dashed line indicates chance levels. Panels C, G show effective training duration for Ctrl and Intv groups, while D, H show test-phase response time (RT, correct trials only). Panel I visualizes differences in prostate Test 1 accuracy (Intv minus Ctrl) stratified by test case difficulty. Test cases were binned by the quartile of their error rate among Ctrl participants ("Ctrl Err. Quartiles"). Positive values of "Avg. Δ Test Acc. (%)" (in red) indicate higher Intv accuracy. Error bars represent 95% bootstrap confidence intervals. Panels J (prostate) and K (breast) show the association between Test 1 case-wise accuracy and model-estimated difficulty percentile (averaged among 4 images in each case), with linear regression fit lines, 95% confidence intervals, and R^2 values (coefficient of determination). Horizontal dashed lines indicate chance levels. Panels L (prostate) and M (breast) show the association between Test 1 case-wise RT (log-transformed; incorrect trials included) and model-estimated difficulty percentile. Panels J-M include Ctrl data only. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. P values are Bonferroni-adjusted within primary study outcomes (accuracy in Test 1 and Test 2a across both tasks). All other comparisons are not adjusted and should be interpreted as exploratory.

Outcome and Time Point	Control		Intervention		Between-Group Differences		
	N	Mean (SD)	N	Mean (SD)	Diff. in Means	Est. (98.75/95% CI)	Adj. P Val.*
Prostate Task (DiagSet)							
Test Accuracy (%)						*Odds Ratio*	
Test 1 (Immediate)	61	69.7 (11.7)	58	74.3 (9.5)	4.6	1.39 (1.05 to 1.85)	**0.01**
Test 2a (2 weeks)	55	66.4 (12.2)	51	68.9 (12.5)	2.5	1.17 (0.88 to 1.56)	0.63
Test 2b (Intv booster)	—	—	50	79.0 (9.5)	—	—	—
Test 3 (2 months)	54	61.8 (12.0)	49	66.5 (11.8)	4.6	1.31 (1.03 to 1.67)	**0.03**
Test Response Time (sec)						*Geo. Mean Ratio*	
Test 1 (Immediate)	61	4.13 (1.90)	58	3.06 (1.09)	-1.07	0.76 (0.66 to 0.88)	**<0.001**
Test 2a (2 weeks)	55	4.07 (1.97)	51	3.71 (1.32)	-0.35	0.95 (0.82 to 1.10)	0.48
Test 2b (Intv booster)	—	—	50	2.46 (0.97)	—	—	—
Test 3 (2 months)	54	3.71 (1.68)	49	3.38 (1.35)	-0.32	0.92 (0.78 to 1.10)	0.39
Effective Training Duration (min)						*Geo. Mean Ratio*	
Training Phase 1 (Initial)	61	10.6 (4.4)	58	8.6 (4.1)	-2.0	0.79 (0.68 to 0.91)	**0.001**
Training Phase 2 (Intv Booster)	—	—	50	6.8 (3.6)	—	—	—
Breast Task (BRACS)							
Test Accuracy (%)						*Odds Ratio*	
Test 1 (Immediate)	61	58.4 (12.2)	58	59.8 (9.7)	1.5	1.13 (0.86 to 1.49)	1.00
Test 2a (2 weeks)	55	53.2 (13.4)	51	58.6 (10.8)	5.4	1.36 (1.03 to 1.79)	**0.02**
Test 2b (Intv booster)	—	—	51	62.4 (9.8)	—	—	—
Test 3 (2 months)	54	56.9 (12.2)	50	58.7 (10.9)	1.7	1.14 (0.92 to 1.42)	0.22
Test Response Time (sec)						*Geo. Mean Ratio*	
Test 1 (Immediate)	61	4.19 (1.97)	58	3.94 (1.33)	-0.25	0.95 (0.83 to 1.10)	0.51
Test 2a (2 weeks)	55	4.35 (1.87)	51	4.11 (1.22)	-0.24	1.00 (0.86 to 1.15)	0.95
Test 2b (Intv booster)	—	—	51	2.68 (0.78)	—	—	—
Test 3 (2 months)	54	3.82 (1.35)	50	3.63 (1.39)	-0.19	0.95 (0.82 to 1.10)	0.50
Effective Training Duration (min)						*Geo. Mean Ratio*	
Training Phase 1 (Initial)	61	12.1 (5.7)	58	11.2 (4.3)	-0.9	0.94 (0.82 to 1.07)	0.36
Training Phase 2 (Intv Booster)	—	—	51	8.6 (4.1)	—	—	—

Table 4.1: **Primary study outcomes suggest an overall benefit of the AI-based intervention, with mixed results across tasks and time points.** Primary endpoint row headings are in bold. "Est." refers to exponentiated model coefficients: odds ratios of correct vs. incorrect responses in Intervention vs. Control for accuracy endpoints (analyzed using a generalized linear mixed model with logistic link function), and geometric mean ratios for response time and effective training duration (linear mixed models and linear models respectively, on log-transformed time endpoints). Mean and standard deviation (SD) values are not log-transformed. *P values for primary endpoints are adjusted with Bonferroni correction (4 comparisons). Adjusted (primary endpoints) or unadjusted (secondary endpoints) P values less than 0.05 are bolded. "Est. (98.75/95% CI)" includes Bonferroni-adjusted 98.75% confidence intervals for primary endpoints or 95% CIs for secondary endpoints.

4.4.2 AI assistance yields task-specific benefits for skill acquisition and retention.

Our primary endpoints compared test-phase accuracy between Control and Intervention groups on the immediate Test 1 and the delayed Test 2a (Table 4.1, row headings in bold; Figure 4.3B,F). In the prostate task, Intervention participants outperformed Controls on Test 1 (74.3% vs. 69.7%; OR, 1.39; 98.75% CI, 1.05 to 1.85; adjusted $P = 0.01$). However, the difference at Test 2a was not significant (68.9% vs. 66.4%). Conversely, breast task accuracy was similar at Test 1 (59.8% vs. 58.4%), but was higher for the Intervention group at Test 2a (58.6% vs. 53.2%; OR, 1.36; 98.75% CI, 1.03 to 1.79; adjusted $P = 0.02$).

4.4.3 Secondary and sensitivity analyses: response time, training duration, and additional Intervention training session

We secondarily analyzed test-phase response times (RT) for trials with correct responses, and effective training duration, defined as the total time for which training stimuli were displayed (clipping to maximum 60 seconds on the feedback screen to remove long breaks, $< 0.5\%$ of trials). For the prostate task, in addition to having higher accuracy on Test 1, Intervention participants had faster Test 1 RT (geometric mean ratio (GMR), 0.76; 95% CI, 0.66 to 0.88; $P < 0.001$) and shorter effective Training Phase 1 durations (GMR, 0.79; 95% CI, 0.68 to 0.91; $P = 0.001$) than Controls. However, no significant differences in these metrics were observed on Test 2a, or on either of Test 1 and Test 2a for the breast task.

In the second study session, Intervention group participants completed an additional "booster" training phase (Training Phase 2) and post-test (Test 2b) for both tasks, while the third session included an identical delayed post-test (Test 3) for both groups (Figure 4.1). Training Phase 2 had a significantly shorter effective training duration than the initial session (within the Intervention group) for both prostate (22% shorter; 95% CI, 14% to 30%; n=50) and breast (24% shorter; 95% CI, 16% to 30%; n=51; paired t-tests on log-transformed times). Training Phase 2 significantly increased accuracy in Test 2b relative to the Test 2a baseline for both tasks (+10.4% for prostate; 95% CI, 7.7% to 13.0%; +3.8% for breast; 95% CI, 1.1% to 6.5%), while also decreasing Test 2b RT (32% faster for prostate; 95% CI 27% to 37%; 34% faster for breast; 95% CI, 30% to 37%). Intervention participants had higher accuracy than Controls on Test 3 for the prostate task (OR, 1.31; 95% CI, 1.03 to 1.67; $P = 0.03$). However, accuracy did not differ at Test 3 for the breast task, and RT did not differ for either task. Secondary analyses were not corrected for multiple comparisons and should be interpreted as exploratory. The imputation-based sensitivity analysis yielded results consistent with the main analyses (Appendix D, Section D.8.1 and Table D.6), although the long-term accuracy benefit observed in prostate Test 3 was no longer significant ($P = 0.08$).

4.4.4 Mechanistic ablation study: effects of image selection and image enhancement

We recruited 469 lay participants via Prolific (demographics in Appendix D, Table D.7). Accounting for pre-randomization technical errors (n=3), incomplete sessions (n=11), and quality control exclusions for prior histology familiarity (n=4) and excessive window switching (n=5), the analytic sample comprised 446 participants (96% inclusion; Appendix D, Figure D.8). Participants were randomized to four groups: (i) Control, (ii) Intervention, (iii) Select-Only, and (iv) Enhance-Only. Control and Intervention groups completed tasks identical to those in the main resident study. Select-Only and Enhance-Only were ablated versions of the Intervention, with only either the model-based image selection or the feedback-phase image enhancement component active respectively.

Ablation study results are shown in Figure 4.4 and Appendix D, Tables D.8-D.9. Intervention, Select-Only, and Enhance-Only groups all had significantly higher test accuracy than Control (OR [95% CI], 1.48 [1.24 to 1.76] 1.40 [1.18 to 1.67], and 1.21 [1.02 to 1.45]; Tukey-adjusted P, < 0.001, < 0.001, 0.02). Intervention group participants had significantly higher test accuracy than Enhance-Only (OR, 1.22; 95% CI, 1.02 to 1.45; $P = 0.02$).

Regarding efficiency, Intervention and Select-Only groups completed the training phase significantly faster than Control and Enhance-Only groups (all pairwise $P < 0.001$). Intervention participants also had faster test-phase response times than Enhance-Only (GMR, 0.83; 95% CI, 0.72 to 0.95, $P = 0.003$), but no other pairwise response time comparisons

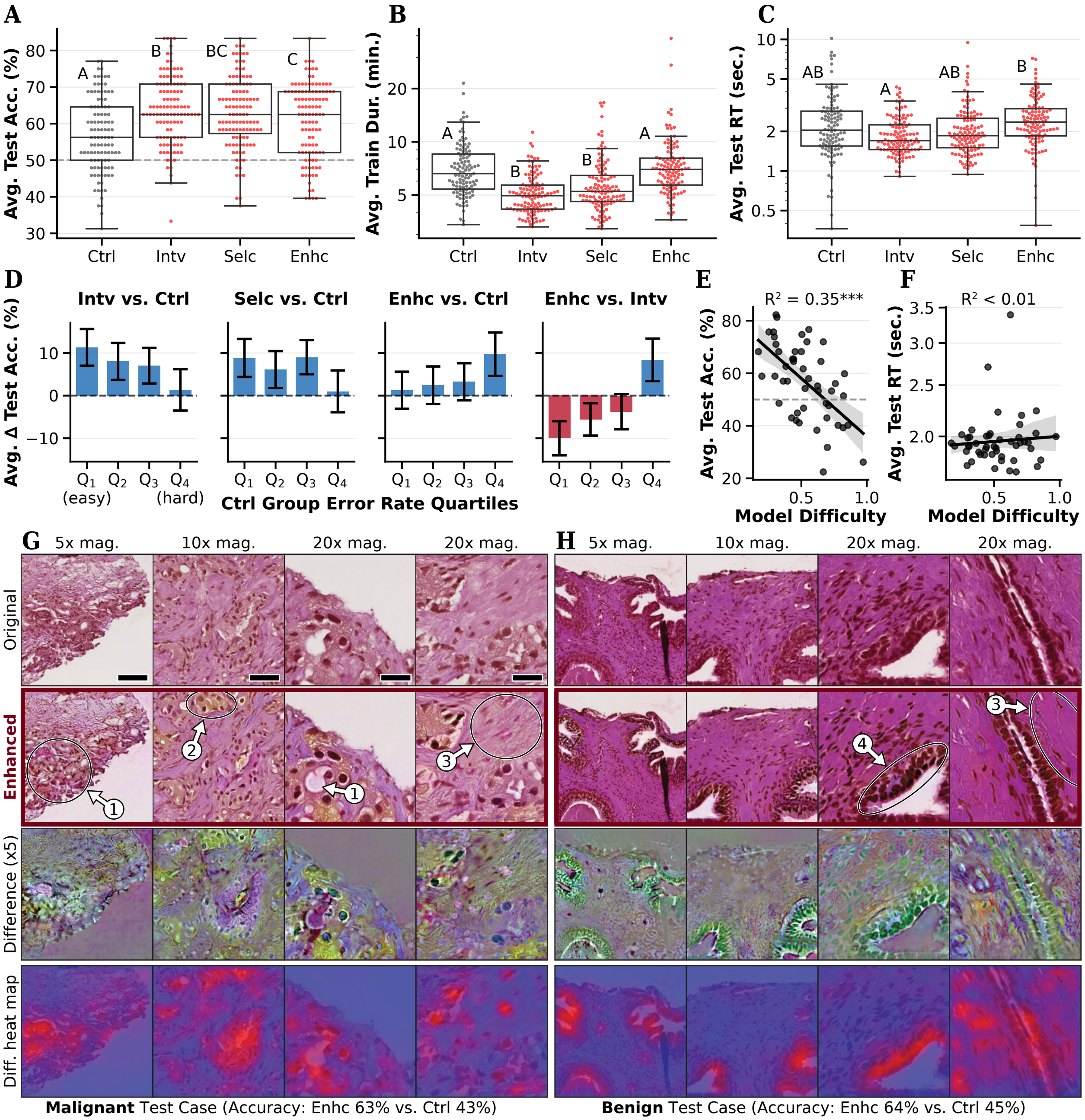


Figure 4.4: **Curriculum sequencing primarily benefits easy test cases, while histological feature highlighting improves performance on difficult cases.** In a mechanistic ablation study of the prostate task (N=446), lay participants were randomized to Control (Ctrl), Intervention (Intv), Select-Only (Selc), or Enhance-Only (Enhc). Panels A-C show test accuracy, effective training duration, and test response time (RT; correct trials only). Axes in B and C represent logarithmic scales. Each point represents one participant. Box plots show median and interquartile range (IQR), whiskers show farthest data points within 1.5×IQR of box edges. Groups sharing a letter label are not significantly different (Tukey-adjusted $P > 0.05$); disjoint letters indicate significant differences. Panel D shows differences in test accuracy between conditions, stratified by quartiles of case difficulty (Q1=easiest, based on error rates among Control participants). Panels E-F show associations between model-estimated difficulty (averaged among 4 images per case) and test accuracy/RT in the Control group (*** $P < 0.001$). Panels G-H visualize feature enhancements in example test cases on which the Enhc group was more accurate than Control (Panel G OR, 2.2; 95% CI, 1.3 to 3.8; Panel H OR, 2.2; 95% CI, 1.2 to 3.7; uncorrected CIs). Rows show original and model-enhanced images, difference images (intensity magnified 5x), and heat maps illustrating which image regions are most/least altered. Columns correspond to the 4 multi-magnification patches presented simultaneously for the test case. Scale bars are 100 µm (5x), 50 µm (10x), and 25 µm (20x). Arrows indicate biologically relevant examples of enhanced features. 1: Accentuated lumina. 2: Epithelial cytoplasm is highlighted, making it easier to distinguish from surrounding stroma. 3. Accentuated differences in collagen fibers, which are thicker, less dense, and less aligned in prostate adenocarcinoma [163]. 4: Accentuated double-layered epithelium (normal).

were significant after correction.

For the test accuracy endpoint, a significant interaction effect was observed between experimental group and test trial difficulty (estimated empirically by the error rate for each test image among Control participants; $\chi^2_3 = 44.3$, $P < 0.001$). On easy test trials (estimated for an image 1 SD below the mean empirical error rate), Intervention and Select-Only outperformed Control (OR, 1.96 and 1.68 respectively; both $P < 0.001$) while Enhance-Only did not ($P = 0.99$). However, for difficult trials (+1 SD), Enhance-Only was superior to both Control (OR, 1.43; 95% CI, 1.13 to 1.80; $P < 0.001$) and Intervention (OR, 1.29; 95% CI, 1.02 to 1.62; $P = 0.02$).

4.5 Discussion

This study shows that providing AI-based assistance during a histology learning activity can accelerate pathology residents' acquisition of visual perceptual abilities and/or improve retention, depending on the specific task and tissue type. Intervention group participants spent less time than Controls learning the same prostate task in Session 1, but reached higher test accuracy and faster test response times (indicative of enhanced fluency or automaticity, a hallmark of perceptual expertise [151,164].

While the intervention did not demonstrate significant immediate benefits for a breast histology task, higher accuracy on the 2-3 week delayed test suggests improved retention. Examination of response patterns by class (Supplementary Figure D.5) suggests that this difference was primarily driven by better discrimination between "atypical" and "malignant" presentations. Although these findings are counterintuitive, we speculate that learning this distinction was overly challenging at baseline but rendered more tractable with ANN assistance. The intervention may have sufficiently reduced cognitive load to bring participants down to an appropriate level of "desirable difficulty," which evidence suggests can support longer-term retention without obvious immediate gains [165]. Under this interpretation, the already easier binary prostate task may have been rendered suboptimally easy by the fading curriculum, resulting in rapid initial gains but with weaker retention due to shallower processing. Our approach could potentially be improved via adaptive methods that dynamically tailor difficulty for each learner [152].

The mechanistic ablation study indicated that the two intervention components, fading based on predicted difficulty and feature highlighting based on image enhancement, both enhance learning but via distinct processes. Our fading curriculum (capped at the 60th difficulty percentile) improved accuracy on easy-to-moderate test cases. Conversely, feature highlighting primarily benefited difficult test cases, but this advantage vanished when combined with fading. Since fading restricts training to easier images, this implies that highlighting is redundant for obvious features and most effective when applied to challenging training cases. In retrospect, the resident curriculum's difficulty cap likely blunted the impact of feature highlighting: future iterations should extend curricula to higher difficulty levels to maximize the complementary effects of both strategies.

Our work has several limitations. The primary training period for each task was less than 15 minutes on average. While this is comparable to previous work [151], the brevity of the intervention makes it difficult to extrapolate the effects of AI assistance during longer-term usage. Additionally, although randomization mitigates systematic biases, variable clinical exposures between sessions may have introduced unmeasured heterogeneity. The tasks in this study focus on recognizing features of malignancy in only two tissue types (plus atypia in breast), with varied results between breast and prostate tissues. While previous studies suggest that histopathological feature recognition is the main driver of diagnostic accuracy for residents, even with multiple magnification levels in each case, our tasks do not emulate visual search and integrative clinical reasoning processes that are integral to real pathology workflows (e.g., with whole-slide images). Finally, our methodology relies on task-specific neural networks trained on large, expert-labeled histology datasets, the public availability of which varies among tissue types and conditions. We speculate, however, that our approach could be adapted to use adversarially-trained foundation models with capabilities across a broad range of histology tasks.

Overall, our results suggest that the incorporation of AI-assisted perceptual learning tools could save time during increasingly crowded residency curricula. Most residents found the tool both relevant to their training and highly engaging, supporting the feasibility of its adoption (questionnaire results in Supplementary Figures D.6- D.7). Our work opens a new frontier of human-AI collaboration during clinical perceptual learning, with the potential to accelerate human expertise development and address educational bottlenecks.

Chapter 5

Optimizing Curricula for Human Visual Category Learning Via Image-Computable Surrogate Learners

5.1 Abstract

Task-optimized artificial neural networks have become the leading models of human vision: optimizing parameters for ethologically relevant tasks such as visual categorization causes convergence toward brain-like configurations. These models also show promise for accelerating human visual learning, but existing approaches rely on heuristics such as selecting examples based on model-predicted difficulty. This work proposes a goal-driven optimization approach to both learning and teaching. Human image category learning is simulated by image-computable "surrogate learner" models that learn visual tasks from scratch under human-like constraints. Curricula are then optimized to maximize surrogate learner performance using an evolutionary algorithm. For a dermoscopy classification task with four categories, a curriculum optimized using simple linear surrogates trained on neural network embeddings significantly enhanced human performance: participants achieved higher accuracy and faster response times, and also completed training faster, than controls who learned from randomly sampled curricula. Optimized curricula featured naturally-emerging easy-to-hard progressions, with ablations showing that both the selection of examples and their relative ordering contributed to curriculum efficacy. These results demonstrate that image-computable simulations of human learning can directly optimize educational interventions, an approach that will benefit from increasingly accurate models of learning processes.

5.2 Introduction

The acquisition of visual expertise is a key bottleneck in domains such as medical training, astronomy, and biological taxonomy. Achieving high levels of proficiency requires humans to internalize subtle, high-dimensional visual patterns, a process that often demands years of training. While automated machine learning-based tutoring systems hold the promise of accelerating this process, designing optimal instructional curricula for complex visual tasks remains an open challenge.

Many prior works that apply machine learning techniques to enhance visual learning in humans focus on specific educational software features informed by heuristics and prior evidence regarding conditions that enhance learning. These include progression from easy to more difficult examples (fading) [156] (see Chapter 3 of this thesis), delineating relevant image regions [127], and highlighting relevant features by exaggerating them through category-specific image enhancement (see Chapters 3 and 4 of this thesis). Such approaches are vulnerable to subtle confirmation biases. For example, in Chapter 3, although a selection-based fading procedure that progressed from easy examples to the full difficulty distribution yielded significant enhancements in human accuracy, the improvement was maintained or even strengthened when the curriculum sequences were shuffled, suggesting that the benefit was related to an easier distribution of teaching examples overall rather than the easy-to-hard progression as hypothesized. Such challenges may be partially mitigated by approaches such as that of Lindsey et al (2013, [166]), who showed how parameters of instructional policies, such as the rate of difficulty increase during fading, can be optimized using models that approximate the results of human experiments as a function of the parameters in a low-dimensional policy space. The effectiveness of certain heuristics may also vary in a task-dependent manner that is challenging to predict, exemplified by effects observed in Chapter 4 that depended on both task type and the timing of evaluations. Chapter 4 also demonstrated difficulty-dependent interaction effects between two different heuristic mechanisms (fading and feature highlighting) when applied simultaneously. These complex dependencies are likely to undermine the generalizability of heuristic-based learning enhancement findings across studies, tasks, datasets, learner populations, and other shifts.

This chapter aims to move beyond heuristic approaches to learning enhancement by simulating the entire end-to-end process of visual category learning with image-computable "surrogate learner" models, and using evaluation metrics derived from these models as objective functions towards optimization of the curricula they are trained on. A surrogate learner is defined as a machine learning model that simulates the process of human learning through two key properties. First, the surrogate learns a task of interest under similar conditions and constraints to a human learning the same task. In a typical visual category learning task, the human learner completes a series of trials in which an image is presented, a judgment is made about its category, and feedback is received indicating the ground truth category and whether the initial judgment was correct. Similarly, a surrogate model in this setting makes an inference about one image at a time before updating its state (parameters) based on that image, its own inference, and the ground truth label. This constraint derives from the hypothesis that the specific ordering of stimuli presented to the learner (the "curriculum") plays an important role in determining learning outcomes (e.g., accuracy on a held-out test after a period of learning).

The second desired property of a surrogate learner is that its learning dynamics, or behavior, resemble those of a human learning the same task. In particular, a valid surrogate should show a pattern of predictions correlated with those of a human who learns from the same curriculum. This may be reflected in (i) predicted class labels, (ii) confidence (which may be indirectly estimated in humans via response times for category judgments [137]), and (iii) aggregate statistics such as estimated mean accuracy at a particular point in (or after) the curriculum sequence [167]. Notably, this operational definition does not necessarily require that the surrogate has the same hypothesis class, learning algorithm, or learning mechanism as a human. A valid surrogate learner of sufficiently low computational expense can be used to rapidly evaluate the efficacy of a large number of curricula, far more than would be feasible by comparing curricula in experiments with human learners.

An additional strategy to quantify both a surrogate's behavioral resemblance to human learners, while also gauging its practice usefulness, is as follows: having somehow discovered a curriculum that leads to high surrogate performance on a held-out test set, evaluate whether the optimized curriculum enables improved human performance relative to non-optimized control curricula. While on the one hand this is merely a special case of the approach described above, the distinction is of practical significance because the space of all possible curricula is extremely large. For example, for a sequence of 50 images drawn from a modest dataset of 100 images in total, the number of possible curricula is around 3×10^{93}, much larger than estimates of the number of atoms in the observable universe [168]. A useful analogy is the vast parameter space of deep neural models: just as gradient descent can efficiently converge upon solutions with levels of performance that would be astronomically unlikely to be found through random initialization, optimization approaches might exist that can design curricula with levels of efficacy that would be effectively impossible to discover by chance.

One such optimization approach is developed in the present work: a genetic algorithm in which a population of curricula compete with each other to reproduce on the basis of evolutionary "fitness." Fitness is operationalized as curriculum efficacy, which is rigorously estimated by training a cohort of 50 surrogate learner models with different initial conditions on each curriculum and then evaluating their average accuracy on a large dataset. Notably, this gradient-free optimization approach treats the surrogate as a black box, making it readily extensible to any surrogate learner. It is also uniquely suited to the discrete, combinatorial nature of curriculum selection and ordering, a permutation problem where gradient-based optimization is not readily applicable. During the evolutionary curriculum optimization, image sequences are considered as analogous to genetic sequences: a crossover operation (analogous to crossing over in meiosis) allows highly effective curricula selected to mate with each other to be recombined into novel sequences, while point mutations that randomly replace individual images or swap their positions within a curriculum create random genetic variation in the population. Curricula are optimized over many generations of evolution, culminating in the selection of a single "champion" curriculum of maximal fitness, which is then evaluated to determine its effectiveness in teaching human learners.

As a starting point for surrogate learner modeling, I employ a linear classifier trained using a simple plasticity rule on image embeddings from a frozen artificial neural network (ANN) feature extractor, an approach shown in a previous study to exhibit a surprising degree of alignment with human learning dynamics [167]. While the evolutionary optimization approach is bottlenecked by the need to train and evaluate a large number of surrogates, a simple linear model can be used to evaluate a curriculum at very low cost: for example, the results in this chapter required over 900 million individual linear surrogate training/evaluation cycles within a parallelized and GPU-accelerated pipeline, consuming roughly 5 days of computation time on a desktop machine with a 20-core CPU and four 2080 Ti GPUs.

This work compares the performance of human learners on a medical image classification task (dermoscopy), having been trained with either a surrogate-optimized champion curriculum or a randomly-sampled one. The results show that a champion curriculum enables more rapid learning with a fixed curriculum length, and results in both higher accuracy and faster response times on a subsequent test with held-out examples. Surrogate-based analyses of champion curricula following optimization demonstrate that both the selection of images and their specific ordering is optimized. Further, champion curricula disproportionately contain relatively easy training examples, and feature a difficulty progression that varies as a function of the surrogate learning rate during optimization. To illustrate the general applicability of the present approach, I also demonstrate curriculum optimization in silico for a novel rule-based algorithm that tests explicit verbalizable hypotheses (as opposed to the linear surrogate, which integrates information across a high-dimensional embedding space), aiming to simulate explicit rule-based learning in humans [169]. To the best of my knowledge, this is the first demonstration of enhanced human visual learning for an image-based task using direct, image-computable simulations of the human learning process. I hope that this finding provides strong motivation for the development of increasingly high-fidelity models of human learning, with the knowledge that these models can be directly incorporated into educational interventions in important visual domains such as medical imaging.

5.3 Related Work

The present work is conceptually related to the framework of machine teaching, which aims to solve an inverse problem to that of machine learning: given a known learning algorithm and a target model, what is the optimal training set to reach that target? [121]. This approach has been successfully applied to select or generate optimal "teaching sets" for training machine learning models (e.g., [122,123,170]). However, applying these analytical frameworks to human learning of complex tasks, such as natural image classification, presents a unique challenge: the mismatch between the high-dimensional, non-linear complexity of visual learning and the need for analytically invertible learning algorithms. Leading image-computable models of human visual perception [11] rely on deep neural networks that lack closed-form analytical inverses. While recent work has begun to extend deep network representations to model the dynamics of visual category learning in humans [167], the dynamic plasticity rule updates used by these models render them analytically intractable within the original machine teaching formulation.

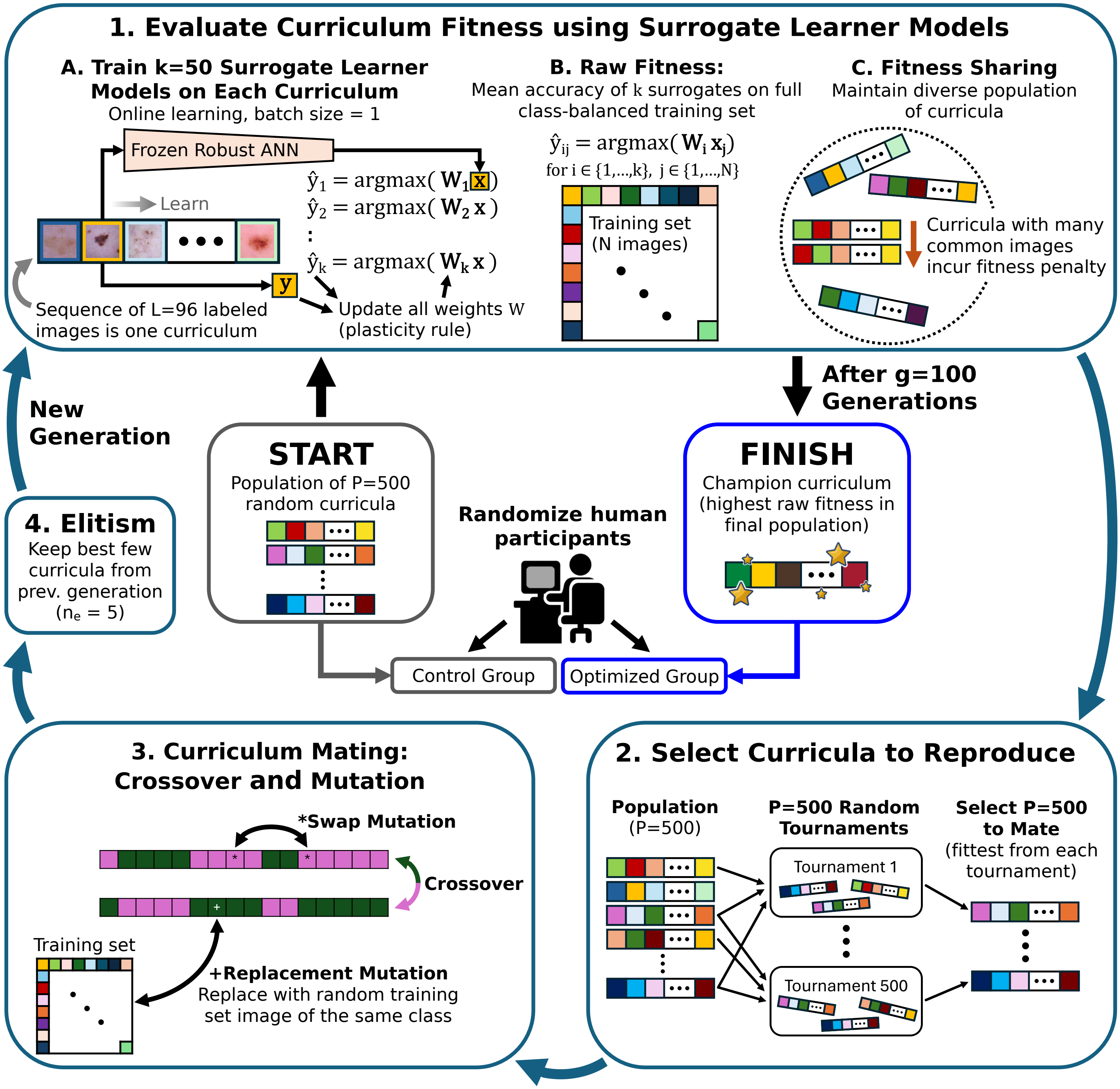
Early attempts to apply machine teaching to human learners focused on simple, low-dimensional tasks such as classifying lines by length or learning discrete mappings between numbers and letters, enabling the use of analytically tractable learning algorithms such as generalized context models [171] and partially observable Markov decision processes [172]. Singla et al. (2014, [124]) demonstrated marginal performance improvements in humans learning a natural image classification task, by selecting an unordered teaching set of 1-7 images that effectively constrains linear classification hypotheses based on crowd-sourced image embeddings (comparing with randomly-selected teaching sets of the same size). Johns et al. (2015, [125]) proposed an interactive machine teaching approach for human learners that adaptively selects teaching images to maximize expected error reduction on unseen images. This was accomplished by using Gaussian Random Fields to propagate learners' responses through a similarity graph constructed from fine-tuned convolutional neural network embeddings (CaffeNet/AlexNet [173,174]). Aggregating results across 4 challenging image classification tasks that human participants learned from short curricula (9-15 images), this approach showed an overall accuracy gain compared to random sampling and other baselines. Wang et al. (2021, [126]) modeled human visual category learning as optimal empirical risk minimization, iteratively selecting a non-ordered set of teaching examples that maximize the risk gradient of a convolutional neural network (ResNet-18 [85]) trained for 10 epochs on the current teaching set. Human participants who learned from 20 images selected using this approach numerically outperformed participants trained via random sampling and other baselines across two image classification tasks, although statistical significance was not reported. Singh et al. (2023, [24]) investigated the impact of task ordering in online class-incremental learning on a custom dataset of novel 3D objects ("Fribbles" [175]), and demonstrated a significant association between the class orders that facilitated effective learning in continual learning algorithms and those that benefited human learners.

The present work is unique in its use of explicit image-computable simulations of learning to augment human performance in complex tasks, unlike the white-box analytical inversion approach of Singla et al., the similarity-based future error minimization of Johns et al., and the offline ANN training approach of Wang et al, which also requires calculating risk gradients using the model. The evolutionary optimization strategy presented here does not require model gradients, and is applicable in principle to any black-box learner. While most previous learner modeling approaches applied to visual category learning in humans are limited to selecting a non-ordered optimal teaching set, or an optimal sequence of entire object classes in the case of Singh et al, this work optimizes both the selection and ordering of individual training examples in a curriculum. Finally, the results demonstrate enhancement of human learning with considerably longer curricula than previous human-facing machine teaching studies involving visual categorization (96 in this work vs. 1-20 in previous work), an important step towards machine teaching applications in domains of practical importance.

5.4 Methods

5.4.1 Overview

This project developed a novel approach to designing optimized curriculum sequences for a challenging visual category learning task, and evaluated an optimized curriculum in an experiment with human participants. To estimate the efficacy of a given curriculum, a surrogate learner model was trained on the curriculum and evaluated on a held-out test set. This estimate served as a fitness metric within a genetic algorithm framework, which was used to jointly optimize a large population of curricula, ultimately producing a single "champion" curriculum to be evaluated in human experiments in comparison to randomly-initialized control curricula. The task used in this study involved distinguishing four categories of lesions in dermoscopy images drawn from the HAM10000 dataset [130]: benign nevus, benign keratosis, basal cell carcinoma, and melanoma.



1. Evaluate Curriculum Fitness using Surrogate Learner Models
A. Train k=50 Surrogate Learner Models on Each Curriculum
Online learning, batch size = 1
Frozen Robust ANN
Learn
$\hat{y}_1 = \text{argmax}(\mathbf{W}_1 \mathbf{x})$
$\hat{y}_2 = \text{argmax}(\mathbf{W}_2 \mathbf{x})$
:
$\hat{y}_k = \text{argmax}(\mathbf{W}_k \mathbf{x})$
y
Sequence of L=96 labeled images is one curriculum
Update all weights W (plasticity rule)
B. Raw Fitness:
Mean accuracy of k surrogates on full class-balanced training set
$\hat{y}_{ij} = \text{argmax}(\mathbf{W}_i \mathbf{x}_j)$
for i ∈ {1,...,k}, j ∈ {1,...,N}
Training set (N images)
C. Fitness Sharing
Maintain diverse population of curricula
Curricula with many common images incur fitness penalty
After g=100 Generations
New Generation
START
Population of P=500 random curricula
FINISH
Champion curriculum (highest raw fitness in final population)
Randomize human participants
Control Group
Optimized Group
4. Elitism
Keep best few curricula from prev. generation (n_e = 5)
3. Curriculum Mating: Crossover and Mutation
*Swap Mutation
Crossover
Training set
+Replacement Mutation
Replace with random training set image of the same class
2. Select Curricula to Reproduce
Population (P=500)
P=500 Random Tournaments
Tournament 1
Tournament 500
Select P=500 to Mate (fittest from each tournament)

Figure 5.1: **Surrogate learner models are used to optimize image curricula to teach humans, in a genetic algorithm framework.** Colors represent different images from the training set, except in box 3 where colors are used to illustrate crossover. A held-out validation set is reserved to check for overfitting and to evaluate human participants who have learned the categorization task from random or optimized curricula.

5.4.2 Surrogate learner models

The surrogate model consisted of (i) a frozen ResNet-50 feature extractor pretrained on ImageNet [80], and (ii) a linear classifier trained on the extracted image features (Figure 5.1, 1.A). This modeling approach was previously shown by Lee and DiCarlo (2023, [167]) to predict human responses while learning a binary visual category learning task with novel categories. The feature extractor used in the present work is a "robustified" ResNet-50 model (adversarially pretrained on ImageNet), which multiple lines of evidence indicate exhibits strong behavioral and representational alignment with human/primate vision [101,156,176–178]. The feature extractor produces 2,048-dimensional embeddings. The embeddings were centered by subtracting the mean of all training set embeddings, and scaled by dividing by the L_2 norm of training set embeddings at the 99th percentile; any embeddings with norms above 1 after this transformation were rescaled to a norm of 1 [167].

A cross-entropy plasticity rule was used to update the weight matrix W of the linear classifier while learning from each image trial. A category prediction $\hat{y}$ for image embedding vector $\vec{x}$ was calculated as follows:

$$\hat{y} = \mathrm{argmax}(W\vec{x}) \tag{5.1}$$

After making a prediction during training, the model was given a "reward" of $r = 1$ for correct responses or $r = -1$ for incorrect. An update to weights $\vec{w}_c$ for the chosen class c (the cth row of W) was then calculated as (following [167]):

$$\vec{w}_c \leftarrow \vec{w}_c + \eta r \sigma(-r\vec{w}_c \cdot \vec{x})\vec{x} \tag{5.2}$$

where η is the surrogate model's learning rate and σ is the sigmoid function.

To evaluate the efficacy of a given curriculum, the surrogate model was trained on one image at a time in the order of the curriculum sequence (online learning with a batch size of 1). The trained surrogate was then tested on a class-balanced subset of the entire training set (during curriculum optimization) and on the held-out validation set (during final post-optimization evaluation). To stabilize the curriculum efficacy estimate, $k = 50$ surrogate models with different random weight matrix initializations were trained on the same curriculum, and their evaluation metrics were averaged. To maximize throughput, all image embeddings were pre-computed using the feature extractor, and many surrogates were trained in parallel using a multi-threaded, GPU-optimized pipeline.

5.4.3 Curriculum optimization with surrogate learner models and an evolutionary algorithm

Curricula were optimized using a genetic algorithm framework, incorporating surrogate model-derived estimates of curriculum efficacy as a measure of evolutionary "fitness." The optimization loop is illustrated in Figure 5.1. A population of $P = 500$ class-balanced curricula was initialized by random sampling from the HAM10000 training set, without replacement per curriculum such that no image appeared more than once in a single curriculum.

The raw fitness of each curriculum in the population was calculated as the mean accuracy (between 0 and 1) among $k = 50$ surrogates on a class-balanced subset of the entire training set (having subsampled each class to the size of the smallest class), which is much larger than the number of images in any one curriculum ($L = 96$). Since only validation set images were used for the test phase of the human experiment, this prevents curricula from being over-optimized specifically for images to be used for evaluation (no “teaching to the test”). To help retain a diverse set of unique images in the overall population throughout optimization, the fitness value of each curriculum was adjusted using a fitness sharing mechanism [179], imposing a penalty on curricula with high Jaccard similarity to other curricula in the population. The Jaccard similarity between two curricula is defined as the number of shared images divided by the total combined number of unique images, irrespective of ordering in the sequences. The shared (penalized for similarity) fitness f'_i of the i-th curriculum is given by:

$$f'_i = \frac{f_i}{\sum_{j=1}^{P} \mathrm{sh}(d_{ij})} \tag{5.3}$$

where f_i is the raw fitness, P is the population size, and d_{ij} is the distance between curriculum i and j (defined here as $1 -$ Jaccard similarity). The sharing function $\mathrm{sh}(d_{ij})$ is defined as:

$$\mathrm{sh}(d_{ij}) = \begin{cases} 1 - \left(\frac{d_{ij}}{\sigma_{\text{share}}}\right)^{\alpha} & \text{if } d_{ij} < \sigma_{\text{share}} \\ 0 & \text{otherwise} \end{cases} \tag{5.4}$$

where σ_{share} is the niche radius (threshold of dissimilarity beyond which no penalty is incurred) and α is a shape parameter. In this work, $\sigma_{\text{share}} = 0.1$ and $\alpha = 1$.

The calculated fitness values were used for a tournament selection procedure [180,181] that sampled curricula to “reproduce” with each other. To select $P = 500$ individuals to reproduce, the following procedure was repeated P times: randomly sample $T = 16$ curricula, and choose the curriculum with the highest fitness among them to reproduce. Curricula selected for reproduction were randomly paired. With a probability of 0.75 per pair, a partially mapped crossover operation [182] (analogous to crossing over during meiosis) was applied to randomly swap segments of curricula between each pair while ensuring uniqueness of all images in each resulting curriculum. Crossover allows recombination of local sequences from high-fitness curricula. With a probability of 0.5, each curriculum was subsequently selected for mutation. During mutation, each image was positionally swapped with another image in the same curriculum with probability 0.05, and each image was replaced with a different image from the training set with probability 0.05. Each replacement image was randomly sampled from the same class as the image being replaced (for class-balanced curricula), or from the smallest minority class within the curriculum (for curricula with imbalances introduced by the crossover operation). Like in biological evolution of DNA sequences, mutation allows stochastic discovery of sequences that increase fitness by introducing random variation in the population. Applying crossover and mutation in a stochastically sporadic manner enables a degree of stability across generations for high-fitness sequences.

The P curricula resulting from this process formed the new population for the next generation of optimization. After fitness evaluation, the $n_e = 5$ individuals with the lowest raw fitness were replaced with the n_e individuals with the highest raw fitness from the previous generation (an “elitism” mechanism [183] that prioritizes preservation of the most efficacious curricula irrespective of fitness sharing penalties). After repeating this optimization for $g = 100$ generations, the curriculum with the highest raw fitness was selected as the "champion" to be used to teach human participants in the subsequent experiment.

5.4.4 Task for human participants

Human participants (n=70) were recruited using the online platform Prolific. Participants learned to distinguish four categories of dermoscopy images (benign nevus, benign keratosis, basal cell carcinoma, and melanoma) drawn from the HAM10000 dataset [130]. The task interface was similar to that used for the dermoscopy task in Chapter 3 ("L-WISE" chapter; see Appendix C, Figure C.4 for an example visualization [156]). While viewing the image in each trial, participants were given 10 seconds to click one of four buttons corresponding to the four categories. After each trial in the initial training phase, participants received immediate feedback indicating the correct category and whether their response was correct. To mitigate the effects of any prior knowledge or preconceptions based on category names, the four categories were assigned aliases based on Greek names ("Ajax," "Eris," "Leda," and "Tyro"), with randomized alias assignments and category response button positions for each participant.

The task was divided into two phases: a training phase with feedback after each trial, and a test phase without feedback. The training phase was a single category-balanced block of 96 trials, and the test phase was a category-balanced block of 48 trials. Participants were randomized to either an optimized curriculum group (for whom the training phase comprised a single optimized "champion" curriculum, from the first optimization run) or a control group (a unique, randomly initialized curriculum for each participant). The testing phase contained an identical sequence of images, none of which appeared in any participant's training phase, for all participants across both groups.

The experiment followed a protocol that was approved by the Institutional Review Board of Boston Children's Hospital. Participant compensation was calibrated to $15.00 USD per hour. To encourage engagement in learning the task, participants with high test phase accuracy were given bonuses ($5.00 above 60%, $10.00 above 70%).

5.4.5 An alternative surrogate model: learning with verbalizable decision rules

To demonstrate the feasibility of optimizing curricula for different types of surrogate learner models, I implemented a neural embeddings-based adaptation of the explicit rule-based system envisioned as part of the Competition between Verbal and Implicit Systems (COVIS) framework [169,184].

The COVIS framework posits that humans learn visual categorization tasks using two distinct systems simultaneously: an implicit, procedural system that integrates information across multiple perceptual features (analogous to the weighted linear combination of embedding dimensions in the linear surrogate), and a separate explicit system based on declarative, verbalizable rules. Consistent with this account, converging evidence from event-related potentials, patients with neurodegenerative, psychiatric, and learning disorders, and comparisons between humans and non-human animals suggests that these systems are mediated by partially distinct neural substrates. Specifically, explicit rule-based learning is linked to prefrontal cortex and hippocampus, while implicit procedural learning depends primarily on striatal structures [169,185–187]. The explicit and implicit systems are theorized to "compete" to determine the behavioral response output based on which system has higher confidence on a given example [169].

The explicit system is modeled here as a rationally designed hypothesis testing algorithm, where each hypothesis is an explicit rule (threshold) along a single, verbalizable dimension that aims to differentiate between classes. The system maintains a small buffer of recently-seen labeled examples ("working memory"), and uses these examples to probabilistically activate a single verbal rule that maximizes class separation. The classification threshold of the rule is adjusted with each new example, as is an estimate of the rule's trustworthiness. When trustworthiness drops below a certain threshold, the active rule is discarded and a new active rule selected based on the examples in working memory. The approach is detailed below, and is also described as pseudocode in Appendix E (Algorithm 2). This implementation is applicable only to binary tasks; it was evaluated using the MHIST dataset [131], which contains colon histology image patches sourced from either a hyperplastic polyp or a sessile serrated adenoma.

A library of 50 candidate verbal rules was generated using a large vision-language model, which was presented with 20 labeled, randomly-sampled images from each of two categories, and prompted to produce class-discriminative verbal statements in plain English without domain-specific terminology or jargon. The complete prompt and other details are available in Appendix E. Example generated sentences for the MHIST dataset were "the geometry is dominated by circles and smooth curves" and "the pattern is repetitive and grid-like, rather than chaotic or branching." The degree to which each verbal rule k aligns with each image x in the dataset is pre-calculated as the similarity score $s_k(x)$, defined as the cosine similarity between the CLIP embedding [188] of k from the language encoder and that of x from the vision encoder.

State Variables

The explicit system maintains the following state variables:

- **Global Salience (w):** A vector $w \in \mathbb{R}^K$, representing the prior probability of each rule k based on estimated perceptual saliency. Initialized as the mean cosine similarity between the rule embedding and a balanced subset of 500 randomly-sampled training images ($\mathcal{X}_{init}$), normalized such that $\sum_k w_k = 1$.

- **Active Rule (k_t):** The verbal rule currently guiding the explicit system's predictions at trial t.

- **Decision Bias (b):** A scalar threshold determining the boundary between categories for the active rule.

- **Rule Trust (τ):** A scalar $\tau \in [0, 1]$ representing confidence in the active rule.

- **Working Memory ($\mathcal{B}$):** A class-stratified first-in-first-out buffer retaining the $N = 2$ most recent exemplars of each class (4 in total), consistent with estimates of visual working memory capacity limits in humans [189,190].

Rational Rule Selection

When trust in the active rule falls below a threshold τ_{min}=0.3, a switch to a different rule is triggered. To choose a new rule to activate, the model employs a rational strategy that maximizes class separation over the contents of working memory.

A local discriminability score (margin) M_k is calculated for each rule k based on the images currently in working memory plus the current image x_t (in a behavioral experiment, x_t would be displayed on the screen and thus accessible to the learner). These 5 examples form the exemplar pool $\mathcal{P}$. The margin $M_k(\mathcal{P})$ for rule k is calculated as follows with respect to classes A and B (in this setting, hyperplastic polyps vs. sessile serrated adenomas):

$$D_k(\mathcal{P}) = \frac{1}{|\mathcal{P}_A|} \sum_{x \in \mathcal{P}_A} s_k(x) - \frac{1}{|\mathcal{P}_B|} \sum_{x \in \mathcal{P}_B} s_k(x) \tag{5.5}$$

$$M_k(\mathcal{P}) = |D_k(\mathcal{P})| \tag{5.6}$$

Each rule is also associated with a polarity parameter $d_k = \text{sign}(D_k(\mathcal{P}))$, which indicates whether the rule is positively or negatively associated with the (arbitrarily) first class in a binary classification task. The set of top candidate rules $\mathcal{K}_{top}$ is defined as the indices of the $r = 5$ rules with the highest combined scores $(w_k \cdot M_k(\mathcal{P}))$. The probability of selecting rule k is given by a truncated softmax function with temperature T:

$$P(k) = \begin{cases} \dfrac{\exp((w_k \cdot M_k)/T)}{\sum_{j \in \mathcal{K}_{top}} \exp((w_j \cdot M_j)/T)} & \text{if } k \in \mathcal{K}_{top} \\ 0 & \text{otherwise} \end{cases} \tag{5.7}$$

Prediction and Confidence

The explicit system calculates its prediction $\hat{y}_{exp} \in \{-1, 1\}$ by comparing the similarity $s_{k_t}(x_t)$ (between current image x_t and active rule k_t) to the threshold b_{k_t} of the active rule. The explicit system's prediction is calculated as:

$$\hat{y}_{exp} = d_{k_t} \cdot \text{sign}(s_{k_t}(x_t) - b_{k_t}) \tag{5.8}$$

Confidence in $\hat{y}_{exp}$ is modeled as a sigmoidal function of the distance from the decision threshold, with slope parameter γ:

$$C_{exp} = \frac{1}{1 + \exp(-\gamma |s_{k_t}(x_t) - b_{k_t}|)} \tag{5.9}$$

Learning and Update Dynamics

Given ground truth label $y_t \in \{-1, 1\}$, the model updates its state:

1. **Trust Update:** If correct, trust τ increases linearly; if incorrect, trust decays multiplicatively:
$$\tau_{t+1} = \begin{cases} \min(1.0, \tau_t + \alpha_{pos}) & \text{if } \hat{y}_{exp} = y_t \\ \alpha_{neg}\tau_t & \text{if } \hat{y}_{exp} \neq y_t \end{cases} \tag{5.10}$$

2. **Bias Adaptation:** The decision boundary b is updated via a weighted moving average toward the optimal separation point b^* calculated over the current exemplar pool, $\mathcal{P} = \mathcal{B} \cup \{x_t\}$:
$$b^*(\mathcal{P}) = \frac{1}{2}\left(\frac{1}{|\mathcal{P}_A|}\sum_{x\in\mathcal{P}_A} s_k(x) + \frac{1}{|\mathcal{P}_B|}\sum_{x\in\mathcal{P}_B} s_k(x)\right) \tag{5.11}$$
$$b_{t+1} = \lambda b_t + (1-\lambda)b^*(\mathcal{P}) \tag{5.12}$$
where $\lambda = 0.8$ determines the persistence of the previous boundary estimation. When a new rule is first selected, the boundary is initialized to b^*.

3. **Global Weight Update (w):** The global salience of the active rule k_t is reinforced to reflect long-term utility:
$$w_{k_t} \leftarrow \max(w_{k_t} + \eta(y_t \cdot \hat{y}_{exp}), 0) \tag{5.13}$$
where η is a small learning rate. The weights w are re-normalized to sum to 1 after the update.

Pseudocode describing the verbal rule surrogate learner is available in Appendix E (Algorithm 2). Notably, future work could easily combine the verbal rule surrogate with the linear surrogate described earlier, which is closely analogous to the implicit information-integration system in the COVIS framework in that it combines information across multiple features (unlike the unidimensional decision boundary of the verbal rule surrogate). To construct a model that implements the full COVIS framework, the explicit prediction $\hat{y}_{exp}$ would compete with the implicit system's prediction $\hat{y}_{imp}$ (with confidence C_{imp}) to determine the model's overall response:

$$\hat{y}_{overall} = \begin{cases} \hat{y}_{exp} & \text{if } C_{exp} > C_{imp} \\ \hat{y}_{imp} & \text{otherwise} \end{cases} \tag{5.14}$$

5.5 Results

This project used a simple linear "surrogate learner model," which is hypothesized to emulate aspects of human visual category learning behavior, to optimize curriculum sequences for teaching humans. Optimized curricula for a four-class dermoscopy classification task were then compared with random control curricula in a randomized experiment with lay human participants. The curriculum optimization approach was also tested in an additional surrogate designed to simulate verbalizable rule-based learning in humans.

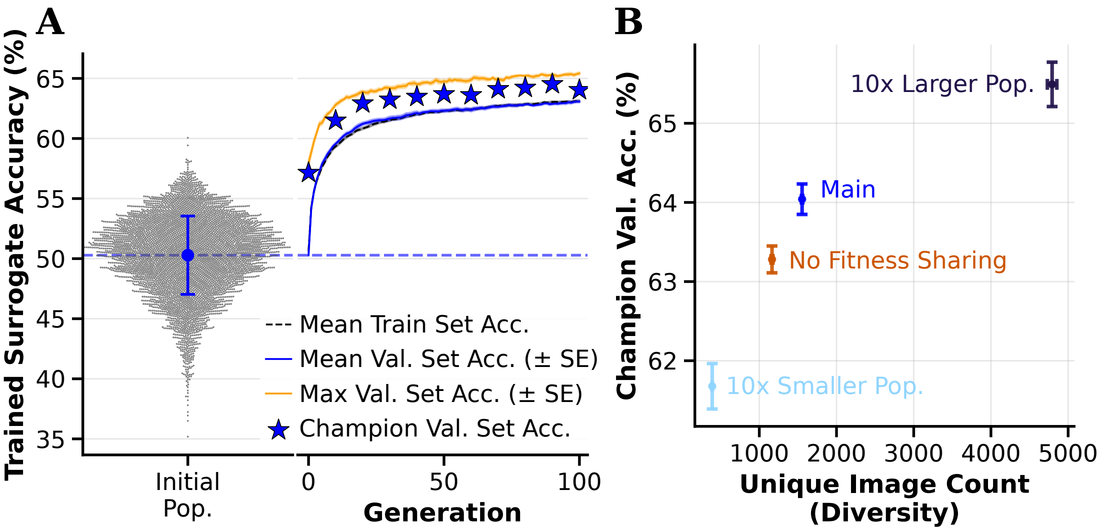


Figure 5.2: **A genetic optimization approach yields highly effective curricula for training a linear surrogate model.** Panel **A** illustrates the results of curriculum optimization for a dermoscopy task with four classes. The plot is based on 10 independent runs with starting populations of $P = 500$. The "Initial Pop." swarm plot shows the *validation set* accuracy of surrogate models trained on these 5000 randomly initialized curricula: each dot represents the average of 50 randomly-initialized surrogates trained on one curriculum (same initializations re-used). The blue dot plot and dotted line show the mean validation set accuracy, with the error bar showing the standard deviation among random curricula. Line plots show the progression of mean and maximum validation set accuracy among the population as genetic optimization proceeds through 100 generations. Lines show the mean among 10 runs, and (thin) shaded areas show standard error of the mean among runs. To avoid optimizing the curriculum design for the validation set, the "champion" is defined as the curriculum yielding trained surrogates with the highest overall training set accuracy. The held-out validation set accuracy of the champions (averaged across 10 runs) is plotted as a star for every 10 generations. Panel **B** illustrates the relationship between population diversity and the validation set accuracy of the final champion. Diversity is measured as the total number of unique images present across the entire population of curricula, averaged across all evolutionary time poins (generations). Each dot plot shows the mean and standard error across 10 runs (5 runs for "10x Larger Pop." due to resource constraints). "Main" is the run depicted in panel A and used for subsequent experiments. Diversity is increased at larger population sizes and when the fitness sharing mechanism is active. All pairs among the 4 conditions are significantly different from each other in both unique image count and champion validation accuracy (Kruskall-Wallis omnibus test with post-hoc pairwise Mann-Whitney U tests, $p < 0.05$ following Bonferroni correction for multiple comparisons).

5.5.1 Surrogate-based curriculum optimization with a genetic algorithm yields large improvements in trained surrogate accuracy

The efficacy of each curriculum at teaching humans was predicted by training a set of surrogate models on it, one image at a time (online machine learning), and then evaluating the mean accuracy of the surrogates on a large, class-balanced set of images. To calculate fitness during genetic optimization, surrogates were evaluated on a maximal class-balanced subset of the training set; a maximal class-balanced subset of the validation set was then used to assess generalizable curriculum efficacy (i.e., for images outside the training set).

Surrogates trained on randomly-initialized, class-balanced curricula reached an average validation set accuracy of 50.3% on the 4-class dermoscopy task, with a standard deviation (SD) of 3.3% (see swarm plot in Figure 5.2A). Genetic optimization across 100 generations, with a population of 500 curricula, yielded champion curricula with a mean validation set accuracy of 64.0% across 10 independent optimization runs (4.2 SDs above the mean for random curricula). Surrogate training set and validation set accuracy were virtually identical throughout the optimization (see blue and black-dotted lines in Figure 5.2A), suggesting that curricula were not optimized in such a way that was overfitted to the training set. However, the champion curriculum, which was selected based on training set accuracy to avoid overfitting, consistently yielded lower validation set accuracy than the maximum among all curricula in the optimized population.

To evaluate the impact of key hyperparameters of the genetic algorithm, the optimization was repeated with the fitness sharing mechanism disabled, and with ten times smaller ($P = 50$) and ten times larger ($P = 5000$) population sizes (Figure 5.2B). The results were averaged over 10 runs for all configurations, except that only 5 runs were used for $P = 5000$ due to high computational expense. The results indicate that larger population sizes accommodate a larger total number of unique images across curricula (a simple measure of curriculum diversity), ultimately yielding increased champion curriculum efficacy. Accordingly, activating the fitness sharing mechanism had a positive impact on both diversity and champion efficacy. All pairs among these four conditions had statistically significant differences ($p < 0.05$) in both unique image count and champion validation set accuracy, following Bonferroni correction for multiple comparisons (see Figure 5.2 caption).

5.5.2 A surrogate-optimized curriculum enhances human accuracy and response time in a test following visual category learning, while also decreasing training time

Seventy lay human participants were recruited using the online platform Prolific, and randomized to either a Control group (randomly-initialized curricula) or an Optimized group (champion curriculum from a single optimization run, identical across participants). All participants learned a 4-class dermoscopy classification task from a sequence of 96 training trials with feedback (which differed between experimental groups), followed by a 48-trial test with no feedback (identical for all participants). Compared to the Control group, participants in the Optimized group had higher test accuracy (mean 43.9% vs. 36.8%, $p = 0.01$, Mann-Whitney U test) and faster test-phase response times (mean 1.85 seconds vs. 2.17 seconds, $p = 0.02$), and also completed the training phase of the experiment more quickly (mean 6.6 minutes vs. 8.2 minutes, $p = 0.005$; Figure 5.3). The linear surrogate models significantly outperformed human participants across both groups (see horizontal blue and black dotted lines in Figure 5.3B). For example, surrogates trained on randomly initialized curricula reached 50.3% validation accuracy on average, compared with 43.9% for humans in the Optimized group ($p = 0.001$, Mann-Whitney U test).

5.5.3 Both image selection and relative ordering are optimized by the surrogate-based genetic algorithm, with natural emergence of easy-to-hard trends

To assess which qualities of optimized curricula are important in determining surrogate performance, a series of experiments were conducted involving systematic ablation of champion curricula followed by training of newly-initialized surrogates on them (Figure 5.4). Each

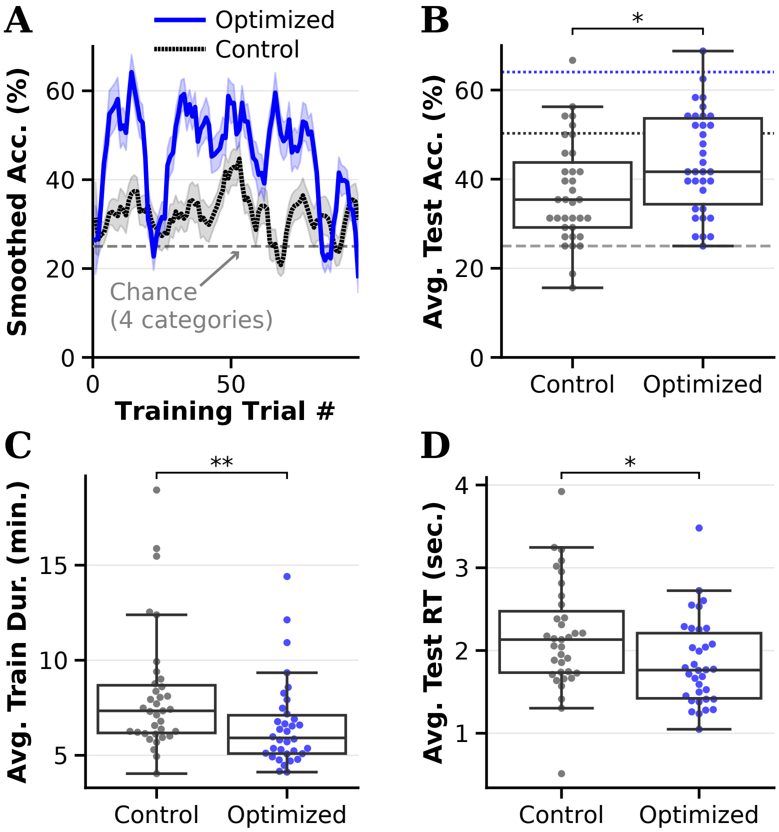


Figure 5.3: **A curriculum optimized by surrogates allows humans to learn a visual task more efficiently, to higher accuracy, and to shorter reaction times.** Panel **A** shows average human learner accuracy curves during the 96-trial *training phase* (smoothed with Savitzky-Golay filter). Shaded sleeve regions represent standard error of the mean. Chance is at 25% with four evenly-balanced categories (benign nevus, benign keratosis, basal cell carcinoma, and melanoma). Panel **B** shows *test phase* accuracy of participants who learned from random control curricula or surrogate-optimized curricula. Each dot represents one participant. The blue dotted line (at 64.0%) and black dotted line (50.3%) indicate mean surrogate model accuracy on optimized champion curricula and random curricula, respectively. Panel **C** shows effective training duration of participants in the Control and Optimized Curriculum groups. This was defined as the total time for which stimuli were displayed on the screen, clipping times on the feedback screen after each trial to a maximum of 60 seconds to remove any long breaks. Panel **D** shows mean test phase response time of participants in the two groups, counting only trials with correct responses. All box plots show the median and inter-quartile range (IQR), with whiskers extending to the farthest data point within 1.5×IQR of the box edges. * $P < 0.05$, ** $P < 0.01$.

of these experiments employed champion curricula from 50 independent optimization runs, which had the same hyperparameters as the run used to produce the curriculum for the human experiments. Judgments of statistical significance in this section are descriptive/exploratory, and based on visual inspection of confidence interval overlap.

Replacing images in champion curricula with random class-matched images from the training set decreased trained surrogate accuracy in a roughly linear fashion (orange line plot with triangles in Figure 5.4A). Shuffling the positions of an increasing, randomly-sampled subset of the images in champion curricula also yielded an approximately linear accuracy decrease, but with a significantly shallower slope than image replacement (green line with circles in Figure 5.4A). Completely reversing the order of curricula yielded a significantly larger accuracy decrease than complete shuffling (“Reversed Order” line in Figure 5.4A), but

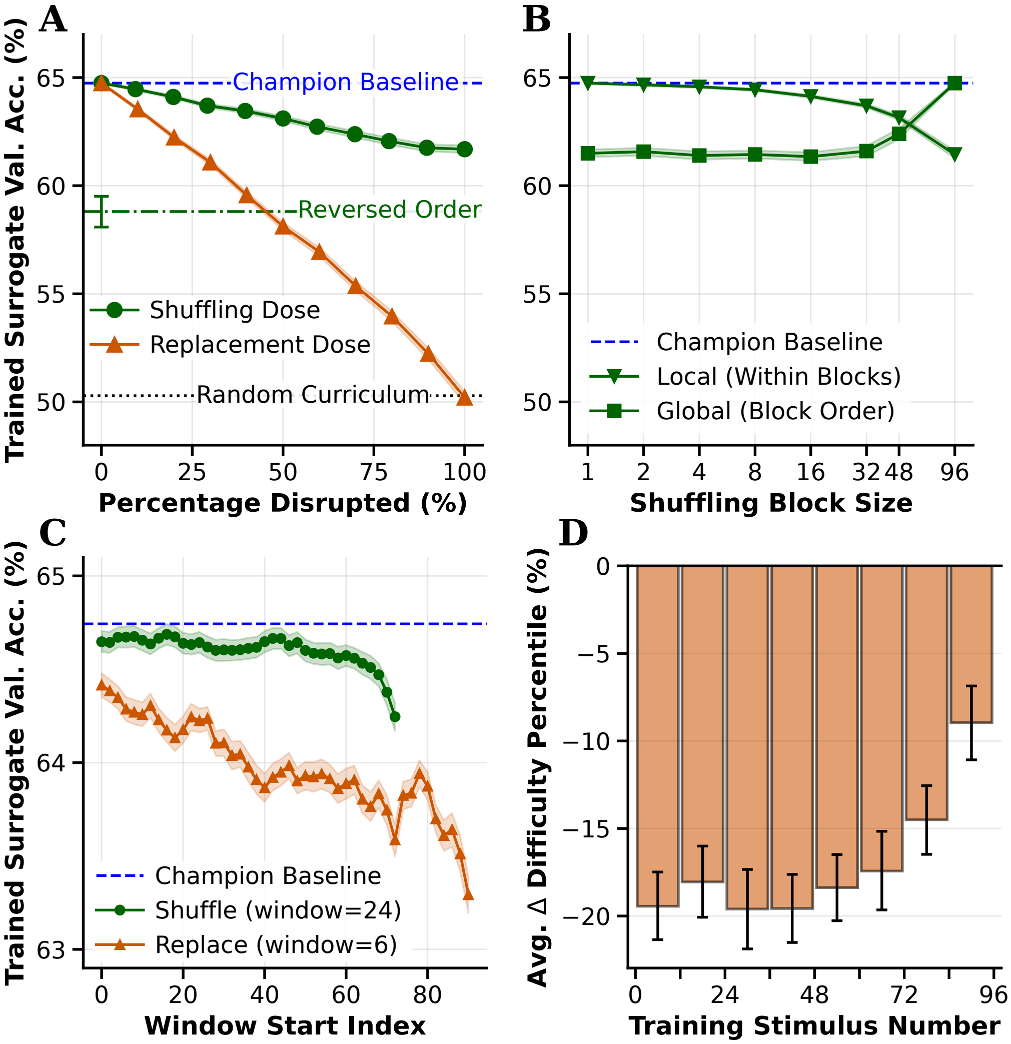


Figure 5.4: **Surrogate-based curriculum optimization naturally selects relatively easy images and sequences them by increasing difficulty.** Panel **A** shows the effect of ablations to optimized "champion" curricula on the validation set accuracy of surrogate models trained on it. The orange line (triangles) shows the effect of replacing a percentage of the images in the curriculum with images (of the same respective classes) randomly drawn from the HAM10000 dataset's training set. The green line shows the effect of shuffling the order among a randomly-selected percentage of the images. "Reversed Order" shows performance of the champion curriculum with the image sequence reversed. Panel **B** shows the effect of shuffling the order either within ("Within Blocks") or among ("Block Order") evenly-sized blocks within 96-item champion curricula. Panel **C** shows the effect of either shuffling or replacing all champion curriculum items within a sliding window (window width=24 items for shuffling, 6 items for replacing). Panel **D** shows the difference between champion curricula and random curricula in mean percentile difficulty (estimated via the ground truth logit of a robust ResNet-50 model) within fixed blocks of 12 curriculum items. All plots use the mean of 50 independent curriculum optimization runs. All sleeves/error bars represent 95% bootstrap confidence intervals.

still less than half of the accuracy decrease from complete replacement. Together, these results suggest that most of the benefit of curriculum optimization is mediated by selection of appropriate images, while relative ordering plays a lesser but still significant role in determining curriculum efficacy.

To better characterize the influence of relative ordering by separating the effects of overall global image position in the curriculum from those of specific sequences within local regions/segments, an additional ablation experiment was performed in which images were shuffled within or among blocks of varying sizes (Figure 5.4B). For example, for a block size of 4, the champion curricula were partitioned into segments of 4 images each. Local shuffling consisted of shuffling the positions of the (e.g.) 4 images within each block, while global shuffling involved shuffling the relative order of the entire blocks. The results indicate that global shuffling had a significantly larger detrimental impact on surrogate performance than local shuffling, suggesting that the overall global position of each image in an optimized curriculum is more important than its local position relative to nearby images.

To evaluate whether some regions of the optimized curriculum sequences are more important than others in determining surrogate performance, additional replacement and shuffling ablations were performed within sliding windows to focus the ablations on specific sections of the curriculum (Figure 5.4C). A window size of 6 was used for replacement and 24 for shuffling due to the relative effect sizes of these ablations. The results show that, while ablating any region of the curriculum through either shuffling or replacement can detrimentally affect surrogate performance, the effect of ablation is larger for later curriculum regions.

Previous work has shown that selecting relatively easy images for a visual category learning curriculum can enhance subsequent test performance in human learners [156] (see Chapters 3 and 4). Figure 5.4D divides curricula into contiguous blocks of 12 images each, and plots the mean difference in model-estimated difficulty percentile of images in each block between champion curricula and random curricula. This plot shows that champion curricula contain significantly easier images on average across all curriculum regions. Additionally, an easy-to-hard trend is observable within the final 2 blocks of the curriculum (24 images total) relative to earlier blocks. Notably, this effect emerged naturally with no difficulty-related guidance of the optimization process.

5.5.4 The relative impact of curriculum optimization depends on curriculum length and surrogate learning rate

The main curriculum optimization experiment was repeated for a wide range of curriculum lengths and learning rates. Figure 5.5A shows that, as curriculum length increases, surrogate accuracy increases while the effect size of curriculum optimization generally decreases. The effect size is measured as the difference between mean surrogate validation set accuracy after training on randomly initialized curricula vs. the champion curriculum, in units corresponding to the standard deviation among random curricula. An exception to this trend is for a minimal curriculum with only 4 images in total, which shows a lower optimization effect size than a curriculum with 12 images. A similar trend is observed as a function of surrogate learning rate in Figure 5.5B: surrogates perform better at higher learning rates (up to an apparent

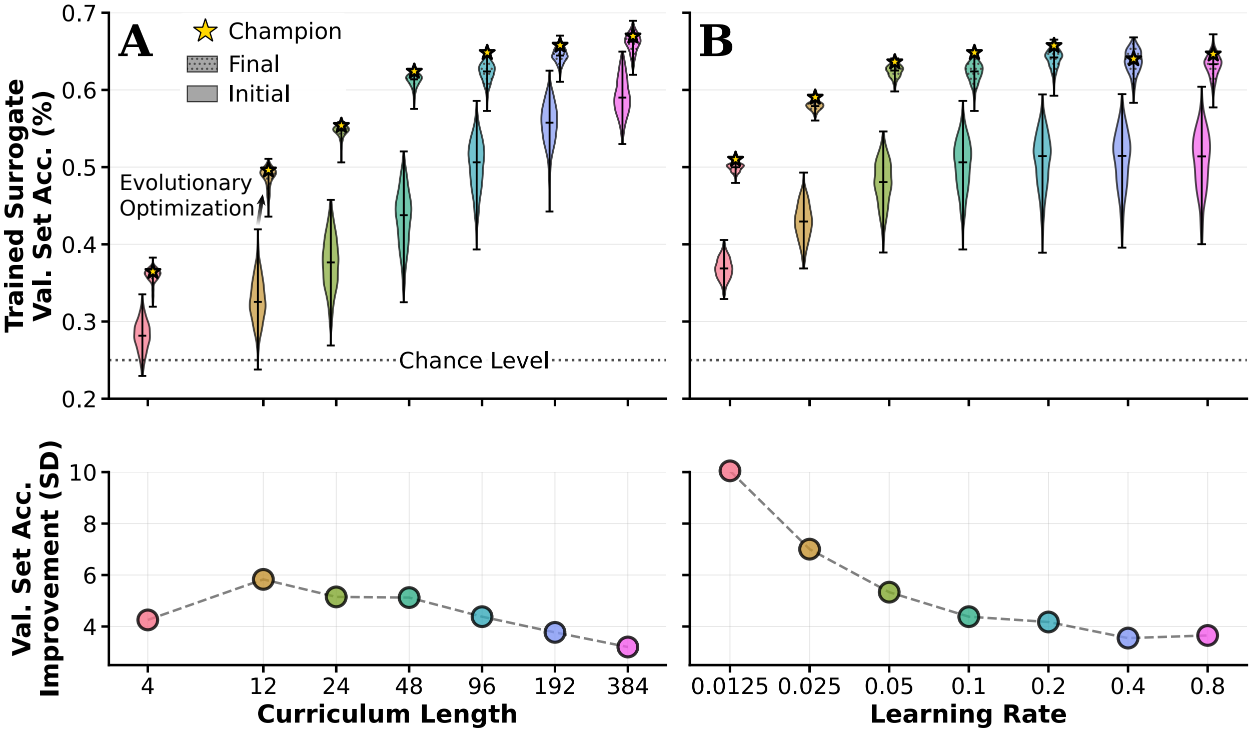


Figure 5.5: **Curriculum optimization has a larger effect size for shorter curricula and surrogates with lower learning rates.** Panel **A** compares validation set accuracy of surrogates trained on initial (random) and final (genetically optimized) curricula of different lengths (number of images in sequence). Each violin plot shows median, minimum, and maximum surrogate validation set accuracy within a single optimization run's population of 500 curricula each, with stars marking accuracy of the champion (defined as the curriculum yielding a surrogate with the highest final training set accuracy). The violin plot for the final population is plotted slightly to the right of that for the initial population for visibility. The lower plot shows the improvement, measured in initial population standard deviation units, from the mean validation set accuracy of the initial population to that of the champion curriculum. Panel **B** compares improvement from randomly initialized to champion curricula among optimization runs with different surrogate model learning rates. This refers to the size of the learning rule update step that the surrogate model takes when learning from an individual image in the curriculum.

saturation point), but the effect of curriculum optimization is largest at low learning rates. However, this trend appears to be driven in part by low variance among curricula at low learning rates, rather than dramatically larger absolute increases in accuracy.

A related question to the overall impact of surrogate learning rate is whether curricula optimized for surrogates at one learning rate are transferable to surrogates with a higher or lower learning rate. Figure 5.6 shows that optimized curricula can partially generalize to learning rates up to 256 times larger or smaller than their "native" learning rate used for optimization, in that the validation set accuracy of trained surrogates remains significantly higher than those trained on random curricula. However, surrogates at non-native learning rates generally underperformed surrogates trained at each native learning rate, and this disparity increased with larger positive or negative shifts between the optimization learning rate and the evaluation learning rate.

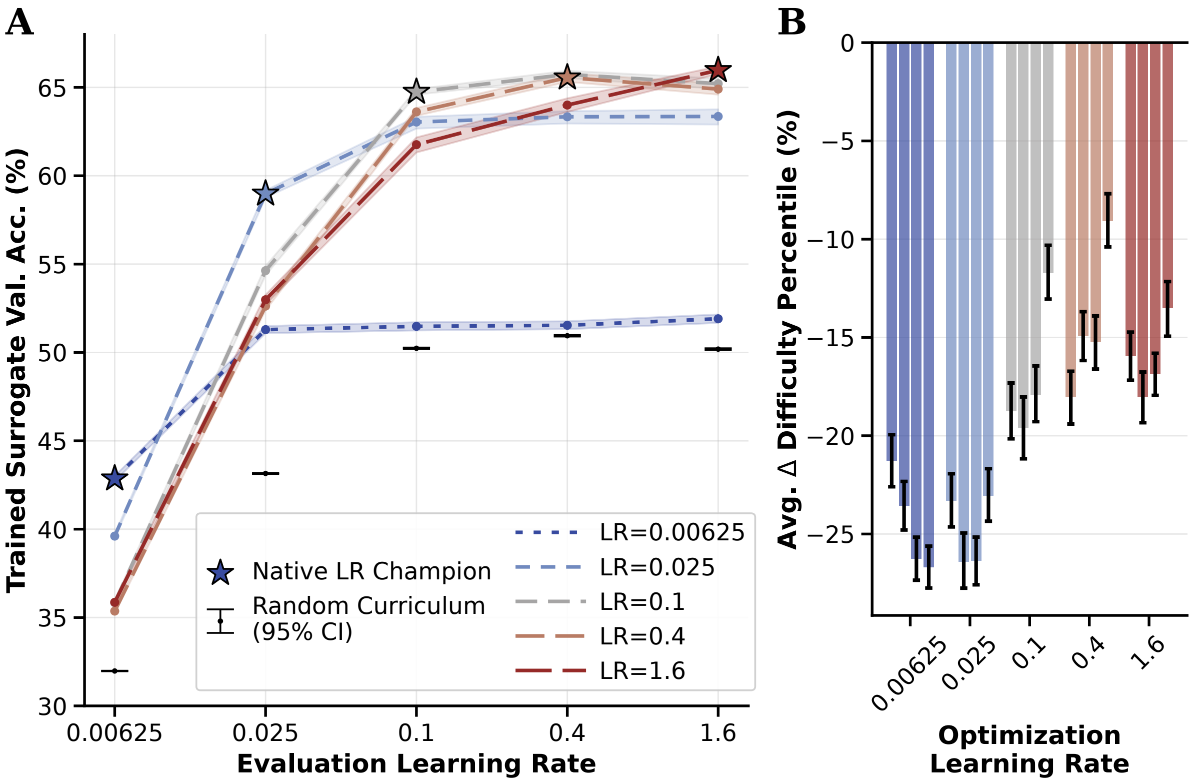


Figure 5.6: **Curriculum optimality depends on surrogate model learning rate, with curricula optimized for low learning rates containing easier images on average.** Each line plot in panel **A** represents champion curricula optimized using a different surrogate learning rate (50 independent runs per learning rate): validation set accuracy varies as a function of the learning rate of surrogates subsequently trained on these champion curricula (the "Evaluation Learning Rate"). Stars indicate the performance of surrogates trained on champion curricula at the native learning rate (the same learning rate at which the curricula were optimized). Shaded regions indicate 95% bootstrap confidence intervals (CIs). Black error bars show 95% bootstrap CIs for surrogate performance after training with each learning rate on randomly initialized curricula. The horizontal axis is a $\log_2$ scale, with increments such that each subsequent learning rate is 4 times larger than the previous one. Panel **B** shows the mean difference in model-estimated image difficulty percentile (normalized by class) within champion curricula vs. randomly initialized curricula. Each group of bars represents curricula optimized for a different learning rate (50 runs each). Each curriculum is a sequence of 96 images: within each learning rate group, bars represent images 1-24, 25-48, 49-72, and 73-96 from left to right. Error bars represent 95% bootstrap confidence intervals.

5.5.5 Curriculum optimization depends strongly on surrogate type

Optimization of curricula for a binary histology discrimination task (benign hyperplastic polyp vs. sessile serrated adenoma in the MHIST dataset [131]) was conducted for both the linear surrogate learner and the verbal rule-based surrogate introduced in Section 5.4.5. The population size was set to $P = 100$ and the curriculum length to $L = 48$; all other hyperparameters were the same as for the dermoscopy task. All curricula were class-balanced, as were the training and validation subsets used for fitness calculation and final evaluation.

As was the case for dermoscopy, evolutionary curriculum optimization was effective for the linear surrogate in that the validation set accuracy of surrogates trained on the final champion curriculum significantly exceeded the entire distribution among random curricula at baseline (Figure 5.7A, top). However, curricula near the performance ceiling

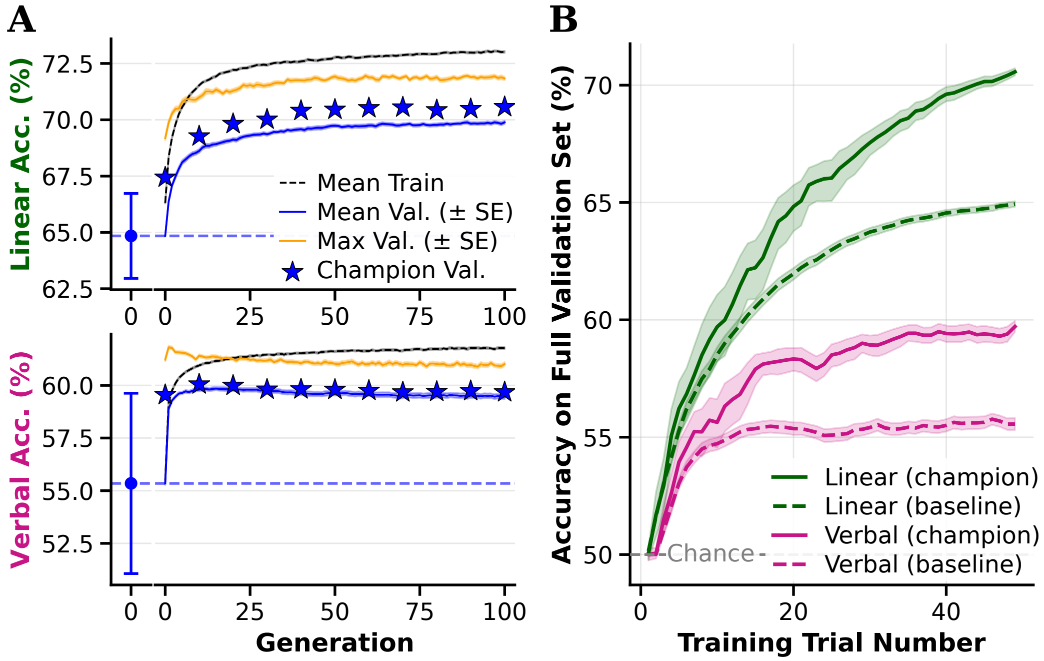


Figure 5.7: **Curriculum optimization is effective for multiple surrogate types, with a linear "information-integration" learner outperforming a verbalizable rule-based learner.** Panel **A** shows the class-balanced accuracy of fully trained surrogates on the MHIST binary histology task as a function of generations of evolutionary optimization (similar to Figure 5.2A). The top plot shows accuracy data for trained linear surrogates and the bottom plot shows verbal rule surrogates. The blue dot plot on the left (separated for visibility) shows the mean (shown also by the blue dotted line) and standard deviation of validation set accuracy for surrogates trained on 10,000 randomly-initialized curricula, averaged among 50 surrogates per curriculum. Line plots show the progression of mean and maximum validation set accuracy of trained surrogates, and mean train set accuracy, averaged among 50 independent optimization runs with 50 surrogates per curriculum. Shaded regions indicate the standard error among optimization runs. Stars indicate the validation set accuracy of champion curricula, which are selected based on training set accuracy every 10 generations. Panel **B** shows the accuracy of surrogates on the entire class-balanced validation as they progress through training iterations within their curricula. Dotted lines show mean accuracy across 1000 randomly-initialized curricula, and solid lines show mean accuracy across 50 independently-optimized champion curricula (50 surrogates per curriculum in both cases). Shaded regions represent 95% bootstrap confidence intervals. The chance level is 50% (grey dotted line) for the binary histology task.

for the verbal rule surrogate were found by mere surrogate-based selection within the initial set of randomly-initialized curricula, with little to no benefit from additional evolutionary optimization (Figure 5.7A, bottom). Comparing against a similar plot for the dermoscopy task in Figure 5.2A, Figure 5.7A also shows strong clear evidence of "curriculum overfitting" for both surrogate types, in that mean training set accuracy greatly exceeds mean (and ultimately maximum) validation set accuracy as evolution progresses. This may be related to the smaller size of the MHIST training set (2175 images in total, compared with 6592 for the 4-class subset of the HAM10000 dermoscopy dataset [130]).

Despite overfitting and the sharply diminishing returns while optimizing for the verbal surrogate, champion curricula of both surrogate types significantly improved validation set performance relative to random curricula. Improvement in validation set accuracy as a function of trials completed, for 1000 randomly-intialized and 50 independently-optimized champion curricula of both surrogate types, is plotted in Figure 5.7B. Champion curricula, which are selected based entirely using the training set, resulted in significantly higher validation set accuracy than random curricula for both surrogate types. The verbal rule surrogate learned to classify significantly above the chance level of 50% accuracy, but significantly underperformed the linear surrogate.

Taken together, these results show that the performance of both surrogate types is curriculum-dependent. However, an additional exploratory analysis suggested that curricula effective for one surrogate type may not be effective for the other. For a fixed set of 1000 randomly-initialized curricula, the validation set accuracies of trained verbal rule surrogates and trained linear surrogates (learning rate=0.1) were not correlated (Spearman's $\rho = 0.02$, $p = 0.52$). In contrast, the Spearman correlation between linear surrogates at learning rates of 0.1 vs. 1.6 was 0.92 ($p < 0.001$), and that between learning rates of 0.025 and 1.6 was 0.43 ($p < 0.001$). However, verbal rule surrogates did show a weak positive correlation with linear surrogates at a smaller learning rate of 0.00625 (Spearman's $\rho = 0.12, p < 0.001$), the latter of which was uncorrelated with linear surrogates at higher learning rates.

5.6 Discussion

This work demonstrates that image-computable models can be used to simulate human visual category learning end-to-end, in a manner that accounts for the effects of curriculum design on learner performance. Human training efficiency and post-training accuracy and response time in a dermoscopy task were all enhanced by a curriculum that was evolutionarily optimized against the post-training performance of "surrogate learner" models, which took the form of linear classifiers trained on neural network embeddings. Curriculum ablation experiments showed that both example selection and relative ordering were important for surrogate performance, and that optimized curricula naturally select easy images with a difficulty progression that depends on the surrogate's learning rate during optimization.

A variety of ablation-style experiments provide insight into the relationship between curriculum and the learning dynamics of surrogate models. Observations from Figures 5.5 and 5.6 show that longer curricula and higher learning rates both increase the final accuracy of trained surrogate models. Curriculum optimization appears to have its largest effect in settings with short curricula or low learning rates. This seems to suggest that the surrogate models were bottlenecked during training by the need to travel a certain distance in parameter space from initialization to regions producing high-performing classifiers, a distance that can be more fully traversed with a larger number of steps (longer curricula), or with a larger step size (higher learning rate).

This framework could also help explain why curricula that were optimized at one learning rate were no longer optimal for surrogates with much higher or lower learning rates (Figure 5.6A). Curricula optimized at low learning rates included easier images on average than curricula optimized at high learning rates, and high-rate curricula (arbitrarily, 0.1 and above) featured a trend towards more difficult images towards the end of the curriculum while low-rate curricula did not (Figure 5.6B). I speculate that we can conceptualize pairs of classes as being distinguished from each other along multiple principal components. For example, in the dermoscopy task, melanomas are distinguished from benign nevi by both asymmetry and the presence of multiple colors (among other features [136]). Given a limited number of steps and/or a low learning rate, it may be optimal to prioritize parameter updates along the first principal component, which is most consistently represented by easy examples positioned far away from class boundaries. At a higher learning rate, the classification hyperplanes parameterized by the model can become aligned with respect to the first principal component in fewer steps. This allows later steps to prioritize other, secondary components that correspond with more subtle distinctions between classes, distinctions best emphasized by more challenging examples that may be less straightforwardly classifiable using the first principal component.

In more mechanistic terms, for each image in a curriculum sequence, the surrogate's parameter vector associated with the ground truth class is incremented in a direction defined by the image's embedding vector. For the trained surrogate classifier to be accurate, each class' parameter vector must produce a relatively high dot product when multiplied by image embeddings of the same class. Images that strongly represent the most important features of a given class (which will tend to be relatively "easy") result in updates that more strongly affect the parameters that multiply with those same features during prediction, with smaller effects on the parameters for less important features.

While surrogate model hyperparameters such as the learning rate can influence performance and curriculum optimization results, comparisons between linear and explicit verbal surrogates revealed striking differences in learning dynamics. The verbal rule-based surrogate tended to reach an accuracy plateau on the binary MHIST histology task [131] early in its course of training, at a significantly lower performance ceiling than the linear surrogate, which continued to improve throughout training (Figure 5.7B). This is partially consistent with the predictions of the COVIS framework [184], which posits that while some tasks can be learned optimally through explicit rule-based strategies (e.g., recognizing red vs. green traffic lights), others require integrating among multiple features through computations closely analogous to the linear surrogate in this study: examples include radiological diagnosis and detecting signs of malignancy in a histology image [169,191,192]. For tasks that are optimally solved with information integration strategies, explicit verbalizable rules are learned rapidly but have a low accuracy ceiling, while the implicit procedural system learns more slowly but with a higher, later plateau [169,193]. Inspection of Figure 5.7B suggests that the linear surrogate is more sample-efficient than the verbal rule surrogate even in the early stages of training, which conflicts with the COVIS prediction that procedural learning should initially be slower than declarative rule acquisition. Future refinements to the verbal rule surrogate could potentially improve its sample efficiency to more closely resemble explicit rule learning in humans.

It is also likely that the relative suitability of different surrogate models depends on the nature of the task: matching the type of surrogate to the strategy human learners are most likely to use for a given task could both better account for human learning dynamics from a modeling standpoint and lead to optimized curricula that better transfer to teaching humans. One intriguing possible approach is that of composite surrogates, such as an image-computable implementation of COVIS in which models similar to the linear and verbal rule surrogates in this work compete to determine an overall prediction based on their relative confidence [169]. Additionally, the fact that curriculum dependency is not correlated between surrogate types (see Section 5.5.5) raises the possibility of dual-objective optimizations that maximize the accuracy of one surrogate while minimizing that of another, potentially guiding human learners towards strategies that are optimal for a given task while discouraging other, less-optimal strategies. Hypothetically, for example, inhibiting explicit rule-based learning could ultimately improve performance on tasks that require information integration [194,195], which includes various forms of image-based medical diagnosis.

Different training settings and learner populations may also benefit from different surrogate modeling approaches. For example, surrogate models for long-term training (e.g., in medical residency programs), perhaps optimally spaced across multiple sessions [196], should ideally account for fatigue within long sessions and forgetting dynamics in between sessions. A feature extractor network pre-trained on ImageNet may be a reasonable model of visual perception in a lay person, but the visual system of a developing expert (e.g., a 2nd-year pathology resident learning advanced histology) would likely be better modeled by a feature extractor with domain-specific pretraining. Future models could also account for inter-individual variation within a given human population, such as differences in prior knowledge, learning ability, and preferred learning strategies as a function of task type [197]: I speculate that the surrogate model that produces the most accurate curriculum efficacy estimates for humans on average may not be an optimal surrogate for any specific individual. Surrogate models could potentially be personalized by adapting to individual learners on the fly, using human response patterns to dynamically learn key hyperparameters such as learning rate, best-fit feature extractor choice, and degree of tendency toward relying on explicit rule-based learning vs. implicit information-integration learning.

This study has several limitations. While the approach to modeling human visual category learning using linear models trained on deep neural network embeddings with simple plasticity rules was validated in a previous study [167], the present work evaluated only a single feature extractor (ResNet-50 adversarially pretrained on ImageNet). Moreover, the main linear surrogate model does not account for several important aspects of human learning, including attention, fatigue, and explicit or verbal reasoning. A novel model of verbal explicit rule learning was introduced, but its ability to predict human learning dynamics has not yet been fully validated.

Although this work demonstrates that genetic curriculum optimization robustly improves surrogate performance across two tasks (a binary histology task and a 4-class dermoscopy task), the human experiment evaluated only a single optimized "champion" curriculum, and only for the dermoscopy task. Additionally, the participants were all lay individuals recruited online, and were only assessed immediately after training. Future work should evaluate optimized curricula across multiple independently optimized champions, in more targeted learner populations such as dermatology, pathology, or radiology trainees, and at multiple post-learning time points (similar to Chapter 4).

Finally, the dermoscopy task used in the human experiment does not fully capture the clinical complexity of real-world dermoscopy, both in terms of the number of lesion categories and the diversity of patient populations. In particular, the HAM10000 dataset [130] is largely limited to images of lesions on lighter skin tones. Validation on more demographically diverse datasets [138], ideally at larger scale, will be necessary before any real-world applications in dermatology training. More broadly, the possibility of curriculum optimization inadvertently amplifying dataset-specific biases through image selection warrants careful investigation.

Ultimately, this chapter demonstrates for the first time that human learning can be accelerated using image-computable machine learning models of the human learning process itself, moving beyond heuristics about what helps humans learn and simulations of visual perception in isolation (as in Chapters 3-4). The results have encouraging implications for the real-world applicability of goal-driven, increasingly high-fidelity models of human visual perception and learning [198], as well as for direct future applications in educational settings.

Chapter 6

Conclusion

The central hypothesis of this dissertation is that artificial neural networks (ANNs), when functionally aligned with the brain, can serve as more than predictive models: they can be repurposed as generative engines for optimizing human perception and learning. I have explored this hypothesis by engineering image-computable models that emulate human constraints and capabilities, and inverting such models to enhance human performance.

In Chapter 2, I addressed a key divergence in the learning dynamics of humans and ANNs: the latter suffer from catastrophic forgetting, especially when constrained to more human-like learning environments. I showed how this can be mitigated by a brain-inspired algorithm ("CRUMB") that displays emergent human-like perceptual characteristics. In Chapters 3 and 4, I leveraged the perceptual alignment of adversarially trained ANNs to predict visual categorization difficulty and generate category-specific image feature enhancements. I applied these predictions and enhancements to demonstrate empirical augmentation of human learning across multiple visual tasks and participant populations, culminating in a longitudinal study of first-year pathology residents learning histology with assistance from AI models. Finally, in Chapter 5, I demonstrated that "surrogate learner" machine learning models can simulate the acquisition of visual expertise end-to-end, enabling evolutionary optimization of curricula that experiments showed were superior to random control curricula for human learners.

This transition to surrogate-based curriculum optimization marks a fundamental shift from the heuristic-based curriculum design approaches in earlier chapters. In the current iteration, surrogate learners are viewed as tireless test subjects that can learn from a curriculum, undergo evaluation, reinitialize themselves as blank slates, and repeat ad infinitum to optimize stimuli for humans. The present work demonstrates the feasibility of this approach with a combination of an inefficient optimization approach (genetic algorithms) and an extremely inexpensive surrogate (a linear classifier with pre-computed image embeddings).

However, the potential of this approach extends far beyond linear probes. It is reasonable to ask how a much heavier ANN model, such as the brain-inspired CRUMB continual learner, might be used to optimize curricula within this framework. Such a model could render more complex and educationally relevant paradigms amenable to surrogate-based optimization, such as maximizing retention of multiple sequentially-learned tasks—a scenario where a surrogate immunized against catastrophic forgetting is a prerequisite. More broadly, well-designed surrogate ANNs could represent non-linear and more functionally realistic simulations of learning, but this comes with a massive increase in computational cost that limits the practicality of evolutionary optimization. As a more efficient alternative, gradients of a measured learning outcome could be backpropagated through an entire learning trajectory, creating pixel-level perturbations similar to those of the enhanced images introduced in Chapter 3, but optimized end-to-end for *learning* instead of categorical perceptibility. Freedom from the constraints of selecting among natural images may be a double-edged sword: I speculate that generating coherent curricula through unconstrained gradient-based perturbations would necessitate more stringent fidelity of the surrogate to human perception and cognition. The evolutionary approach of Chapter 5 implicitly benefited from a powerful inductive bias by limiting curricula entirely to natural images. Hybrid gradient-based selection of natural images [126], if applied to loss functions representing post-curriculum surrogate learning outcomes, could provide an avenue toward efficient gradient-based curriculum optimization within a space defined by natural stimuli.

While this thesis has focused largely on translating advances in image-computable models to practical engineering applications, I anticipate that such efforts will reciprocally drive the development of improved, more human-aligned models. A primary goal of building computational models, and perhaps the gold-standard test of their usefulness, is to generate testable predictions that extend beyond the principles and observations that informed the models' construction: translational applications provide ideal opportunities for testing the predictive capabilities of human-like AI systems. A key question for the field is the extent to which the human-facing effectiveness of stimuli produced by inverting AI models, meaning their measurable effects on neural activity, behavior, or both, is strongly associated with functional similarity between these AI models and the human brain. If we accept this premise, then practical utility can be used as a metric for model fidelity. Measuring the effects of inversion-generated stimuli can provide an important signal to guide the development of increasingly human-like AI. While the proposed operational definition of "human-like" is grounded directly in translational applications, the resulting models may also reveal new insights about underlying neural processes.

One simple example of such a translationally-defined signal can be found in Figure C.9 (Appendix C), representing a controlled experiment showing that some ANN configurations/architectures are more effective than others at producing low-norm image perturbations that enhance human visual category perception. Surrogate learners could provide a similar signal that adjudicates the relative fidelity of competing surrogates posited to simulate human learning dynamics. Ultimately, we might envision benchmarks in the tradition of BrainScore [11]; instead of comparing the predictivity of neural activity and behavioral responses by different models, we would directly benchmark their efficacy towards well-defined translational goals.

In the broader scope, the basic paradigm explored here—that of using surrogates to discover inputs that drive the brain toward a desired state—need not be limited to education. Many neurological and psychiatric conditions can be conceptualized as disorders of learning, plasticity, or maladaptive processing. Furthermore, psychotherapy-mediated recovery from many forms of mental illness could certainly be considered as a kind of learning. Imagine an ANN-based surrogate simulating the neural and behavioral dynamics of a psychiatric condition, or a patient recovering from a stroke, or an elderly person experiencing cognitive decline. If such models were developed, they could potentially be inverted to discover powerful therapeutic stimuli, sensory inputs, or even interactive tasks engineered to drive the brain toward healthier states. Challenges abound in developing reproducible benchmarks for such applications, notably the need for rigorous standardization and the long time scales and high costs of clinical studies. And yet, the potential reward would be a new feedback loop between AI development and clinical medicine, in which models of the brain and its ailments are validated by their therapeutic capabilities. By demonstrating how aligning AI models with the brain and mind can enhance human visual learning, I hope that my work represents an initial step toward that vision.

Appendix A

Statements of Author Contributions

This Appendix details the author's contributions to chapters that are based on multi-author publications and manuscripts.

A.1 Author Contributions for "Tuned Compositional Feature Replays for Efficient Stream Learning"

This chapter is based on the following publication: Morgan B. Talbot*, Rushikesh Zawar*, Rohil Badkundri, Mengmi Zhang†, and Gabriel Kreiman†, "Tuned Compositional Feature Replays for Efficient Stream Learning." *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 2, pp. 3300-3314, Feb. 2025. *Equal contribution, †corresponding authors.

I am a co-first author of this work (equal contribution with Rushikesh Zawar). Our collaboration involved extensive technical discussions throughout the project. Rushikesh Zawar led the baseline comparisons, benchmarking, and figure design, including the implementation and testing/benchmarking of a wide variety of competing baseline models. Rohil Badkundri, Mengmi Zhang, and Gabriel Kreiman developed the initial version of the model. My primary contributions were: (1) iteratively improving the architecture to the final model reported in the paper (which involved designing, implementing, and validating several modifications that simplified the model while preserving or improving performance), (2) designing and running the model analysis experiments reported in the paper (e.g., ablation studies, interpretability analysis, analyses of shape vs. texture bias), and (3) rewriting and preparing the manuscript through several submissions and revisions. Rushikesh Zawar, Mengmi Zhang, and I collaborated particularly closely on technical decisions regarding the model architecture. Mengmi Zhang and Gabriel Kreiman provided mentorship and guidance throughout the project. The work was collaborative and benefited from extensive technical discussions among the co-authors.

A.2 Author Contributions for "L-WISE: Boosting human visual category learning through model-based image selection and enhancement"

This chapter is based on the following publication: Morgan B. Talbot, Gabriel Kreiman, James J. DiCarlo and Guy Gaziv, "L-WISE: Boosting Human Visual Category Learning Through Model-Based Image Selection and Enhancement." *International Conference on Learning Representations (ICLR)*, 2025.

I am the first author of this work. I conceived and developed the methodology, designed and implemented the computational pipeline, designed and ran the experiments, collected behavioral data, performed all data analyses, and prepared the original manuscript and figures. Guy Gaziv led a redesign of the main-text figures. Guy Gaziv and Gabriel Kreiman provided mentorship and guidance throughout the project. All co-authors provided substantial feedback/editing of the main manuscript.

A.3 Author Contributions for "Accelerating Histology Learning with Perceptually Aligned AI: A Randomized Study of First-Year Pathology Residents"

This chapter is based on a manuscript in preparation for journal submission, authored by Morgan B. Talbot, Anurag Vaidya, Kenneth Chian, Marina Muratova, Richard N. Mitchell, Guy Gaziv, and Gabriel Kreiman.

I am the first author of this manuscript. I conceived and developed the study methodology, implemented all computational pipelines and technical systems, ran all experiments and collected all behavioral data, performed all quantitative data analyses, and prepared the original draft and figures. Anurag Vaidya and I collaborated on the design of the automated histology image quality control pipeline, which I implemented. Gabriel Kreiman, Guy Gaziv, and Richard Mitchell provided mentorship and guidance on study design and interpretation. Marina Muratova and Kenneth Chian assisted with interpretation of the study results and analysis of the effects of image perturbations on histology image features, and conducted the initial qualitative data analysis of questionnaire data, which I subsequently revised and expanded.

Appendix B

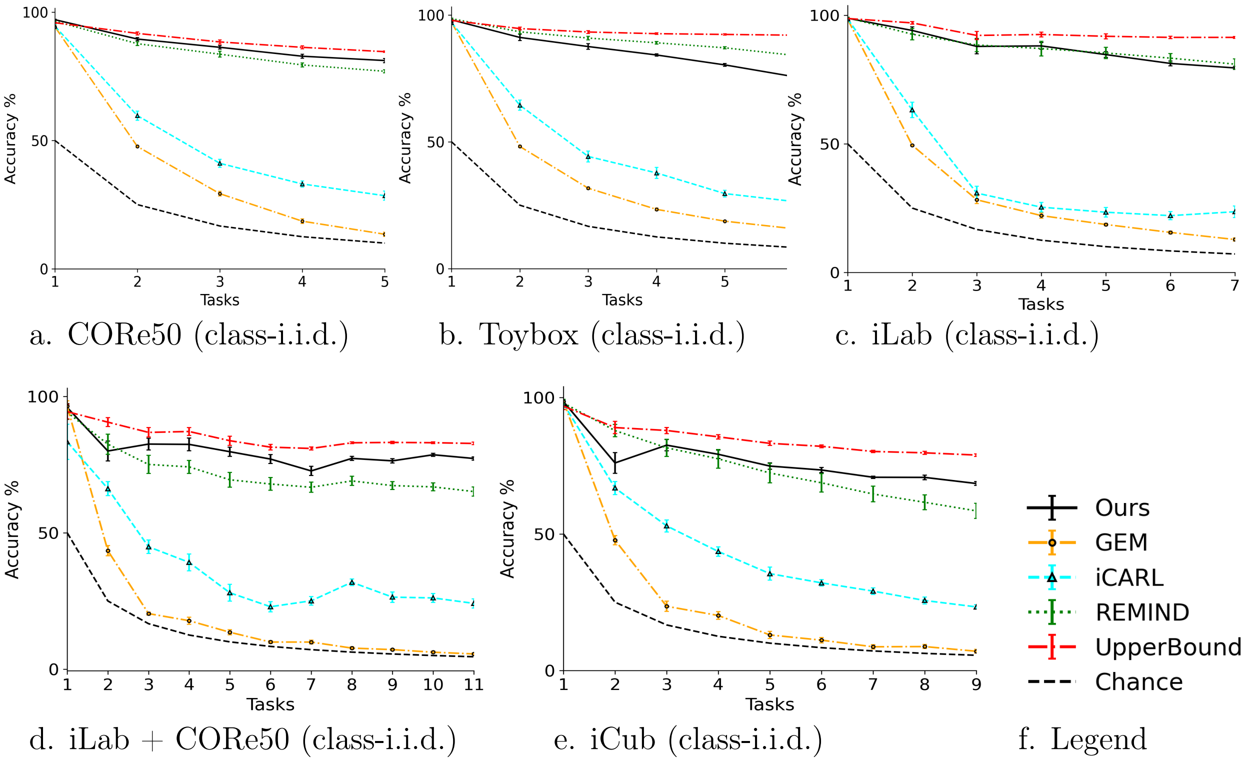
Tuned Compositional Feature Replays for Efficient Stream Learning

B.1 Compositional Replay Using Memory Blocks (CRUMB) outperforms competing algorithms in the class-i.i.d. setting in most cases

We report top-1 accuracy results (measured on all tasks/classes in each dataset at the end of stream learning training) for CRUMB and all competing baseline algorithms on five video streaming datasets (CORe50 [77], Toybox [37], iLab [38], iLab+CORe50, and iCub [78]) and two image datasets (Online-CIFAR100 [79], Online-Imagenet [80]) in both class-i.i.d. and class-instance training protocols in Table 2.1 in the main text. For CRUMB and a subset of baseline algorithms, we illustrate task-by-task top-1 accuracy (on all previously seen classes) for the five video datasets in the class-i.i.d. setting in Fig. B.1. Class-instance plots for video datasets, and class-i.i.d. plots for image datasets, are in main text Fig. 2.3.

B.2 CRUMB's performance is competitive with baseline algorithms even when memory buffer size is unlimited

Our primary baseline comparison experiments in the main text focus on comparing CRUMB with competing algorithms under a fixed memory budget: methods that store and replay entire raw images cannot store as many training examples, affecting their continual learning performance. CRUMB is specifically designed with memory-constrained conditions in mind, and compresses each training example stored in its replay buffer to occupy only 3.6% as much memory as an entire image (e.g., as stored by image-based replay baseline iCARL [57]). Nonetheless, we demonstrate here in Fig. B.2 that CRUMB obtains competitive performance even when memory usage is unlimited, greatly outperforming iCARL. CRUMB's accuracy after training on all tasks is slightly lower than that of REMIND [28] in the class-instance setting under these conditions, with both methods close to the offline upper bound in both class-instance and class-i.i.d.



a. CORe50 (class-i.i.d.)

b. Toybox (class-i.i.d.)

c. iLab (class-i.i.d.)

d. iLab + CORe50 (class-i.i.d.)

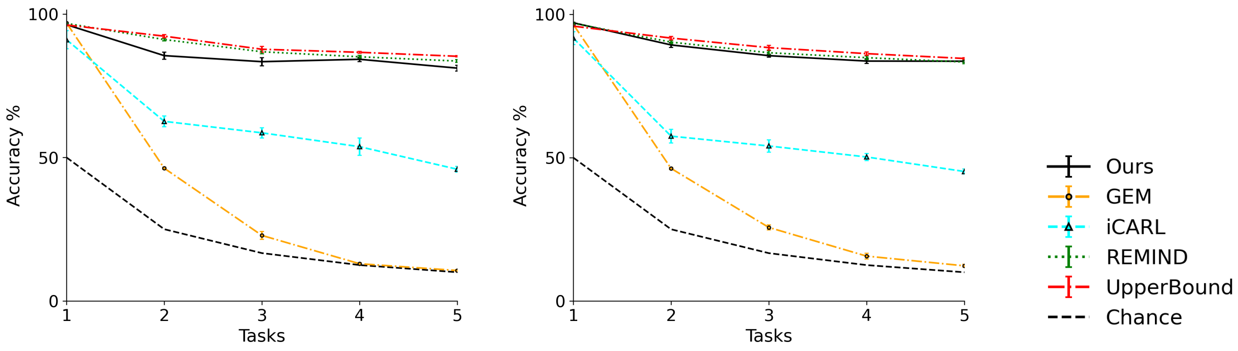
e. iCub (class-i.i.d.)

f. Legend

Figure B.1: **In the class-i.i.d. setting, CRUMB outperforms most baseline algorithms and performs near the the upper bound on some datasets**. Line plots show top-1 accuracy in online stream learning on video datasets (a) CORe50, (b) Toybox, (c) iLab, (d) iLab + CORe50, and (e) iCub in the class-i.i.d. setting. All models train on the first task for many epochs, but view each image only once on all subsequent tasks. Accuracy estimates are the mean from 10 runs, where each run has different class and image/video clip orderings. Error bars show the root-mean-square error (RMSE) among runs. Results for all baselines are in Table 2.1 in the main text.

B.3 CRUMB maintains its bias towards object shape information after class-i.i.d. stream learning on video datasets

In Section 2.5.4 and Fig. 2.4 in the main text, we observe that pretraining CRUMB on ImageNet [80] induces a bias in the CNN towards attending to object shape information more than image texture information, an effect that has been shown to mitigate catastrophic forgetting by flattening the loss minimum of each task [94]. Fig. 2.4 in the main text visualizes the extent of this "shape bias" for CRUMB trained on video datasets in the class-instance setting. Fig. B.3 shows results in the class-i.i.d. setting. As for class-instance, we observe that CRUMB mostly retains its shape bias during class-i.i.d. stream learning.



a. CORe50 (class-instance) b. CORe50 (class-i.i.d.) c. Legend

Figure B.2: **CRUMB attains competitive levels of performance in conditions of unlimited memory usage**. Line plots show top-1 accuracy in online stream learning on the video dataset CORe50. For this comparison among replay methods, all models are allowed to store all previously encountered images in a replay buffer and intersperse them with images encountered while training on new tasks. All models train on the first task for many epochs, but view each image only once on all subsequent tasks. Accuracy estimates are the mean from 10 runs, where each run has different class and image/video clip orderings. Error bars show the root-mean-square error (RMSE) among runs.

B.4 Data analysis

B.4.1 Data cleaning

For our main results on the video datasets CORe50, Toybox, and iLab, we noticed that a small subset of runs for some models had markedly reduced accuracy on the first task compared to other runs. To facilitate fair comparisons among models, we excluded all runs with an initial task accuracy less than 80% from all analysis and results. For the small number of algorithm/dataset/protocol combinations for which no runs exceeded 80% on the first task, we filtered at a 60% threshold, or a 40% threshold if no runs exceeded 60%. We did not encounter this issue for any runs of CRUMB on any dataset, or for any method on Online-CIFAR100 and Online-Imagenet.

B.4.2 Statistics for model analysis experiments

Our model analysis experiments in main-text Section 2.5.4 compared the performance of CRUMB with various ablated or otherwise perturbed versions of CRUMB. For each comparison with the original algorithm, we evaluated statistical significance of pairwise differences using the following method:

i. Divide the test set from the dataset being used into batches of 100 images. The images should be randomly sampled without replacement, and the sampling should be done only once (or, using a fixed random seed) for all experiments such that each version of the algorithm is evaluated on the exact same batches of images.

ii. Evaluate CRUMB and each experimentally perturbed version of CRUMB on the same set of image batches and record mean top-1 accuracy on each batch. This is done for each

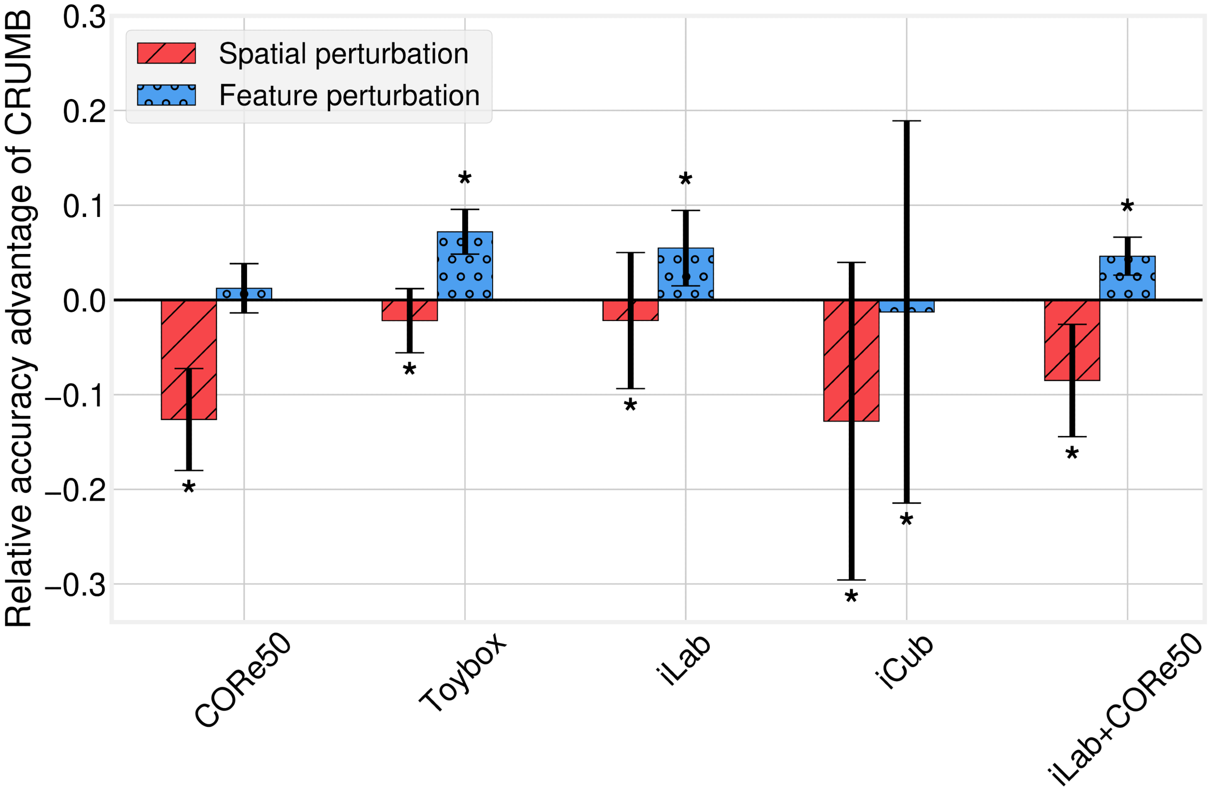


Figure B.3: **The bias towards shape information induced by CRUMB pretraining persists through stream learning in the class-i.i.d. setting**. The height of each bar shows how much smaller (or larger, if negative) CRUMB's drop in normalized test set accuracy under a perturbation is, in comparison to a control network. "Relative accuracy advantage" is calculated by dividing the difference in accuracy caused by a perturbation by the unperturbed accuracy, and then subtracting this result for CRUMB from that of the control network (see main-text Section 2.5.4). "Spatial perturbation" shuffles the spatial positions of all feature vectors in an intermediate feature map (at the same layer where it is reconstructed by CRUMB), "feature perturbation" randomly sets half of the feature map's features to zero, and "style perturbation" uses images from Stylized-ImageNet [93]. Streaming results (to the right of grey dotted line) are in the class-i.i.d. setting: class-instance shape-texture bias results for the video datasets, and class-i.i.d. results for Online-CIFAR100 and Online-ImageNet, are available in Fig. 2.4 in the main text. Error bars are standard errors of the mean of relative accuracy advantage among or 10 independent runs. * denotes a statistically significant difference from 0 (see Section B.4).

of the 5 independent training runs, and accuracies are pooled across runs. Therefore, for each training protocol (class-instance and class-i.i.d., for which all analyses are kept separate), each version of the algorithm has $n_r \times n_b$ top-1 accuracy estimates, where n_r is the number of runs and n_b is the number of 100-image batches in the test set. Conceptually, we treat the accuracy on each batch as an independent sample indicating the accuracy of the corresponding algorithm on a roughly continuous scale, with each run of each algorithm tested on the exact same batches of images.

iii. Perform a paired-samples t-test for each comparison, using accuracy on each image batch of CRUMB and the perturbed version of CRUMB as a sample pair and pooling sample pairs across runs. We used a global p-value cutoff of $p < 0.01$ to report the statistical significance of t-test results for each comparison between CRUMB and a perturbed version of CRUMB.

For our experiments on shape-texture bias, we employ a similar approach by first calculating accuracy on batches of 100 images at a time. For each batch, we subtract the model's perturbed accuracy (i.e., after spatial, feature, or style perturbation, see Section 2.5.4 in the main text) from the unperturbed accuracy, and divide the result by the unperturbed accuracy to obtain the relative accuracy drop for each perturbation. We compare the relative accuracy drops for CRUMB and a control network on all batches, pooling across runs with different data orderings, using the Wilcoxon signed-rank test for paired samples. We apply a global p-value cutoff of $p < 0.01$ to report the significance of any differences, visualized as CRUMB's relative accuracy advantage being either above or below zero in main-text Fig. 2.4 and Fig. B.3 in this Appendix.

B.5 Replay buffer size calculations

For replay-based baseline algorithms, we limit the number of examples that can be stored in the buffer to fit within a memory budget that is held constant for all methods in our main results (main text Table 2.1). We do not apply this constraint for weight regularization approaches. To calculate the maximum number of training examples we can store in the replay buffer for each experiment, we first set the number of examples n_{raw} that raw-image replay methods such as iCARL may store, then calculate how many examples (n_x) CRUMB can fit into the same amount of memory using the formula:

$$n_x = \frac{n_r(3w_ih_i) - bd}{swh/d} \tag{B.1}$$

Where w_i and h_i are raw image width and height respectively (224 × 224 for our experiments), the codebook matrix has dimensions $b \times d$ (b memory blocks, each of dimension d), and the feature map has dimensions $s \times w \times h$ (s features in a $w \times h$ spatial grid). The numerator corresponds to the number of 8-bit RGB values needed to store one image, subtracting a discounting factor for the number of values in the memory blocks themselves. The denominator corresponds to the number of 8-bit integer indices required to encode one feature map. Concretely, the memory budgets are 2.2 MB on CORe50, Toybox, and iLab, 14.3 MB on CIFAR100, and 1.44 GB on ImageNet based on the number of 8-bit integers each method stores per training example.

For direct comparisons between algorithms in our main results, we applied both CRUMB and REMIND to the SqueezeNet network architecture [81]. To calculate n_x for REMIND, we multiplied the compression ratio provided by the REMIND paper (959,665 feature maps/10,000 raw images) by the ratio of values in one feature map from ResNet18 (used in the REMIND paper, $512 \times 7 \times 7$) to those in one feature map from SqueezeNet ($512 \times 13 \times 13$) [28]. We then multiplied the resulting ratio of 278,246 feature maps/10,000 raw images by n_{raw} to obtain the corresponding n_x for each dataset.

Appendix C

L-WISE: Boosting Human Visual Category Learning Through Model-Based Image Selection and Enhancement

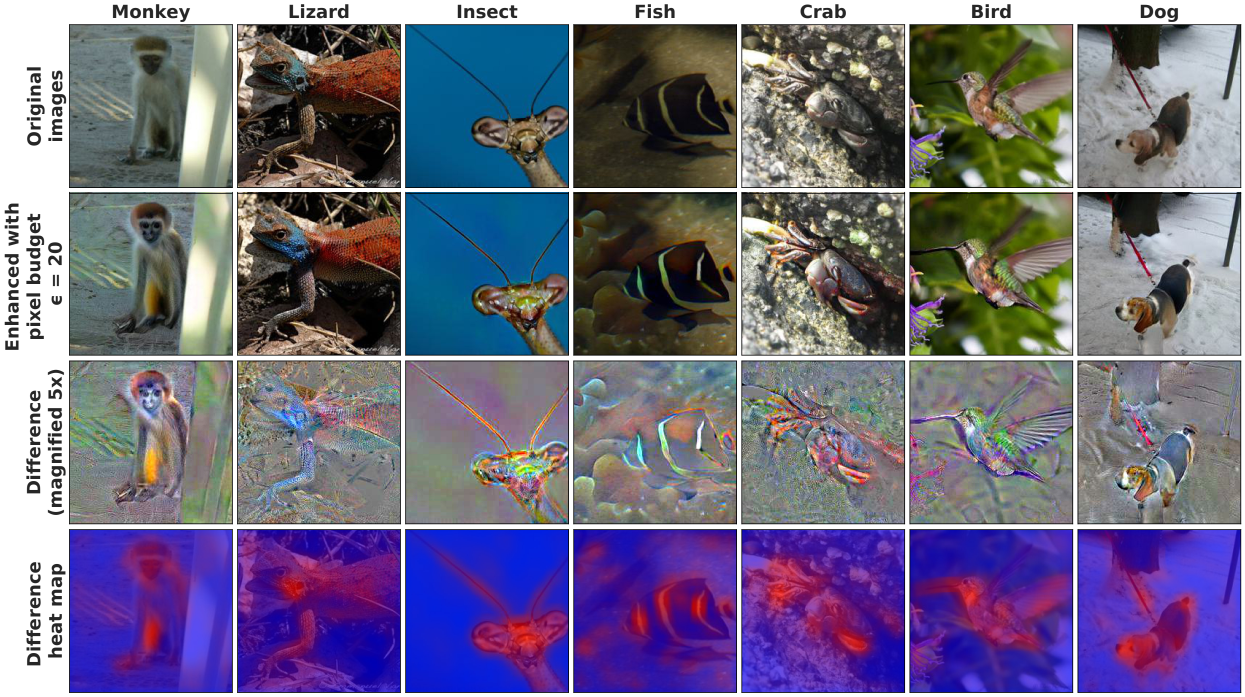


Figure C.1: **Ground truth logit enhancement with robustified ANNs leads to semantically meaningful perturbations.** The top row shows original ImageNet images, and the second row shows the same images after enhancement by robustified ResNet-50 (training $\epsilon = 3$) with a pixel budget of $\epsilon = 20$. 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.

C.1 Details on Training and Using Robustified Guide Models

We adversarially trained a ResNet-50 model on ImageNet [80], and another on iNaturalist 2021 [199], with the hyperparameters following [101]:

- 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 [130], even though we only used 4 of them in the learning task for humans. 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 3.1).

For the moth task, we adversarially fine-tuned the (adversarially) iNaturalist-pretrained model on the four moth classes to be used in the task. We subjectively judged the perturbations from this fine-tuned model to be of higher quality 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 adversarial fine-tuning used $\epsilon = 1.0$ with 7 gradient steps of size 0.3. We used learning rates of 0.0001, 0.0004, and 0.001, and batch sizes of 32, 64, and 16, for *Idaea* moth, dermoscopy, and histology fine-tuning respectively. We fine-tuned the entire network end-to-end for each task.

Our choice of $\epsilon = 3$ for ImageNet pretraining follows [101], who found this to be an optimal choice for generating perturbations that disrupt category perception (relative to $\epsilon = 1$ 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 $\epsilon = 3$ (iNaturalist pretraining, and fine-tuning on moth photos, dermoscopy images, and histology images) - therefore, we reverted to $\epsilon = 1$ for these settings.

To enhance images in a category-specific manner, we perform the optimization of Equation 3.1 (main text) in a series of steps using projected gradient ascent (Equation C.1), where

k denotes the optimization step, η the step size, and Proj_ϵ a projection onto a hypersphere of radius ϵ with original image x at its center (see text below Equation 3.1 for definitions of other symbols).

$$\delta_{k+1} = \mathrm{Proj}_\epsilon\Bigg(\delta_k + \eta\nabla_\delta\Big(L_{\mathrm{gt}}(x+\delta_k) - \frac{\alpha}{|C|-1}\sum_{c\in C:c\neq\mathrm{gt}} L_c(x+\delta_k)\Big)\Bigg) \tag{C.1}$$

Fig. C.1 shows several example images enhanced with $\epsilon = 20$ using logit maximization by adversarially-pretrained ResNet-50 ($\epsilon = 3$), along with difference images and heat maps produced by the same method as Figs. 3.3-3.4 in the main text. Throughout our experiments, robustified ResNet-50 enhancements with pixel budget ϵ used $\mathrm{ceil}(2\epsilon)$ steps of $\eta = 0.5$ in a $224 \times 224 \times 3$ pixel space. This formula seems somewhat model-dependent and sometimes requires adjustment: for example, when enhancing images with robustified XCiT (see Figs. C.9-C.10), $\mathrm{ceil}(4\epsilon)$ steps were required instead of $\mathrm{ceil}(2\epsilon)$ to reach a similar effective perturbation size.

We generate the heat maps in Figs. 3.3, 3.4, and C.1 by subtracting the enhanced image x' from the original image x element-wise: $\delta = x' - x$. We calculate the magnitude of the changes as δ^2. We apply smoothing using 2D convolution, and normalize the result to have all values between 0 and 1, to produce δ_{norm}. We then produce the heat map by setting the red channel to $255 \times \delta_{\mathrm{norm}}$ and the blue channel to $255 \times (1 - \delta_{\mathrm{norm}})$. The resulting image shows red in regions where larger changes have taken place, and blue in regions where smaller or no changes have taken place. In Figs. 3.3, 3.4, and C.1, we superimpose a translucent version of the heat maps ($\alpha = 0.7$) over the original images. Averaging the heat maps across all ImageNet validation set images (Fig. C.2) indicates that changes to the images tend to occur in the central regions of the images more than in the periphery, consistent with the observation that our image enhancement approach tends to primarily change image regions corresponding to the main subject of each image.

C.2 Details on image preparation for experiments

All images presented in all psychophysics 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 [110]), Multi-Scale Retinex with Color Restoration (MSRCR [111,132]), and the "Auto" image tuning feature in Adobe Photoshop Lightroom (Auto-LR [133]) 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 [199]. In this case we used training set images during the training period of the human image category learning experiment and validation images during the

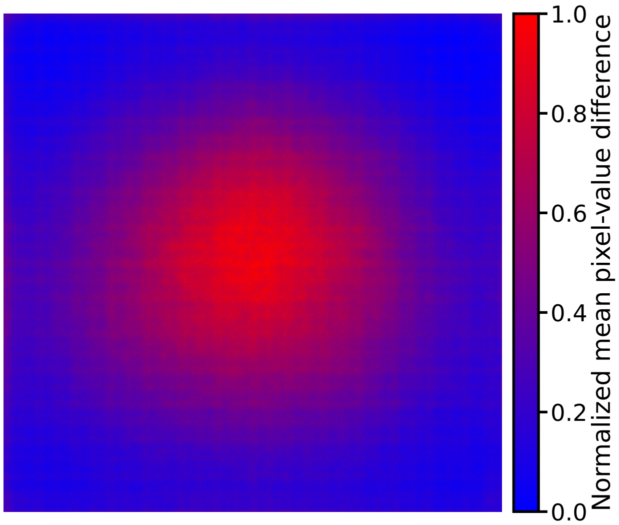


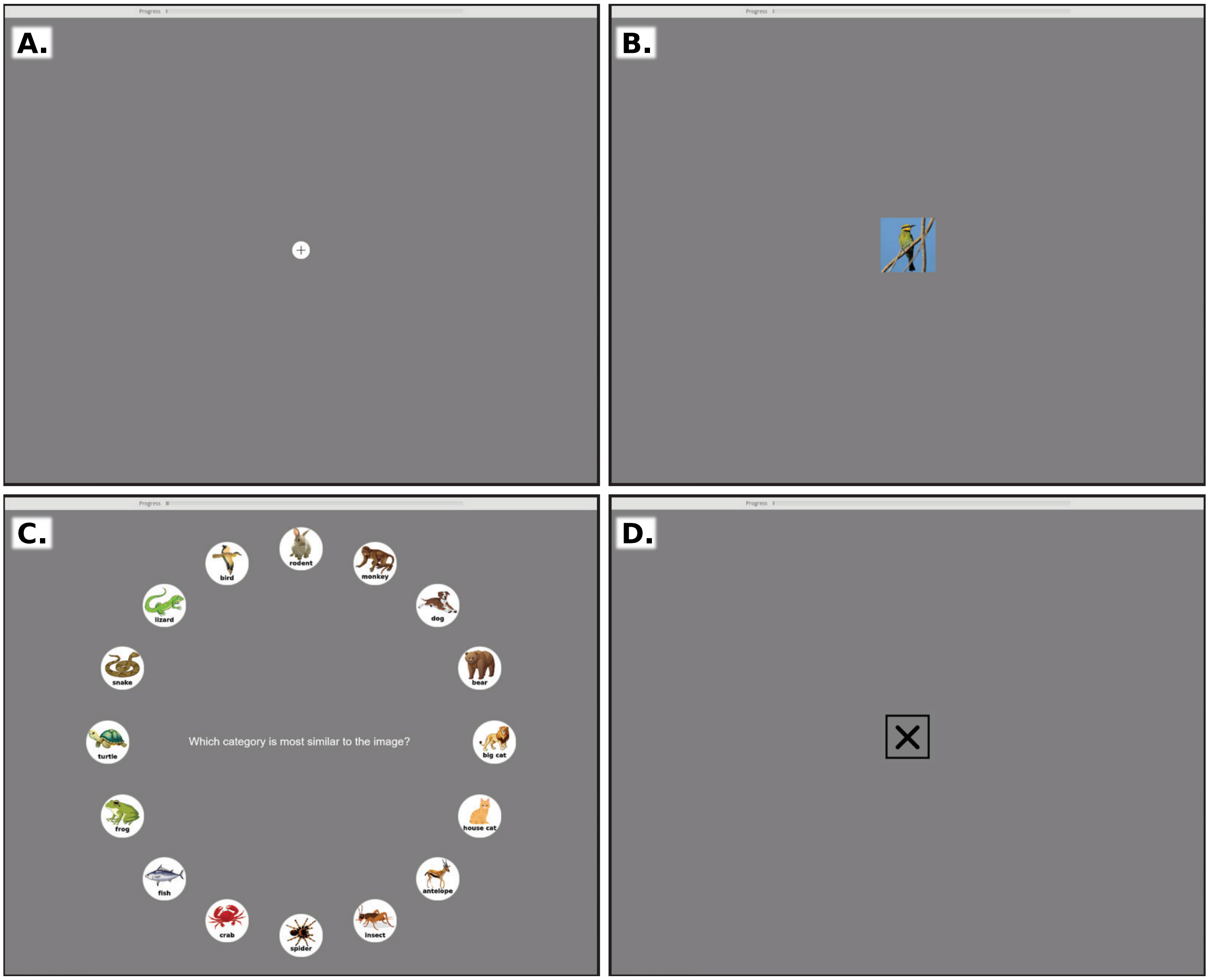
Figure C.2: **Pixel-value changes during enhancement of ImageNet images are biased towards the center of the image.** This heat map indicates which spatial regions of the ImageNet animal images were changed the most on average during logit-maximization enhancement with $\epsilon = 20$. Regions that were changed more on average are more red, and regions that were changed less on average are more blue. The heat map was generated by averaging the normalized absolute pixel value changes across all 2400 ImageNet validation set images that we used for our 16-way animal classification experiments.

test phase. We show that image enhancements are still effective for training set images in Fig. C.11A.

C.3 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 (e.g., see Figs. 3.1 and C.3), adapting and expanding the Restricted ImageNet dataset defined in the Robustness library [200]. 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
- Monkey: classes 369–382
- Fish: classes 0, 1, 389, 391, 392, 393, 394, 395, 396, 397



Which category is most similar to the image?

Figure C.3: **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 displayed 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 the 16 buttons shown in panel **C**, which are randomly rotated in position every trial, within a 15-second time limit. For incorrect responses, 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. C.4D 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.

- Crab: classes 118–121
- Insect: classes 300–320
- Lizard: classes 38–48
- Snake: classes 52–68

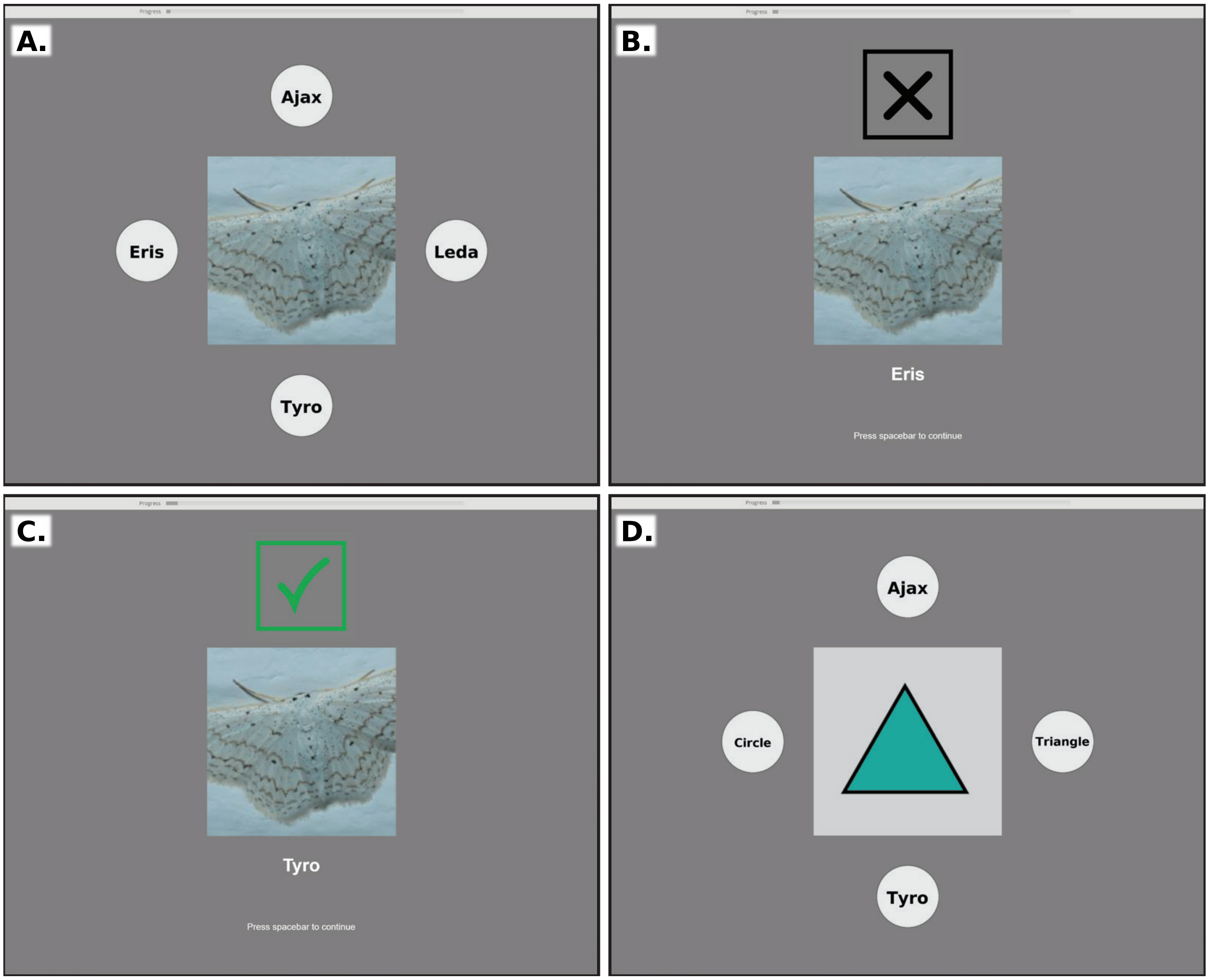


Figure C.4: **Task interface for image category learning experiments.** In the 4-way image category learning experiments (moths, dermoscopy), human subjects learned to classify four types of images that were represented by randomly assigned aliases "Ajax," "Eris," "Leda," and "Tyro." The image was shown 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.

- 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 or beavers in the “rodent” class (as many people may not recognize these as rodents), and we did not include eels in the “fish” class due to the possibility of confusion with snakes. We mistakenly classified rabbits and hares as “rodents” given that they were reclassified to the order Lagomorpha in 1912 [201] (we thank the participant who notified us of this).

C.4 Details on image category learning experiments

Figs. C.3 and C.4 show the task interfaces for the 16-way animal classification and 4-way image category learning experiments with human participants, respectively. For the learning experiments, the positions of the four buttons used to indicate responses are randomly permuted for each participant.

To minimize biases stemming from any priors induced by the class names, for each participant in the 4-category learning tasks we randomly assign a four-letter, two-syllable alias from the set “Ajax,” “Eris,” “Leda,” and “Tyro.” These names are drawn from Greek mythology, and each has 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. C.15 and C.16).

We used a different approach for the binary histology task that employed the MHIST dataset [131], giving “benign hyperplastic polyp” the alias “benign” and sessile serrated adenoma the alias “malignant” (although sessile serrated adenoma is actually a pre-cancerous lesion). The histology task has an interface very similar in appearance to that of the 4-way tasks (as shown in Fig. C.4), except that the “benign” and “malignant” buttons always appear on either 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.

C.5 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 (see Figs. 3.1A, C.5, C.7, and C.14A1,B1,C). We conducted an experiment to compare this ground truth logit score with prior state-of-the-art predictors of image difficulty for humans established by [108]. We apply logistic regression to predict correct v.s. incorrect responses to each (original, unmodified) 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 [106]), (2) prediction depth (earliest layer upon which a linear probe makes the same prediction as the final output [107]), (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. C.6). C-score,

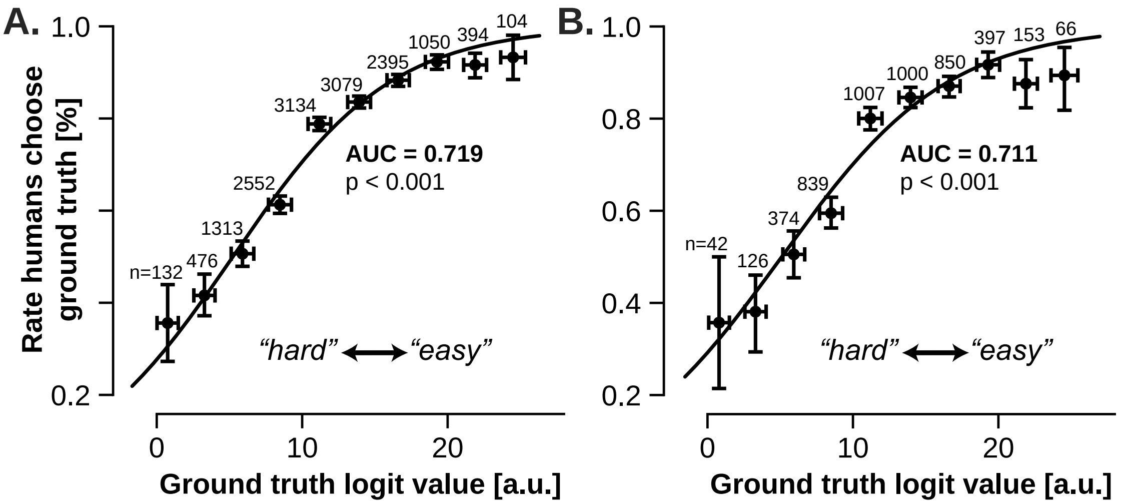


Figure C.5: **The observed relationship between robust model ground truth logits and human error rates is not sensitive to trial inclusion criteria.** For our main analysis regarding predicting ImageNet animal image recognition difficulty illustrated in Fig. 3.1A, we included all trials with original images but also trials featuring images modified by off-the-shelf control enhancement methods (Adobe Lightroom, CLAHE, and MSRCR), and trials with image perturbations from non-adversarially-trained ANNs (i.e., Vanilla ResNet-50 and CutMix ResNet-50 from Fig. C.7). Panel **A** above reproduces Fig. 3.1A1, while panel **B** replicates the same analysis but strictly including only trials with natural, unmodified images. The numbers above each point indicate how many image trial observations were included in the corresponding bin (the size of the vertical 95% confidence intervals is sensitive to this). AUC=Area Under the Receiver Operating Characteristic Curve, p-values derived from Wald statistic on the logistic regression coefficient for the ground truth logit predictor.

prediction depth, and adversarial robustness are implemented following [108]. 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. C.6). Furthermore, combining all other metrics with L_{gt} does not improve performance beyond L_{gt} alone. We also find that using a robustified model rather than a "vanilla" model to generate each metric greatly improves the predictivity of L_{gt} and (marginally) adversarial robustness, but not of the c-score or prediction depth.

C.6 Comparison of different ANN guide models for difficulty prediction and image enhancement

Our main results use robustified ResNet-50 as a guide model for generating image perturbations. To evaluate the importance of the choice of guide model, we compared the accuracy of difficulty prediction (Fig. C.7) and the effects of enhancement with $\epsilon = 20$ (Fig. C.9) 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 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 [101]), while perturbations guided by a non-adversarially-trained "vanilla" model have negligible effects. Model accuracy seems to be less important

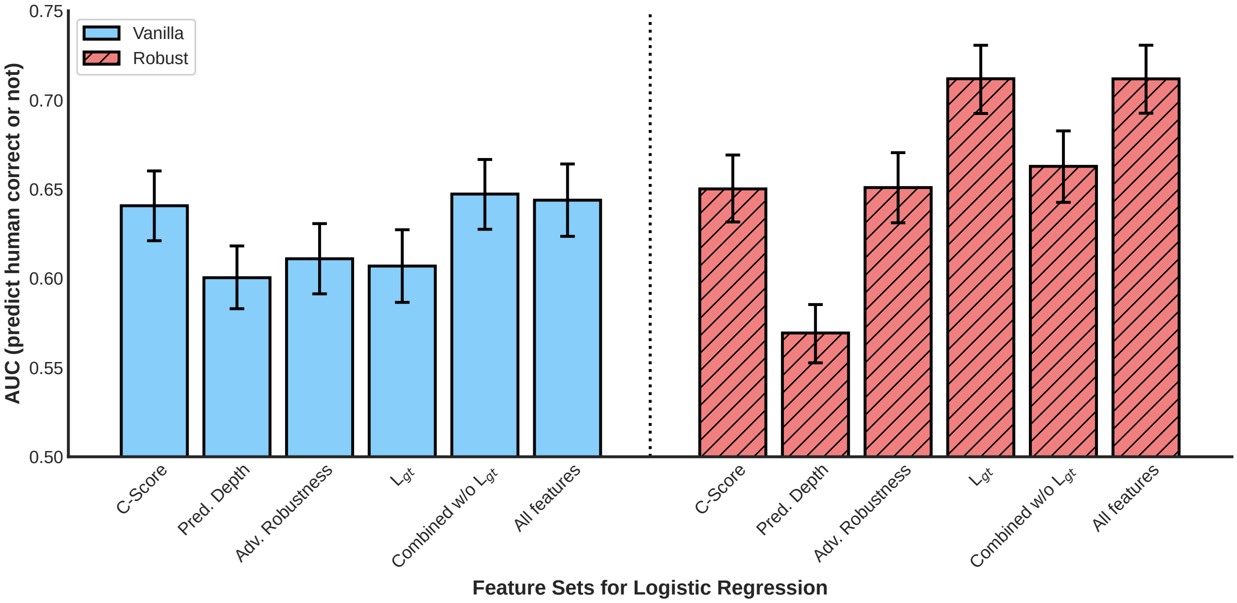


Figure C.6: **Robustified ground truth logit is a state-of-the-art predictor of image difficulty for humans, outperforming the c-score, prediction depth, and adversarial epsilon of both vanilla and robustified models.** AUC estimates are based on fitted logistic regression models using one or more features listed under each bar, with stratified 500-fold cross-validation. Error bars are 95% confidence intervals for the mean from 10,000 bootstrap replicates. The chance level is AUC=0.5.

than robustness, at least for difficulty prediction: ground truth logits from ResNet-50 models at early epochs of adversarial training on ImageNet, which have lower image classification accuracy than models at later epochs, can still yield accurate image difficulty predictions (Fig. C.8A). Models at early training epochs may also be able to generate high-quality image enhancement perturbations (Fig. C.8B), although we did not study this in our experiments with human participants. Although training ResNet-50 with CutMix improves its robustness to adversarial perturbations [202], 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 enhancement capabilities of an adversarially trained vision transformer model, the Cross-Covariance Image Transformer (XCiT) [134]. [203] 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. C.7 and C.9, we used pretrained guide models provided by [101] (Vanilla, $\epsilon = 1$, $\epsilon = 3$, and $\epsilon = 10$ ResNet-50 models), [202] (CutMix ResNet-50), and [203] ($\epsilon = 4$ XCiT). Examples of images enhanced by each of these guide models with $\epsilon = 20$ are displayed in Fig. C.10.

Image perturbations generated with vision transformer models (such as XCiT) typically include grid-like artifacts related to the image patch/grid structure of these models [204]. To mitigate grid artifacts during each step of generating image perturbations with XCiT, we calculated gradients with respect to each pixel value by averaging across ten randomly

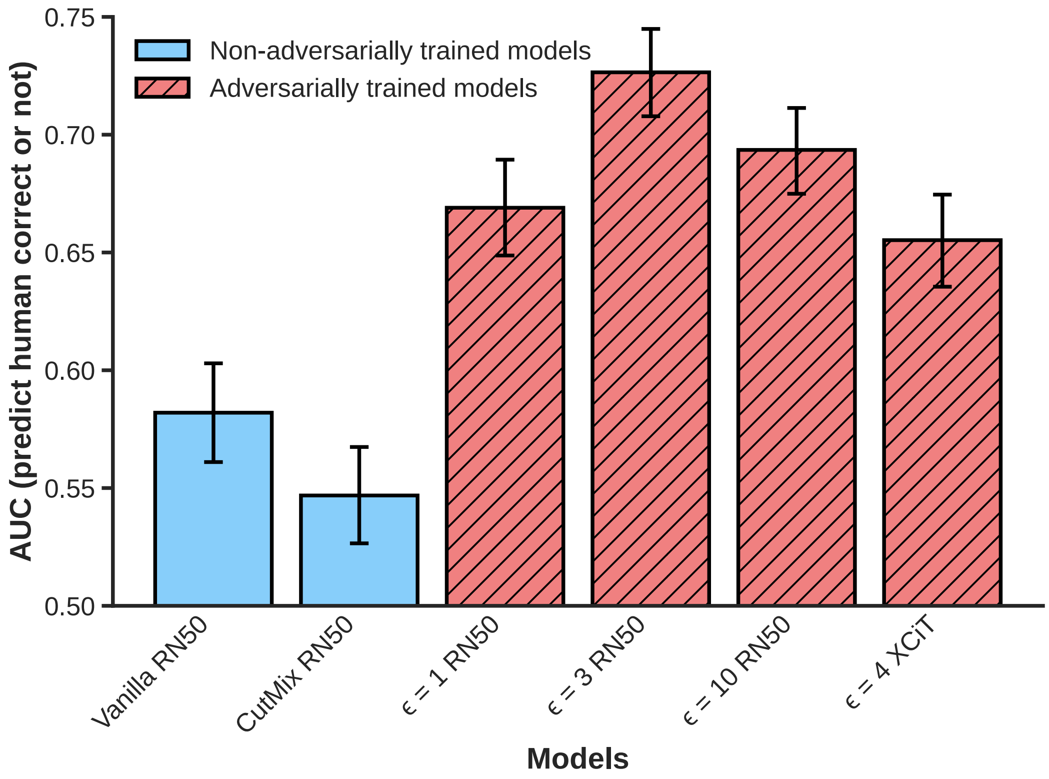


Figure C.7: **Accuracy of image difficulty prediction using ground truth logits from different model types.** AUC estimates and 95% confidence interval error bars are generated by the same procedure as in Fig. C.6 - however, results are not directly comparable between the two figures as different ANN 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.

translated, resized (followed by cropping/padding), color-jittered, and randomly cut-out "views" of each image, a strategy inspired by [205] that extends DiffAugment [206].

C.7 Ablation study on logit maximization approach to enhancement

As 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 $\epsilon = 10$. The results of this experiment are summarized in Fig. C.11A. 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 larger 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 [207], 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 validation set are drawn from a slightly different distribution than those in the training set, resulting in this accuracy discrepancy.

C.8 Additional results from image category learning experiments

Fig. 3.3 in the main text shows learning curves (mean smoothed 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. C.12 and the histology task in Fig. C.13. Panels H-M of Fig. C.12 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. C.13 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 [131] agreed on the same category label. On average, the model is more "confident" in its predictions (higher ground truth logit) on images where experts are more in agreement with each other.

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. 3.1 in the main text, and from the moth photograph, dermoscopy, and histology tasks in Fig. C.14 (Panels A1, B1, and C). For this analysis in the non-ImageNet tasks, we rely on test-phase data from control group participants who had just learned the tasks in question.

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. C.14A2,B2). This analysis is limited to the first training trial blocks in the "ET (shuffled)" participant group in the ablation study (main-text Table 3.1), the only group that viewed enhanced images without ϵ monotonically decreasing over time. 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 both the moth task and the dermoscopy task, participants respond with the original, correct category label statistically significantly more often when viewing images enhanced with greater ϵ (Fig. C.14B1,B2).

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 C.1. The same data are visualized in Fig. 3.4B. 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 a subset of classes.

C.9 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 3.1) were less likely to withdraw.

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 binomial distribution via Equation C.2 below (where p is the probability of being assigned to the control group). Equation C.2 evaluates here to a probability of $p = 0.02$, 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.

$$P(X \geq d) = 1 - P(X \leq d - 1) = 1 - \sum_{k=0}^{d-1} \binom{n}{k} p^k (1 - p)^{n-k} \tag{C.2}$$

Similarly, in the dermoscopy category learning experiment, 13 participants withdrew, of whom 6 were from the control group. In this case, Equation C.2 evaluates to a probability of $p = 0.041$, 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 3.1). 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.

	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 C.1: **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.

C.10 Notes on "hallucinations" in enhanced images

To support our approach to assisting human learners, we demonstrate the ability to enhance category percepts in images using low-norm perturbations. Previous work by [101] 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 disruptive perturbations could be described as "hallucinations:" perceptions (by the model and the human viewer) of objects that are 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: panel **B2** in Fig. 3.1 appears to show bolder contrasts and exaggerated features in class-relevant regions of the perturbed images. Panel **B** (image farthest to the left) in Fig. C.11 in this Appendix, however, 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.

C.11 Participant recruitment and demographics

We recruited a grand total of 521 participants via the online platform Prolific. All participants lived in the United States and were 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 a given 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 3.1) or difficulty selection (equivalent to DS in Table 3.1) participants to be roughly 10%, with a standard deviation in accuracy of roughly 10% in each group.

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = 0 \\ H_1 &: \mu_1 - \mu_2 \neq 0 \end{aligned}$$

$$\begin{aligned} &\text{Given:} \\ &\qquad \delta = 0.1 \text{ (estimated mean difference)} \\ &\qquad \sigma = 0.1 \text{ (estimated standard deviation)} \\ &\qquad \alpha = 0.05 \text{ (significance level)} \\ &\quad 1 - \beta = 0.8 \text{ (power)} \end{aligned}$$

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

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

We provide a demographic breakdown of the participants in our study, aggregated across experiments, in Table C.2. 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 C.2: Demographic characteristics of study participants, aggregated across all experiments.

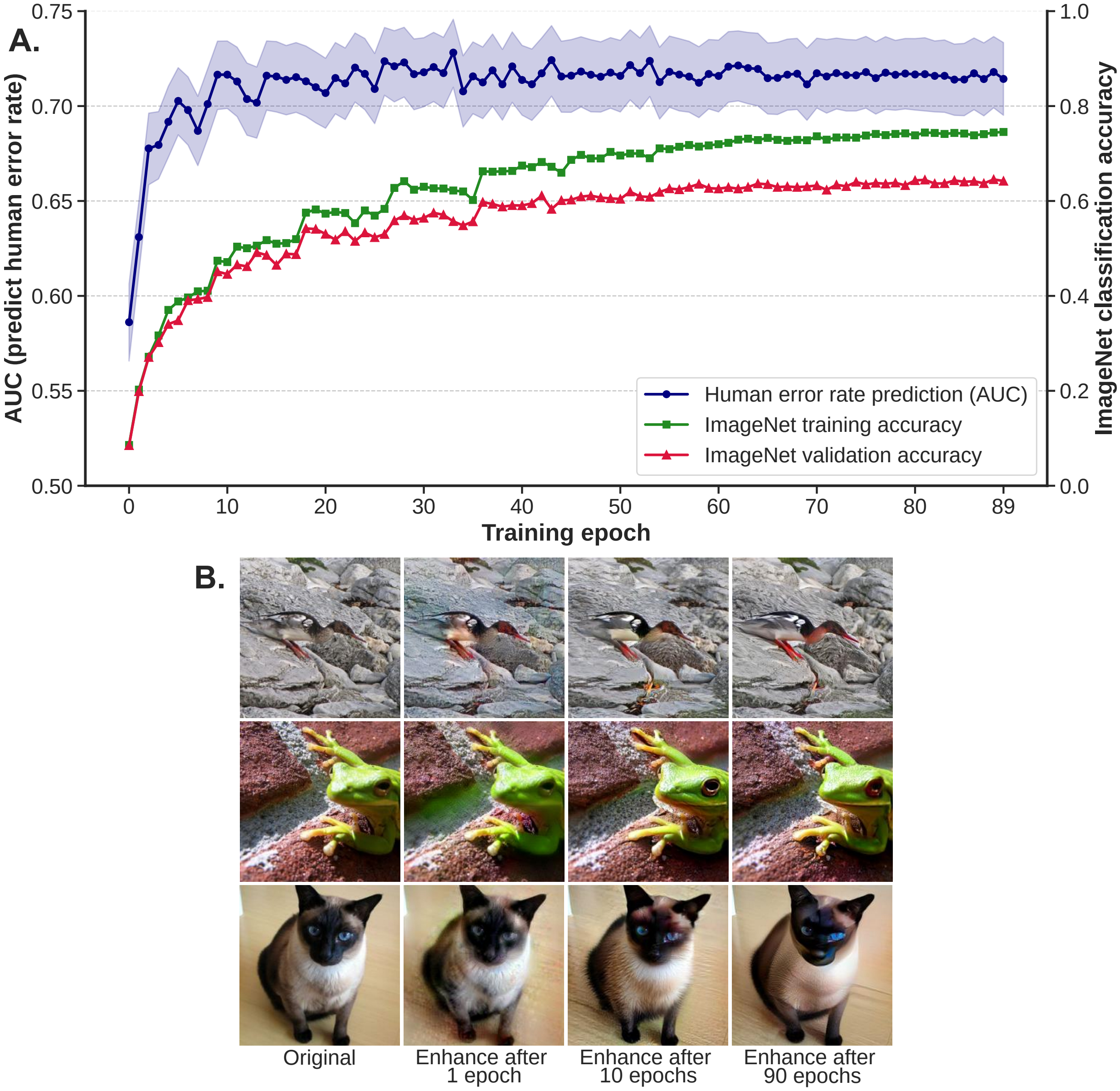


Figure C.8: **Robustified model accuracy weakly affects the relationship between ground truth logit and human image difficulty.** To generate this figure, we retained ResNet-50 model checkpoints from immediately after each of 90 epochs of adversarial training on ImageNet-1K from scratch ($\epsilon = 3$, batch size $= 256$, initial learning rate of 0.1 decreases by half every 9 epochs). In panel **A**, AUC estimates of logistic regression models predicting human trial response correctness were generated by the same procedure as in Figs. C.6-C.7 in this Appendix, using ground truth logits generated by each of the epoch checkpoints. The shaded area denotes 95% confidence intervals around each mean AUC from 10,000 bootstrap replicates. The training set and validation set accuracy at 1000-way ImageNet-1K classification, on non-perturbed images, is also plotted by training epoch. We observe that the AUC of human error rate prediction stops increasing relatively early during training, well before the training/validation accuracy on the ImageNet classification task is saturated. Panel **B** shows example images with $\epsilon = 20$ enhancements generated by checkpoints at various stages of the training process.

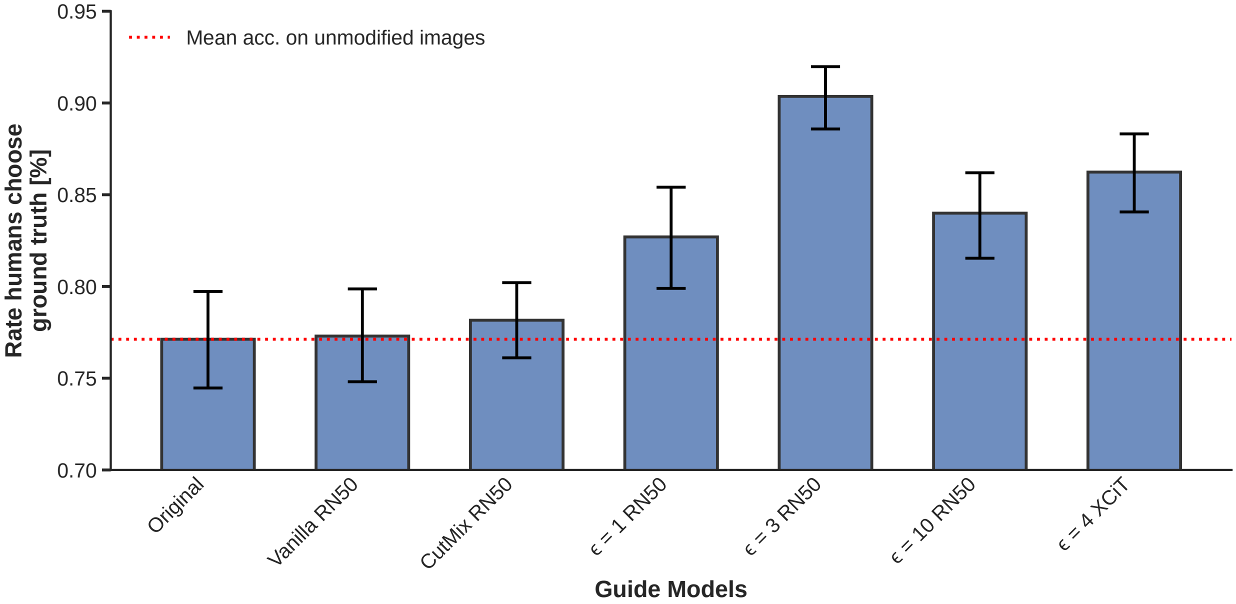


Figure C.9: **Effectiveness of image category enhancement across different guide model types.** Each bar shows the mean and 95% confidence interval (by bootstrap) of the rate at which humans choose the original ground truth label, in a 16-way basic animal classification task using ImageNet images. The "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.

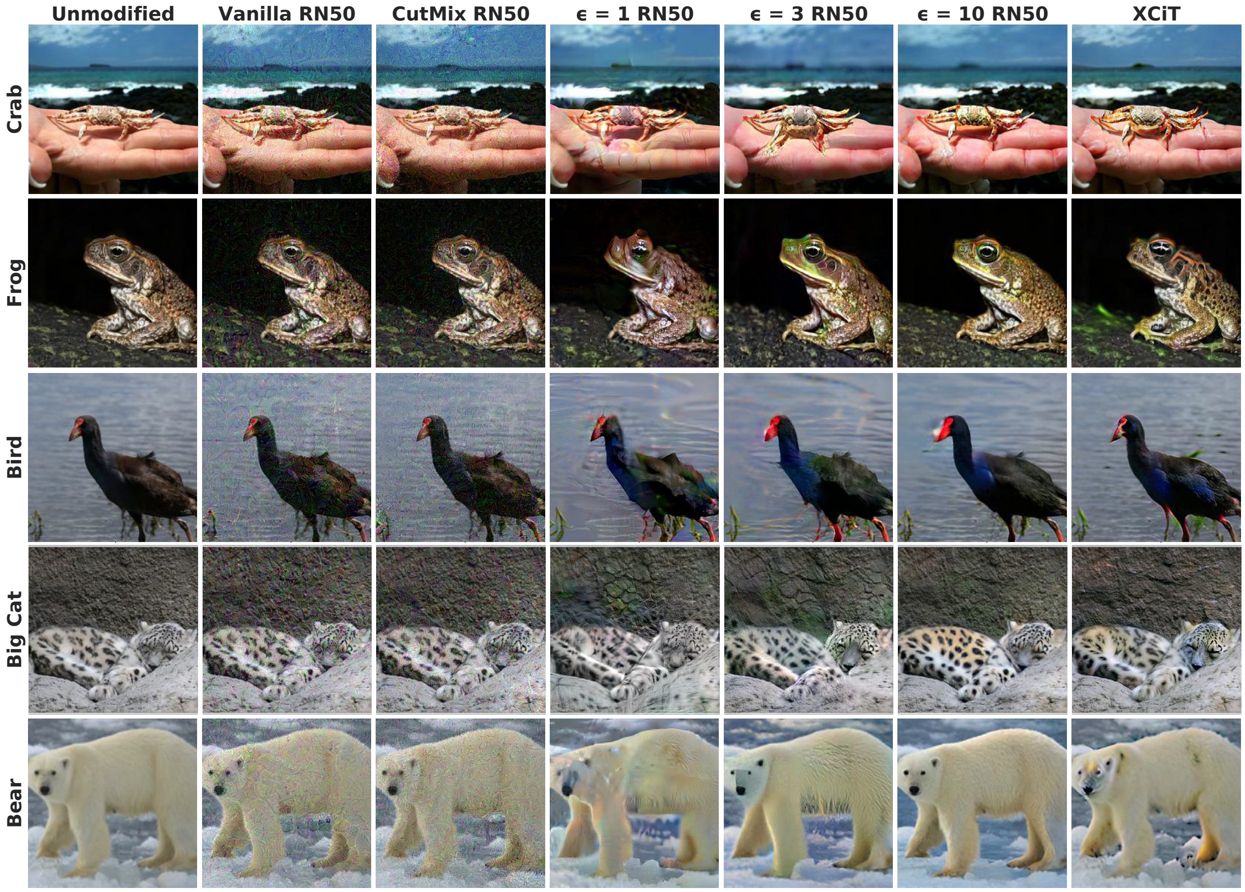


Figure C.10: **Meaningful perturbations require robust models, and are possible with CNN and vision transformer architectures.** Each row shows an image from ImageNet (original on the far left) enhanced with $\epsilon = 20$ by different guide models. A quantitative comparison of different models' perturbation efficacies with regards to improving human classification accuracy can be found in Fig. C.9 in this Appendix.

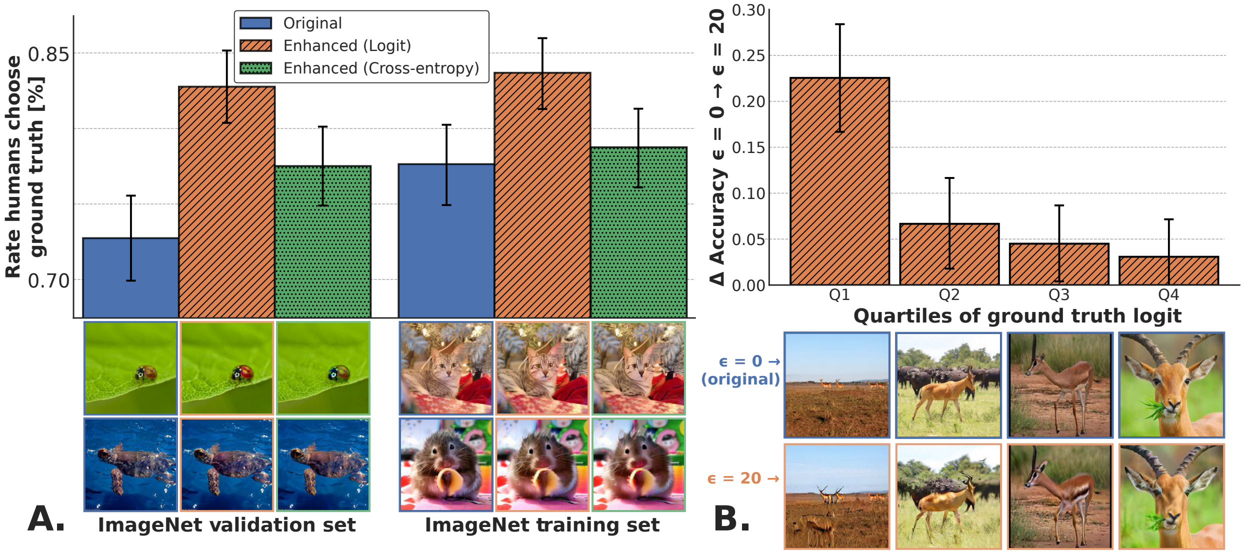


Figure C.11: **Ablation results for image enhancement with ImageNet images.** Logit maximization enhancement is effective for images used to train the robustified CNN used as a guide model, and also for held-out validation images (panel **A**). Logit-max enhancement is more efficient at increasing human 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.

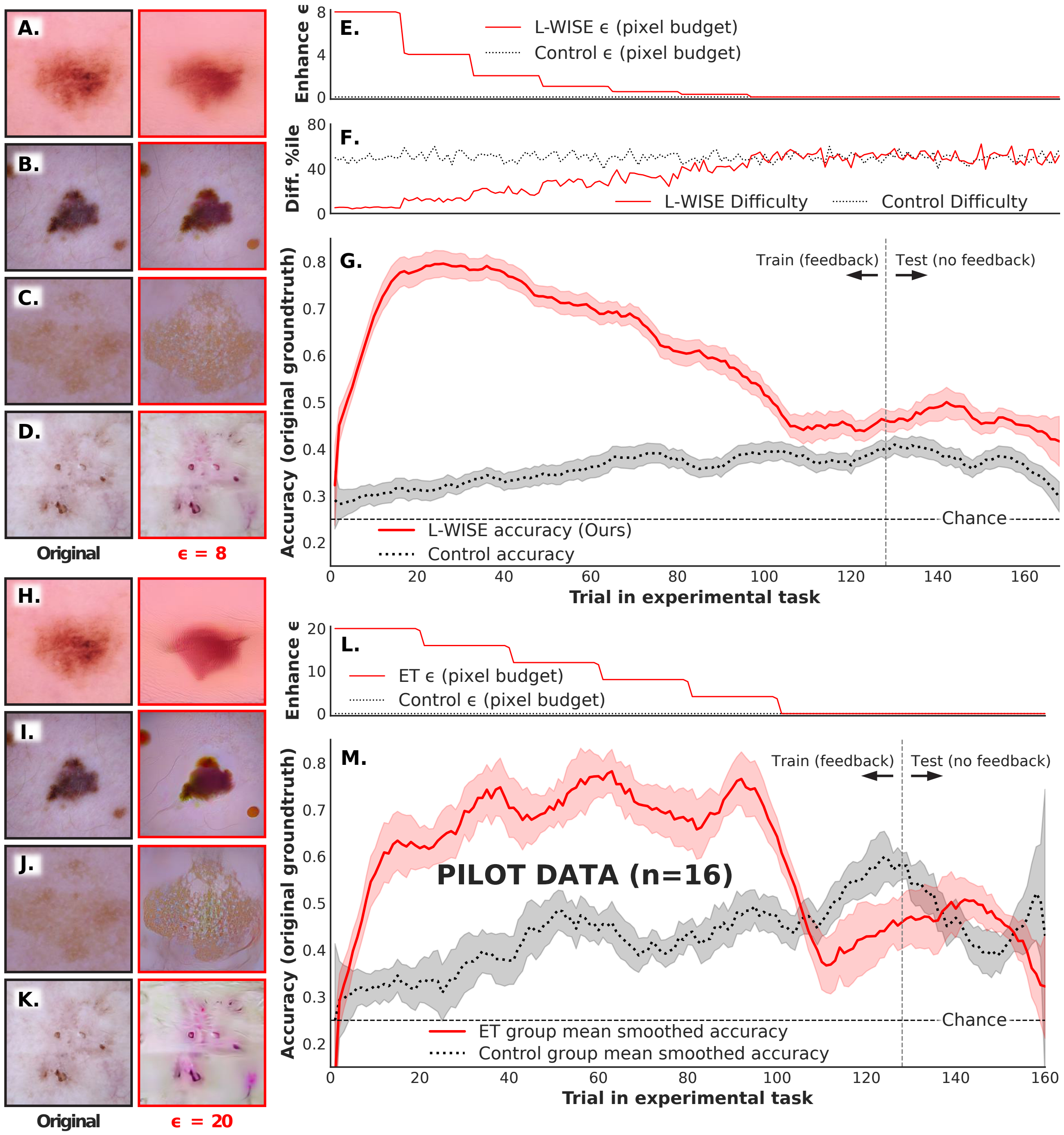


Figure C.12: **Plots showing the accuracy trajectory of human participants throughout training/testing in the main dermoscopy learning experiment (panels A-G) and a preceding pilot experiment after which the epsilon tapering schedule was adjusted (panels H-M).** All conventions are identical to main-text Fig. 3.3. There is a statistically significant difference between the test-phase performance of the L-WISE participants and that of the control participants ($\chi^2(1)$ 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. 3.2F): we suspect that this is the reason for the sudden decline in accuracy in the enhancement group of the pilot experiment (**M**).

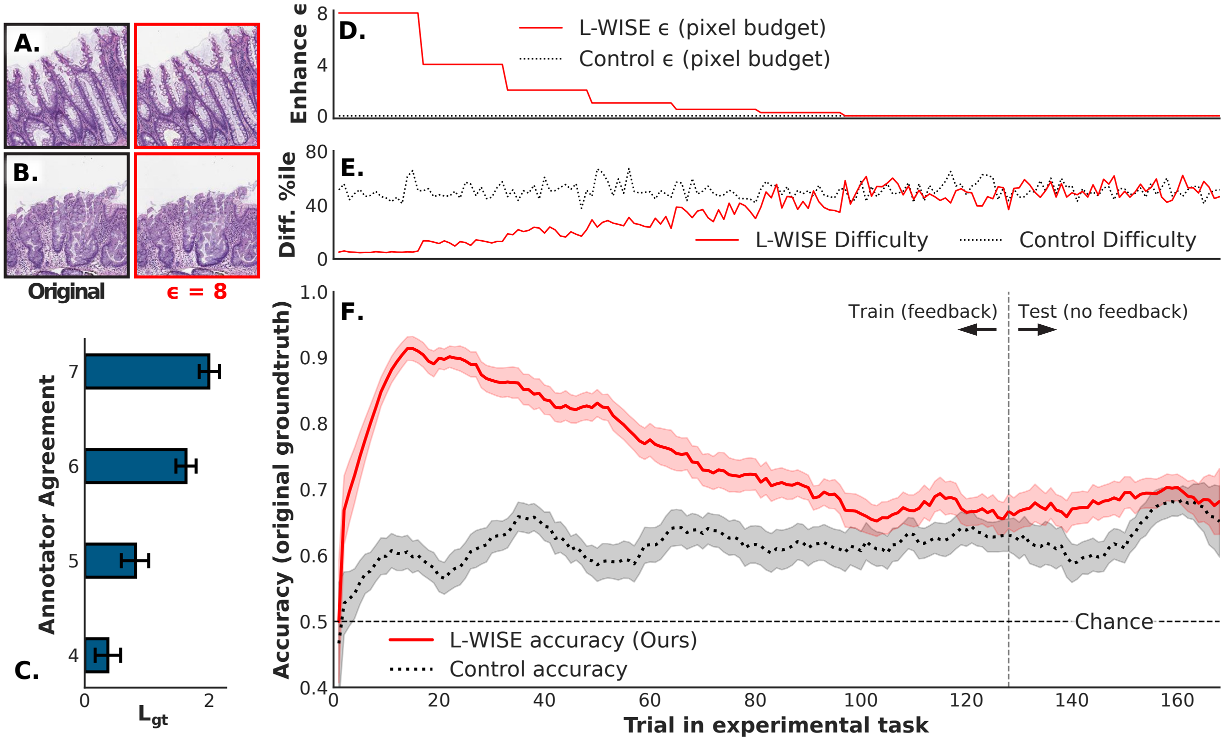


Figure C.13: **Plot showing the accuracy trajectory throughout training/testing of human participants in the histology learning experiment.** All conventions follow Fig. 3.3. Panel **C** shows the association between agreement among the 7 expert pathologist annotators of the MHIST dataset [131] and the 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.

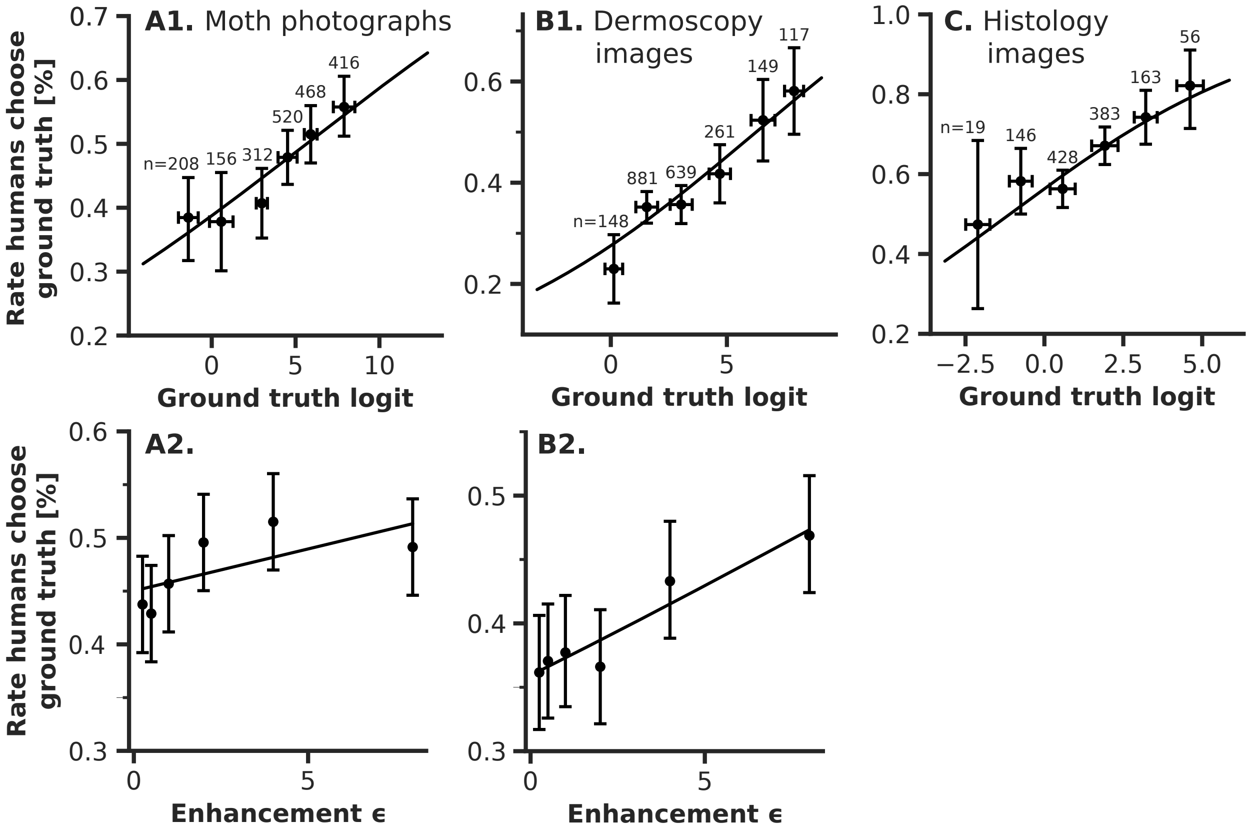


Figure C.14: **Difficulty prediction and image enhancemement are effective across image domains.** Panels **A1**, **B1**, and **C** show the relationship between the ground truth logit from a fine-tuned robustufied ResNet-50 model and the rate at which human participants (from the control groups) choose 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 within each bin. The curved lines illustrate fitted logistic regression models. All logistic regression models had statistically significant coefficients for ground truth logit ($p < 0.001$ from Wald statistic). 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 3.1), the only group that viewed enhanced images without ϵ monotonically decreasing over time. The logistic regression coefficient for ϵ was statistically significant ($p < 0.05$ from Wald statistic) for both moth photographs (**A2**) and dermoscopy images (**B2**).

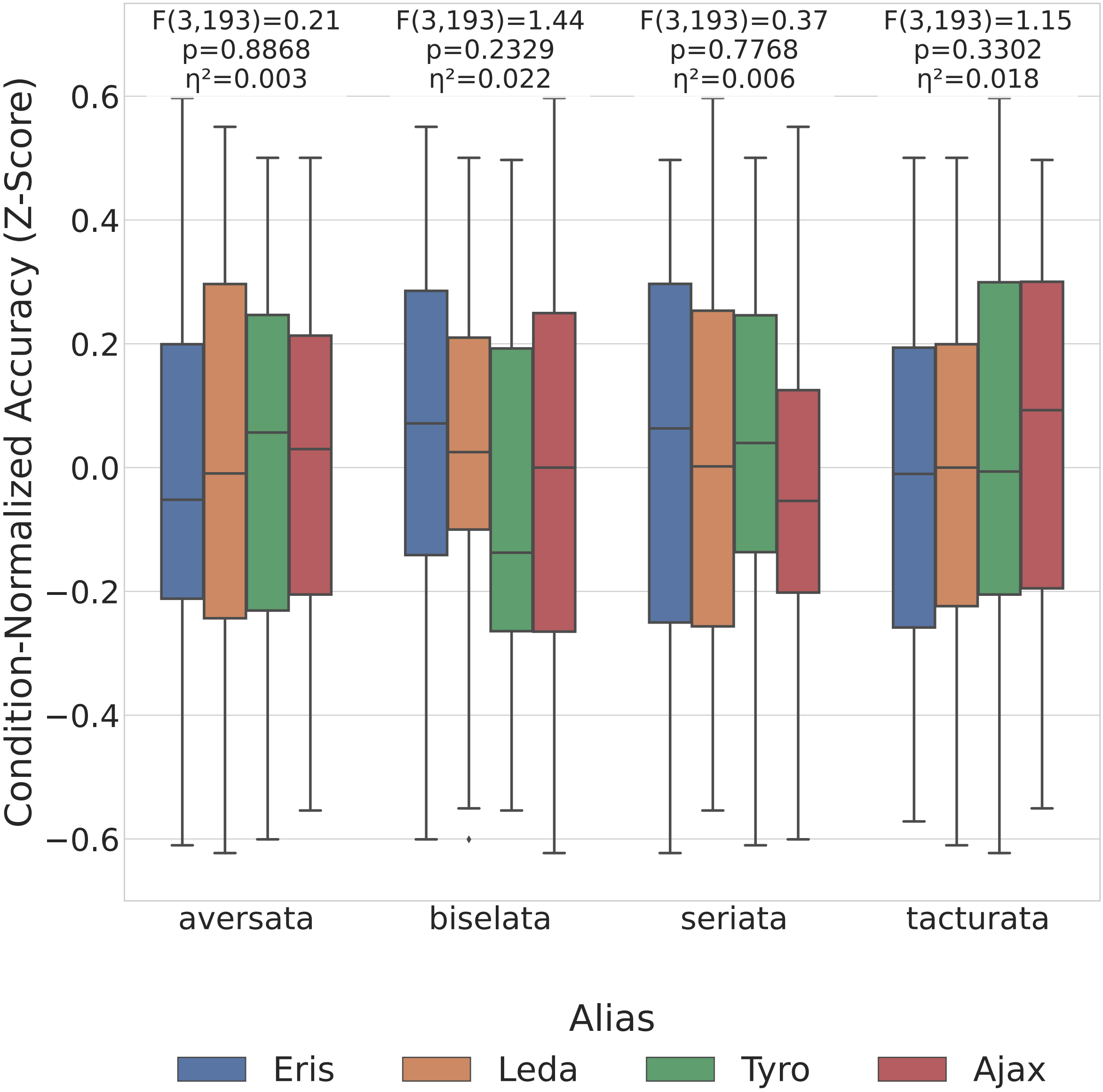


Figure C.15: **Randomized assignment of aliases "Leda," "Ajax," "Eris," and "Tyro" had minimal impact on test-phase accuracy in the moth species classification task.** Each group of four boxplots shows the relative effects of assigning each alias to a specific class from the moth classification experiment. 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 images of the moth species *Idaea tacturata* labelled as "Eris." There is no evidence from one-way ANOVA that the random assignment of aliases to classes influences test-phase performance.

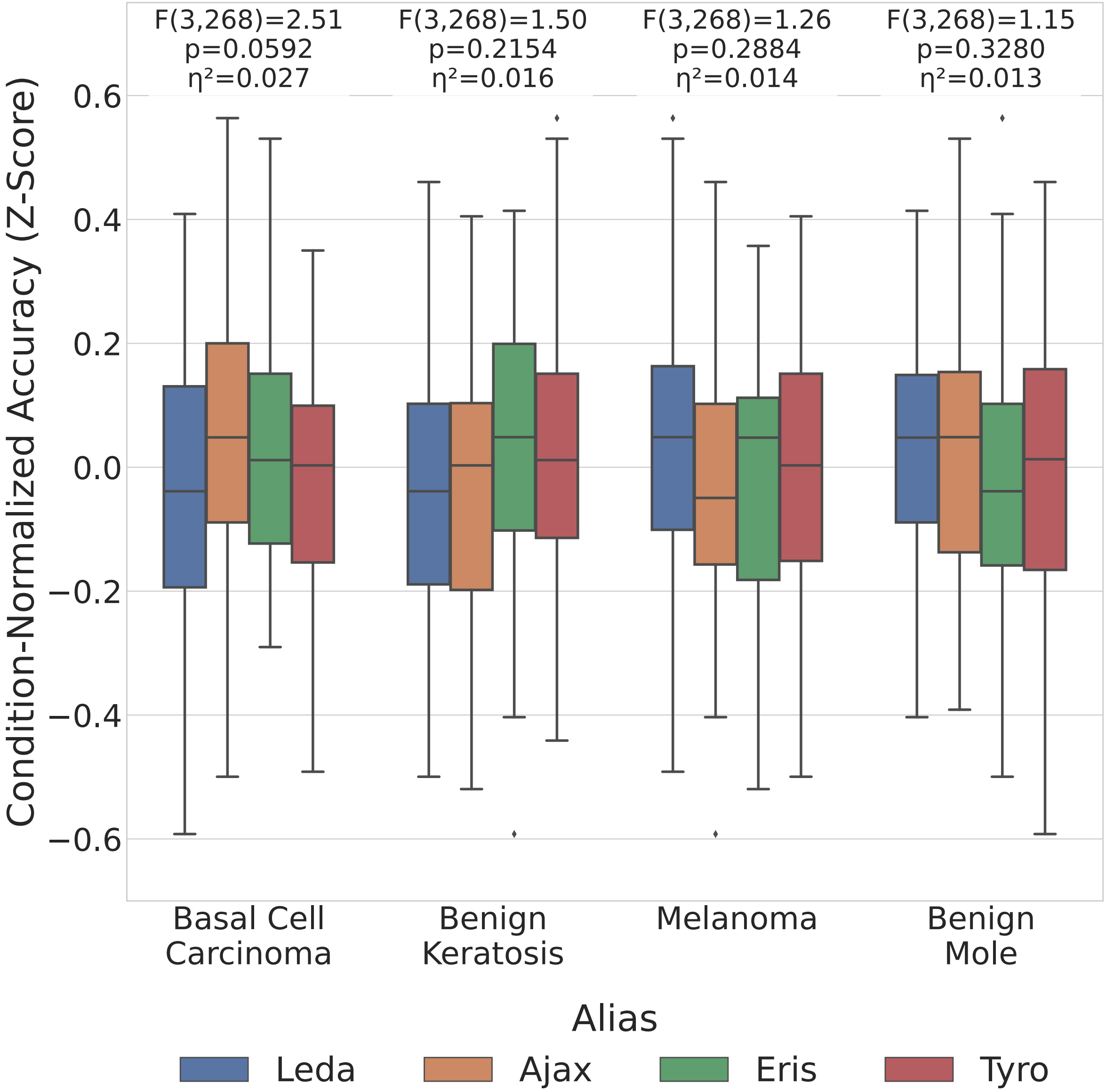


Figure C.16: **Randomized assignment of aliases "Leda," "Ajax," "Eris," and "Tyro" had minimal impact on test-phase accuracy in the dermoscopy task.** After correcting for multiple comparisons, there is no evidence that the random assignment of aliases to classes affects task performance. See also Fig. C.15.

Appendix D

Accelerating Histology Learning with Perceptually Aligned AI: A Randomized Study of First-Year Pathology Residents

D.1 Additional details of longitudinal study design and protocol

A recruitment email linked to an online enrollment form was sent to the Program Directors of all 143 pathology residency programs listed in the American Medical Association's FREIDA Residency Database [208] (as of June 2025), with a request to forward it to eligible first-year residents. All participants provided informed consent, and the study followed a protocol approved by the Institutional Review Board at Boston Children's Hospital.

Participants were assigned to the Intervention or Control arm by alternation based on order of enrollment. Although this allocation approach relies on a deterministic algorithm rather than a random number generator, the unpredictable and asynchronous nature of the online enrollment process ensured that the assignment sequence was effectively random and was concealed from both the investigators and participants. This approach was selected in order to ensure balanced group sizes at baseline, and to facilitate implicit stratification by institution (residents from the same program tended to enroll in clusters following the Program Director's forwarded email). Participants were not informed of the study's hypothesis or that the study included different experimental groups/curriculum formats. Outcome assessment was fully automated, ensuring that all performance metrics were recorded without observer bias.

Each email sent to participants contained a personalized link to a landing page, which contained a link to the histology tasks, contact information for the research team, and a button allowing participants to withdraw from the study. Withdrawn participants were not sent any further communications (other than an email with details of compensation for any completed tasks or sessions), and their data was deleted and not used for analysis.

Participants were able to complete the histology tasks on any laptop or desktop computer with an internet connection, but not on mobile devices or tablets. The tasks were separated into "training" and "testing" phases (Figure 4.1A). Each training phase contained 96 trials with feedback (randomly sampled cases for each participant), and each testing phase contained 48 trials without feedback (identical sequence of cases for all participants, differing between testing phases). All participants completed three task sessions at their convenience within pre-specified time windows (Figure 4.1A). Session 1 became available immediately upon enrollment, and closed 12 hours after the end of the global enrollment period. In Session 1, all participants completed "Training Phase 1" immediately followed by "Test 1" for each of the breast and prostate tasks, with task order counterbalanced among participants and kept consistent throughout the study. Participants also completed a short questionnaire about their experience at the end of this first session. Session 2 was available between 14 and 21 days after the completion of Test 1. In Session 2, all participants completed "Test 2" for both breast and prostate, after which Intervention group participants completed an additional "Training Phase 2" and "Test 2b" for each task. Session 3 was available from 60 to 74 days after Test 1 completion, and included "Test 3" for both tasks as well as a questionnaire pertaining to participants' overall experience in the study. Once the first task for a given session was started, participants were required to complete all tasks in that session within 48 hours. Residents were sent automated email reminders when they became eligible for each session, and if 24 hours had passed since the beginning of a non-completed session. Participants also received a combination of automated emails and batched emails from the study coordinator when they were within a few days of the session expiring. Extensions were granted upon request for participants who missed the eligibility window for a given session; extensions ranged from 1 to 4 days depending on the request, with one outlier at 13 days (Session 2).

Participants were compensated with up to $50 via an e-gift card of their choice: $10 for each of the three experimental sessions, with an added $20 "bonus" for completing all three. To help incentivize recruitment and sustained engagement, two participants with outstanding task performance were offered an opportunity to collaborate on the manuscript. Specifically, these participants were invited to assist with qualitative analyses of model-enhanced images and participant feedback from the questionnaires in the study. Authorship was contingent upon complete fulfillment of ICMJE criteria, including substantial contribution to data interpretation and manuscript drafting/revision. The specific performance metrics used to select these two participants were not disclosed to avoid introducing bias. Selection was based on Z-scored performance metrics normalized by experimental group and task type to ensure equal opportunity regardless of group assignment.

Session 3 also included a task designed to estimate baseline visual category learning ability independent of histology knowledge. Participants learned to categorize photographs of 4 different moth species following [156]. Metrics from the test phase of this task were then used as covariates in a sensitivity analysis (see Section D.8.1).

D.2 Details of artificial neural network model training and evaluation

We adversarially fine-tuned (training pixel budget $\epsilon = 1.0$) ResNet-50 models [85] that had first been adversarially pre-trained on ImageNet ($\epsilon = 3.0$, following [156]), training separate

Dataset/Mag.	No. Classes	N_{train}	Batch Size	LR (Init.)	Val. Acc. (SD)	Balanced Val. Acc. (SD)
DiagSet (Prostate)						
5x	2	32486	64	0.001	89.0 (0.6)	87.1 (1.0)
10x	2	82728	64	0.001	91.1 (0.3)	89.6 (0.4)
20x	2	230240	64	0.0005	92.0 (0.3)	90.5 (1.0)
BRACS (Breast)						
5x	7	1557	16	0.00025	52.4 (2.2)	43.6 (2.4)
10x	7	10190	64	0.0005	54.7 (1.3)	50.7 (1.0)
20x	7	51417	64	0.0005	48.8 (1.6)	46.9 (1.0)

Table D.1: **Robust model hyperparameters and performance summary.** Separate models were trained for each magnification level. "No. Classes" is the total number of classes used to train the models for each dataset. The BRACS models were trained on the original 7 classes, which form the 3 superclasses benign, atypical, and malignant. "N_{train}" is the total number of training images for each magnification level. "LR (Init.)" is the initial learning rate, which is divided by one-half after every 20 training epochs. "Val. Acc." is accuracy on the held-out validation set. "Balanced Val. Acc." is balanced validation set accuracy, defined as the mean of the recall values obtained for each class. Accuracy metrics are reported as mean (SD) among 10 training set splits, but are obtained on the validation set instead of the held-out training data.

models for each magnification level within each dataset. BRACS models were trained using the dataset's original 7 class labels [158], and DiagSet models were trained on binary benign (original class "N") vs. malignant (Gleason grades R3, R4, R5) labels [159]. Model training was not limited to the "teaching set" used for the experiments with human participants (see Section D.3), and instead used the full original datasets. We tuned the learning rate and batch size for each magnification/dataset combination, trained for 60 epochs, and used the training epoch checkpoints with maximum validation set accuracy for all downstream processing. Due to heavy class imbalance in the BRACS dataset (very few images in both "atypical" subcategories, atypical ductal hyperplasia and flat epithelial atypia), the cross-entropy loss for BRACS was weighted inversely proportionally to class frequencies during training. Model hyperparameters and overall performance are summarized in Table D.1, with a breakdown of sensitivity and specificity by class in Table D.2.

The adversarially trained models were used to estimate image difficulty and to produce enhanced images (details below). To avoid performing these operations on images that were also used to train the model, we used a 10-fold cross-validation procedure with no overlap in patient tissue sources between splits. Each image in the training set was later processed by the model corresponding to its held-out split.

D.3 Details of image preprocessing and teaching set curation

We drew H&E-stained histology image patches from publicly available datasets: BRACS [158] for breast tissue and DiagSet [159] for prostate tissue. We standardized all images to 256 × 256 pixels at 0.5 µm/pixel (20x magnification), 1 µm/pixel (10x), and 2 µm/pixel (5x), corresponding to the native resolutions of the DiagSet dataset. The BRACS dataset provides labeled region-of-interest (ROI) images of varying size at 0.25 µm/pixel. We extracted as many non-overlapping square patches (of the corresponding size for each magnification) as possible from each ROI before downsampling to 256 × 256 pixels as needed, applying the overall ROI class label to each patch. The image patches were center-cropped to 224 × 224

	5x Magnification			10x Magnification			20x Magnification		
Dataset and Class	N_{train}	Sens.	Spec.	N_{train}	Sens.	Spec.	N_{train}	Sens.	Spec.
DiagSet (Prostate)									
Benign	16660	93.0 (1.0)	81.1 (2.6)	43379	93.5 (0.5)	85.7 (1.1)	121502	93.9 (0.7)	87.1 (2.5)
Malignant	15826	81.1 (2.6)	93.0 (1.0)	39349	85.7 (1.1)	93.5 (0.5)	108738	87.1 (2.5)	93.9 (0.7)
BRACS (Breast)									
Benign									
Normal	97	58.0 (6.7)	93.3 (2.3)	761	69.5 (3.6)	94.5 (0.9)	4031	75.6 (3.8)	87.7 (1.3)
Pathological Benign	527	43.8 (3.9)	93.5 (1.7)	3129	31.5 (6.3)	93.2 (1.7)	15158	25.9 (3.3)	93.3 (1.3)
Usual Ductal Hyperplasia	50	33.3 (14.2)	89.1 (3.9)	484	40.8 (10.3)	90.7 (2.6)	2758	36.2 (6.7)	88.5 (3.4)
Atypical									
Flat Epithelial Atypia	16	25.8 (18.2)	98.6 (1.8)	308	57.9 (10.1)	95.0 (1.6)	2088	64.0 (4.5)	93.3 (1.3)
Atypical Ductal Hyperplasia	34	2.7 (8.6)	98.0 (1.3)	409	17.6 (5.1)	95.0 (2.3)	2353	9.6 (3.9)	95.4 (1.9)
Malignant									
Ductal Carcinoma In Situ	317	80.6 (8.1)	76.3 (3.9)	2069	70.5 (6.5)	84.7 (2.3)	10523	57.8 (4.2)	88.4 (1.5)
Invasive Carcinoma	516	60.9 (4.2)	97.2 (0.5)	3030	66.9 (2.4)	94.7 (1.1)	14506	59.6 (3.4)	95.2 (0.9)

Table D.2: **Sensitivity and specificity of robust neural networks, by class.** Separate models were trained for each magnification level. "N_{train}" is the total number of training images for each class within each magnification; we report these values here for convenience, as they help explain inter-class differences. "Sens." is sensitivity, and "Spec." is specificity; both are reported as mean (SD) among 10 training set splits (evaluated on validation set).

for artificial neural network (ANN) processing.

We used a multi-step quality control procedure to exclude unsuitable image patches from the "teaching set" of images available for learning in both Intervention and Control groups (Figure 4.1C). To eliminate blurry images, we calculated a focus quality score for each patch using the fast Fourier transform (FFT). Images were converted to grayscale before the FFT operation. To isolate high spatial frequencies associated with edges and fine textures, a high-pass filter was applied to the FFT image by zeroing the central low-frequency components using a square mask (100×100 pixels, on a 256×256 uncropped image). The quality score was defined as the mean magnitude of the remaining high-frequency spectrum. Images with a quality score below 1000 were discarded; this threshold was determined empirically through visual inspection of score distributions (to exclude the lower tail), and of images near the decision boundary.

In addition to blur artifacts, some histology image patches consisted primarily of stroma, glandular lumina, or slide background instead of diagnostically informative epithelial tissue. This occurred because patches from both BRACS and DiagSet were labeled according to their source ROI during a tiling procedure, even for individual patches that lack the features necessary to independently verify that label. To select patches containing breast/prostate epithelium, we trained a linear classifier using UNI2-h embeddings [157] from each patch. Embedding dimensionality was reduced from 1536 to 16 using principal component analysis. We hand-labeled 500 patches from each dataset, balanced among the three magnifications, as containing epithelium or not (defined as epithelium covering $> 10\%$ of the image). Logistic regression classifiers trained on these labeled subsets reached 5-fold cross-validated AUC values of 0.94 (BRACS) and 0.95 (DiagSet), and were used to classify the remaining patches (using a lenient probability threshold of 0.25 to prioritize recall and minimize removal of acceptable images). This procedure eliminated the majority of non-epithelium-containing patches. Figure D.1 shows examples of images that were excluded vs. accepted during both the frequency-based filtering and epithelium detection stages.

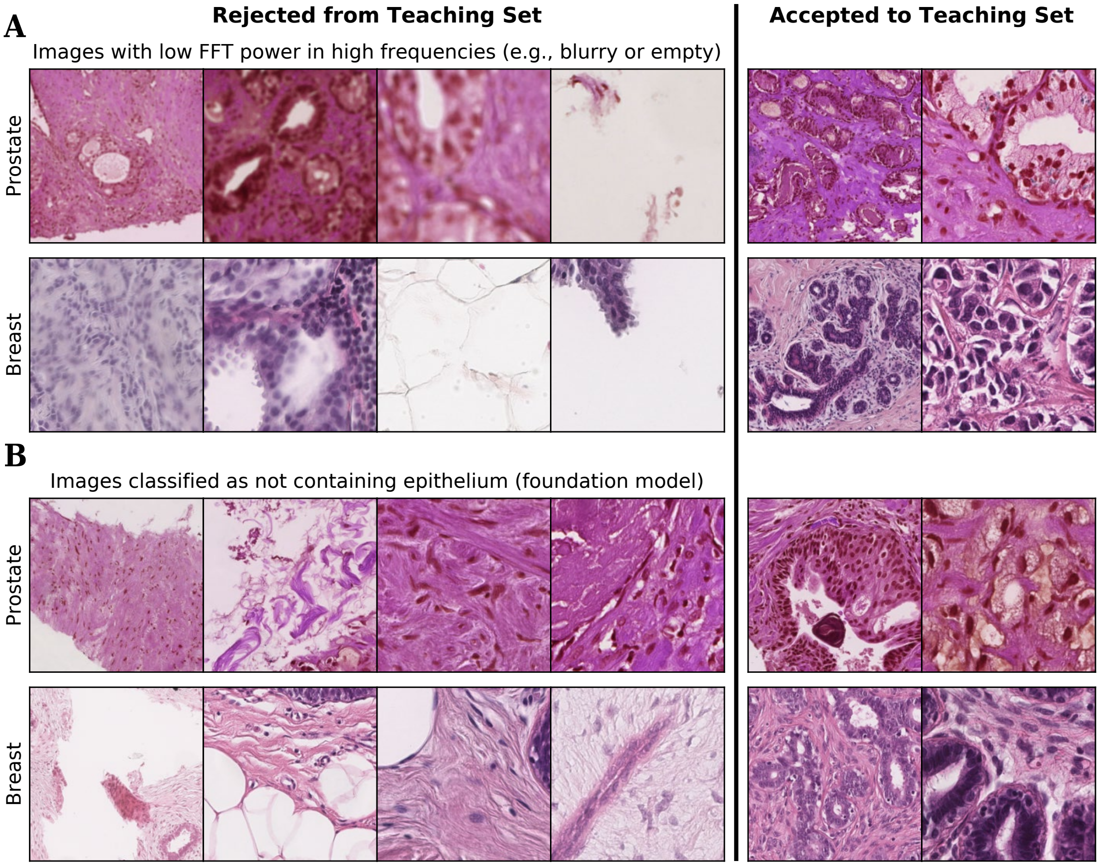

A
Rejected from Teaching Set
Accepted to Teaching Set
Images with low FFT power in high frequencies (e.g., blurry or empty)
Prostate
Breast
B
Images classified as not containing epithelium (foundation model)
Prostate
Breast

Figure D.1: **Quality control pipeline for histology image teaching set curation.** Each row displays examples of prostate (DiagSet) or breast (BRACS) histology image patches that were excluded (left) and accepted (right) during the automated curation of the teaching set. The first stage (panel A) filtered images using fast Fourier transform (FFT) analysis, rejecting patches with mean high-frequency power below an empirical threshold (typically associated with blur or slide background). The second stage (panel B) filtered for image patches that contained epithelial cells using a linear classifier trained on UNI2-h foundation model embeddings. Patches containing primarily stroma, adipose tissue, or other non-epithelial structures (left) were rejected, retaining images with diagnostically relevant features (right).

D.4 Details of model-based difficulty predictions and generation of curriculum sequences

To generate histology curricula with images selected and sequenced according to difficulty, the difficulty of recognizing the correct ground truth class (i.e., benign, malignant, or atypical) was predicted using an ANN trained to distinguish among these classes. Specifically, a ResNet-50 model was first adversarially pretrained on the ImageNet dataset [80] and subsequently adversarially fine-tuned on each histology dataset (BRACS for breast tissue [158], DiagSet for prostate tissue [159]). For a new image not seen during training, the difficulty score was calculated based on the model's logit score (pre-softmax) corresponding to the ground truth class of that image [156].

Separate models were trained for each tissue type (breast and prostate), and for each magnification level (5x, 10x, and 20x). In order to calculate valid logits for every image in the training set without the influence of having trained on any of the images, 11 separate models were trained for each tissue type/magnification combination in a procedure similar to cross-validation. The first of these was trained on the entire training set, and used to generate logit scores for the held-out validation set. The training set was split into 10 cross-validation folds, ensuring that image patches from any given patient tissue source (whole-slide image) were all in the same fold. To obtain logit scores for images in each of these folds, the model trained on the 9 other folds was used, such that no logits were generated for images that were seen by the model during training (similarly, for image enhancement, each image was enhanced by a model that did not see it during training).

To convert raw ground truth logit scores into difficulty percentile scores, the percentile rank of each image's logit was calculated within its class and magnification level. Since higher logit scores correspond to higher model confidence and lower difficulty, the logit percentile was inverted (subtracted from 1) to obtain the difficulty percentile [156].

This procedure allowed direct calculation of the difficulty percentile scores for the DiagSet (prostate) dataset, where models were trained for binary benign/malignant classification. For the BRACS (breast) dataset, models were trained on the original 7 classes [158], which combine to form the superclasses used for the learning tasks: benign (normal, pathological benign, usual ductal hyperplasia), malignant (ductal carcinoma in situ, invasive carcinoma), and atypical (flat epithelial atypia, atypical ductal hyperplasia). A composite superclass logit score was used to calculate the difficulty percentile score for BRACS image patches. The softmax-normalized scores for the 2–3 subclasses in the ground truth superclass were summed and converted back to a logit score by applying the logit function (log-odds). These logits were then converted to percentiles, normalizing among images of the same magnification and subclass (e.g., atypical ductal hyperplasia) to reflect varying difficulty levels within these finer-grained categories.

Curriculum sequences for the longitudinal study training phases were structured as 8 consecutive blocks of 12 class-balanced cases each (96 cases/trials in total), with all image patches drawn from a dedicated training set within each dataset. Test phase sequences consisted of 4 consecutive blocks following the same structure, drawn from a held-out validation set within each dataset to ensure no overlap between training and testing cases.

The histology datasets used in this study contain many image patches extracted via tiling from a smaller number of whole-slide images (patient cases). For example the BRACS training set includes 281 unique cases yielding 51,417 image patches at 20x magnification. To maximize study participants' exposure to a diversity of presentations, case sequences for both training and test phases were generally sampled such that no patient case appeared more than once. However, due to class imbalance in the BRACS dataset (see Table D.2), exceptions were made for the "atypical" class. For this class only, the same patient case was allowed to appear up to 3 times in a training sequence and up to 2 times in a testing sequence, but without repetition of individual image patches.

As described in Section 4.3.3, Intervention group training sequences were sampled using difficulty percentile caps that increased across blocks. Although repeated case sampling effectively increased the pool of available atypical image patches, there were not enough images with sufficiently low difficulty percentiles to be placed in the early blocks (e.g., the first two blocks required images at/below the 10th difficulty percentile). To accommodate this, the percentile cap was increased by 30 percentage points for the atypical class only (e.g., a block normally capped at the 10th percentile was capped at the 40th percentile for atypical cases).

Each participant was assigned to a uniquely sampled sequence of 96 training cases for each training phase. Test phases were identical across participants and groups. Tests 1, 2a, 2b, and 3 contained distinct case sequences. There were not enough image patches available to enforce fully disjoint sets of image patches across all four tests. To create a constant difficulty distribution in expectation throughout the four tests, while avoiding the selection biases inherent in sequential sampling without replacement, each test sequence was constructed as an independent sample from the held-out validation set. Of the 768 total image patches presented across the 4 test phases, 642 (83.6%) were unique for the breast task and 752 (97.9%) were unique for the prostate task. Between Test 2a and Test 2b (the only two tests occurring within the same session), 90.1% of breast patches and 99.5% of prostate patches were unique.

D.5 Details of category-specific image enhancement procedures

The AI-based learning assistance approach in this study employed a model-based, category-specific image enhancement to highlight diagnostically relevant features in histology images during the feedback phase of the task, which takes place after the participant enters their response for each trial. Image x (formulated as a vector in pixel-space) is optimized with a single, high-rate step of gradient ascent, using the gradient with respect to x of the ANN model's logit score corresponding to the image's ground truth label. While the ground truth logit score is maximized, the procedure simultaneously minimizes the logit scores of other competing classes that the model is trained to recognize. The computed gradient step is then projected onto a hypersphere of radius ϵ centered about the origin, effectively constraining the perturbation to have an ℓ_2 norm of at most ϵ. The complete image optimization operation is described by Equation D.1:

$$x^* = x + \text{Proj}_\epsilon\Bigg(\eta \nabla_x \big(L_{\text{gt}}(x) - \frac{1}{|C|-1} \sum_{\substack{c \in C \\ c \neq \text{gt}}} L_c(x)\big)\Bigg) \tag{D.1}$$

where x^* is the final, enhanced image, $\text{Proj}_\epsilon(v)$ projects vector v onto the ℓ_2 ball of radius ϵ centered at the origin, η is the step size (or "learning rate"), $L_{\text{gt}}(x)$ is the logit score produced by the ANN given image x for class gt (x's ground truth class label), and C is the set of all classes that the ANN is trained to recognize. For image enhancement in this study, we set the pixel budget to $\epsilon = 20$ and the learning rate to $\eta = 40$. The approach was further modified for the BRACS breast tissue dataset [158], which includes 2–3 subclasses (e.g., usual ductal hyperplasia, ductal carcinoma in situ) for each superclass used in the learning task ("benign," "malignant," and "atypical"). The ANN models trained on BRACS are trained to recognize each of the 7 original subclasses independently. Rather than minimizing the logits of all competing classes, only the logits for classes that are not part of a given image's superclass were minimized, while maximizing the logit score of the ground truth subclass label (effectively restricting the summation in Equation D.1 to exclude all subclasses belonging to the ground truth's superclass).

Examples of enhanced images, together with "difference images" and heat maps to highlight which image regions are the foci of the perturbations, are shown in Figure 4.4 and Appendix D, Figures D.2- D.3.

D.6 Rationale for single-step optimization strategy

Previous approaches to category-specific image optimization using projected gradient descent/ascent have generally employed multiple optimization steps [101,156]. This iterative refinement of the image perturbation allows for the gradual emergence of complex features and even newly "hallucinated" well-formed objects, to the extent that this technique can also be used to convincingly transform one category into another [101]. However, this poses a potential risk in medical education settings, as trainees are presented with synthetically generated features produced in an unpredictable manner (whereas our goal is to draw attention to diagnostic features that are already present).

We hypothesize that complex hallucinations emerge through a compounding process: initial perturbation steps alter the image features, creating new gradient directions in subsequent steps that depart non-linearly from the original image. In contrast, our proposed approach uses a single, high-rate linear step, relying solely on the gradients of the model with respect to the original, unperturbed image. This restricts the enhancement to coarsely amplifying existing features, rather than iteratively constructing new features. One tradeoff of this approach is that a large linear step seems to produce simpler intensity-based artifacts (such as localized changes in coloration). Selected examples of hallucinated artifacts associated with an iterative optimization approach from prior work [156] are shown in Figure D.4.

D.7 Details of data analysis methodology

Data from the longitudinal study of pathology residents were collected during three sessions (Figure 4.1A). The primary outcome endpoints were (i) accuracy on the test immediately after the initial training session (Test 1), and (ii) accuracy on the delayed test in Session 2 (Test 2a) for both tissue types (breast and prostate). After Test 2a, the Intervention group completed an additional training session followed by Test 2b (neither of which was completed by the Control group). Then, in Session 3, both groups completed Test 3. Secondary outcomes of the study were accuracy on Test 3 (as this is not comparable between groups in the same way as Tests 1 and 2a), response time on all three tests (including only correct responses), and effective training duration in the initial training phase in Session 1. Effective training duration was defined as the cumulative time for which training images were displayed on the screen (i.e., excluding time spent on task instructions, any internet connection-dependent delays between trials), except that any periods longer than 60 seconds spent on the feedback screen after each trial were clipped to 60 seconds. The feedback screen had no time limit, allowing participants to take breaks of unlimited duration: feedback screen durations exceeded 60 seconds in less than 0.5% of all training trials. Analyses included all available data from participants who completed the test sequence for a given session, including cases where one tissue type was completed and not the other (see Figure 4.2 for participant flow and attrition). Longitudinal data were analyzed using likelihood-based mixed models, which yield unbiased estimates under the missing-at-random (MAR) assumption; this allows inclusion of participants with incomplete longitudinal data without requiring imputation of missing values (we also conducted a sensitivity analysis in which multiple imputation was used; please see Section D.8.1). Participants who completed Session 1 but did not complete Session 2 before the deadline were allowed to complete Session 3, including the Session 3 questionnaire and the non-medical visual learning assessment. However, these four participants (two Control, two Intervention) were not included in any analyses of Test 3 accuracy or RT.

Accuracy on Tests 1 and 2a were analyzed jointly, using one logistic generalized linear mixed model (GLMM) per tissue type. Each trial was treated as one data point, with either a correct or incorrect response. Any trials where the participant did not respond within the 15-second time limit were set to "incorrect" for the analysis; depending on the tissue type and time point, this occurred with a frequency between 0.04% and 1.10%. The model's covariates included experimental group (binary), session (binary), an interaction term between group and session, whether the participant had completed a rotation with a focus on the corresponding tissue type, whether the participant was assigned to complete the breast or prostate part of the sessions first (counterbalanced across participants, consistent across within-participant sessions), the class of the test case (treated as categorical, with 3 levels benign, malignant, and atypical for breast), the model-estimated difficulty percentile of the test case (between 0 and 1), and the position of the test case within the standardized test sequence. Test case position (0 to 47, for 48 test trials) was centered at 0. The model also included random intercepts for participants and test cases. Confidence intervals and P values from the fitted models were Bonferroni adjusted for the four primary outcomes (Tests 1 and 2a for breast and prostate). Accuracy on Test 3 was analyzed using a separate GLMM for each tissue type, with the same set of covariates as the models for Tests 1 and 2a. GLMMs were fitted using the BOBYQA optimizer.

Test response time (RT) was log-transformed and analyzed using linear mixed models (LMMs) for each tissue type, with models for Test 3 separate from those for Tests 1 and 2a (mirroring the accuracy analyses). The covariates for the RT LMMs were identical to those for the accuracy GLMMs. Effective training time was log-transformed and analyzed using multiple linear regression, with covariates for experimental group, past rotation experience in the tissue type, and whether breast or prostate was first in the task sequence for that participant. LMMs were fitted using restricted maximum likelihood (REML).

All models converged successfully without warnings (results in Table 4.1). Residual and Q-Q plots for all LMMs (RT) and linear regression models (effective training duration) showed no evidence of heteroscedasticity and approximately normally distributed residuals. Variance inflation factors (VIFs) were calculated for all covariates of all models. A moderate degree of multicollinearity was present between experimental group and session in the models for Test 1 and 2a, for both accuracy and RT. Across these four models, the VIF value was between 1.85 and 1.90 for the experimental group variable, between 1.94 and 2.01 for session, and between 2.80 and 2.87 for the group × session interaction. Across all models, VIF values for other covariates were ≤ 1.09, with the exception of those for the class variable (e.g., malignant, benign), which were ≤ 1.62. Data preprocessing was performed in Python (version 3.12.9), and all GLMM, LMM, and linear regression analyses were completed using R (version 4.4.1) and the *lme4*, *lmerTest*, and *emmeans* packages.

Parameter	**Prostate OR (95% CI)**	**P Value**	**Breast OR (95% CI)**	**P Value**
Fixed Effects				
(Intercept)	12.90 (7.69 to 21.69)	<0.001	3.06 (1.61 to 5.85)	<0.001
Group: Intervention (vs. Control)	1.39 (1.11 to 1.74)	0.004	1.13 (0.91 to 1.40)	0.25
Session: Test 2a (vs. Test 1)	0.81 (0.59 to 1.10)	0.17	0.66 (0.45 to 0.99)	0.04
Interaction: Intervention × Test 2a	0.84 (0.70 to 1.02)	0.08	1.20 (1.00 to 1.43)	0.05
Prior Rotation: Yes (vs. No)	1.75 (1.31 to 2.34)	<0.001	1.81 (1.38 to 2.37)	<0.001
Task Order: Prostate First (vs. Breast)	0.88 (0.72 to 1.08)	0.22	0.99 (0.82 to 1.21)	0.94
Case Sequence Position (Centered)	1.00 (0.99 to 1.01)	0.91	0.99 (0.98 to 1.00)	0.16
Case Difficulty Percentile (0 to 1)	0.04 (0.02 to 0.09)	<0.001	0.05 (0.02 to 0.15)	<0.001
Class: Malignant (vs. Reference Class)	1.32 (0.98 to 1.78)	0.07	4.94 (3.07 to 7.95)	<0.001
Class: Benign (vs. Reference Class)	—	—	1.29 (0.81 to 2.07)	0.28
Random Effects (Variance)	**Variance (Prostate)**		**Variance (Breast)**	
Participant Intercept	0.24	—	0.23	—
Test Case Intercept	0.49	—	0.86	—

Table D.3: **Output parameters from GLMMs used in primary analysis.** Odds ratios (ORs) with 95% confidence intervals (CIs) computed using the profile likelihood method are reported. Prior rotation experience was specific to breast or prostate. The "Reference Class" was "benign" for the prostate task and "atypical" for the breast task.

D.8 Additional results from main resident study

Table D.4 describes the baseline characteristics of the participants in the main study, and institutions with at least one participating resident are listed in Table D.5. Figure 4.2 shows participant flow through the study. Output parameters from the GLMMs used for the primary accuracy analyses are provided in Table D.3, including estimated coefficients for all covariates. The model coefficients and confidence intervals suggest that the probability of a correct response to a given trial was influenced by the model-predicted difficulty of the case and the past rotation experience of the resident.

Confusion matrices from each test phase in the resident study are shown in Figure D.5. Higher prostate task accuracy in the Intervention group vs. Control on Test 1 (Figure 4.1) appears to be driven primarily by higher sensitivity in detecting malignancy; however, the difference conversely appears to be driven by improved specificity on Test 3 (this should be interpreted descriptively, due to marginally overlapping confidence intervals). In the breast task, participants in both groups were significantly more sensitive to the "malignant" class than to "atypical" or "benign" across all test phases. Intervention group participants showed significantly higher breast task accuracy than Control participants on Test 2a (suggesting improved retention following a 2–3 week delay; Table 4.1). The confusion matrices in panels G-H of Figure D.5 suggest that this is primarily driven by better discrimination between the "atypical" and "malignant" classes.

Characteristic	**Total**	**Control**	**Intervention**
No. of participants	147	73	74
Age (years) — Mean ± SD	31.6 ± 4.6	31.6 ± 4.7	31.5 ± 4.5
Gender — No. (%)			
Female	80 (54.4%)	39 (53.4%)	41 (55.4%)
Male	51 (34.7%)	27 (37.0%)	24 (32.4%)
Non-binary	1 (0.7%)	1 (1.4%)	0 (0.0%)
Prefer not to answer	15 (10.2%)	6 (8.2%)	9 (12.2%)
Ethnicity — No. (%)			
Hispanic or Latino	12 (8.2%)	8 (11.0%)	4 (5.4%)
Not Hispanic or Latino	104 (70.7%)	48 (65.8%)	56 (75.7%)
Prefer not to answer	31 (21.1%)	17 (23.3%)	14 (18.9%)
Race — No. (%)			
Asian	46 (31.3%)	23 (31.5%)	23 (31.1%)
Black or African American	10 (6.8%)	8 (11.0%)	2 (2.7%)
Other	12 (8.2%)	6 (8.2%)	6 (8.1%)
White	60 (40.8%)	28 (38.4%)	32 (43.2%)
Prefer not to answer	19 (12.9%)	8 (11.0%)	11 (14.9%)
Residency Type — No. (%)			
AP	3 (2.0%)	2 (2.7%)	1 (1.4%)
AP/CP	136 (92.5%)	66 (90.4%)	70 (94.6%)
AP/NP	8 (5.4%)	5 (6.8%)	3 (4.1%)
Prior Experience — No. (%)			
Completed breast pathology rotation	21 (14.3%)	13 (17.8%)	8 (10.8%)
Completed prostate pathology rotation	20 (13.6%)	12 (16.4%)	8 (10.8%)

Table D.4: **Baseline characteristics of participants (pathology residents in main study).** Data are shown for all enrolled participants (n=147). The response options for the race question included "American Indian or Alaska Native" and "Native Hawaiian or Other Pacific Islander," and those for the gender question included "Prefer to self-describe," but no participants selected these categories. Participants were considered to have completed a rotation in breast or prostate pathology if they either (a) self-reported a rotation in their current program or (b) self-reported a prior completed pathology residency (data available for n=106 participants). One participant with a prior post-sophomore fellowship was considered to have completed a rotation for both tissue types, and one participant with 5 years of breast pathology research experience was considered to have completed a breast pathology rotation. AP=anatomic pathology, CP=clinical pathology, NP=neuropathology.

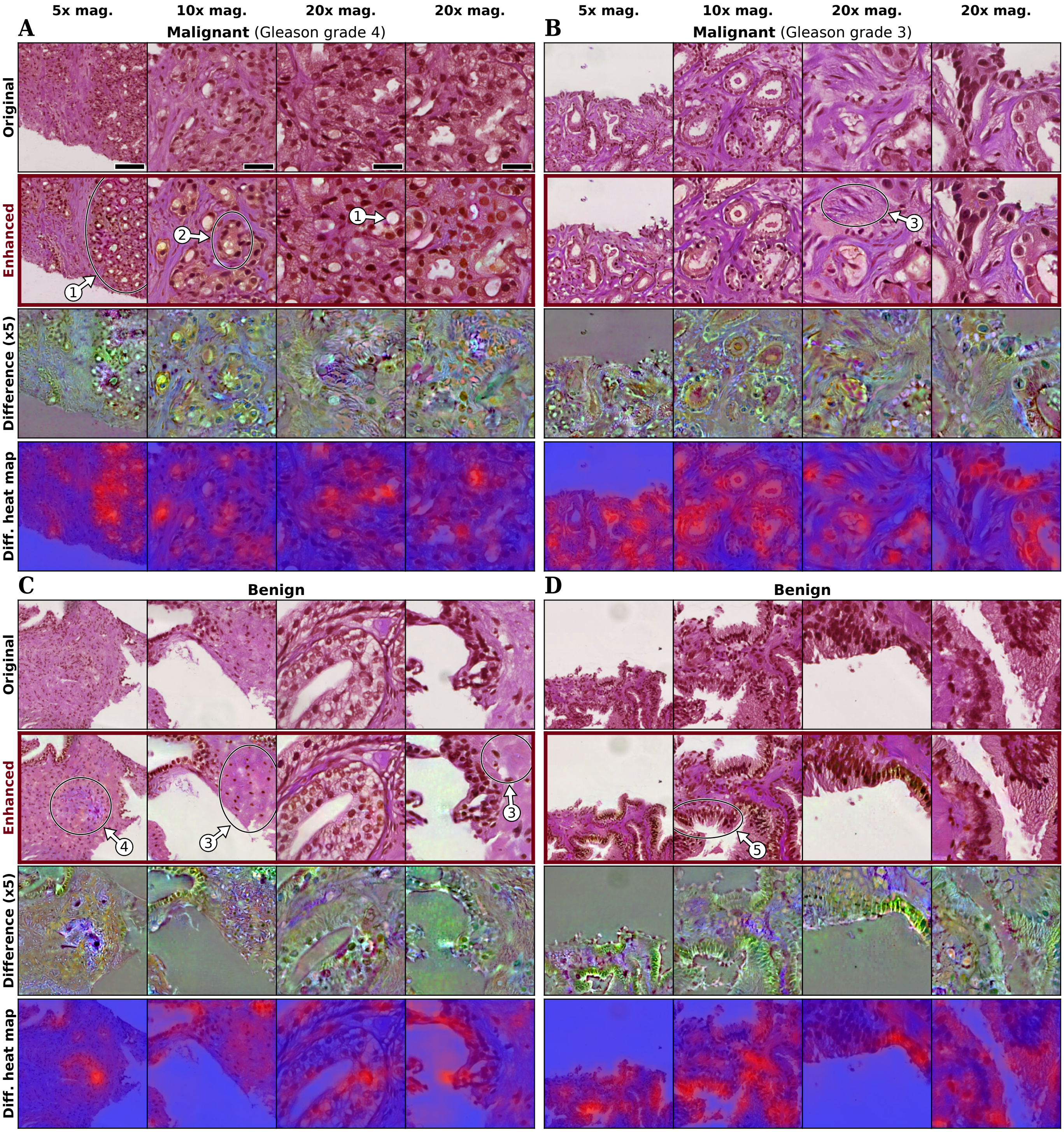


Figure D.2: **Examples of category-specific histological feature highlighting in prostate tissue images.** Arrows indicate examples of biologically relevant enhancements. 1: Accentuated lumina indicative of prostate malignancy. 2: Highlighted cytoplasm of epithelial cells, making them more distinguishable from the surrounding stroma. 3: Accentuated differences in the appearance of stromal collagen fibers: malignant stroma has a more fibrillar appearance with thicker, less aligned, and less densely arranged collagen fibers [163], while enhanced benign stroma appears more dense and hyalinized. 4: An apparent tissue tearing artifact, which might otherwise be mistaken for lumina indicative of malignancy, has been "repaired" during an enhancement of normal tissue, and possible vacuoles have been obscured. 5: Accentuated double-layered epithelium (normal). Scale bars are 100 μm (5x), 50 μm (10x), and 25 μm (20x).

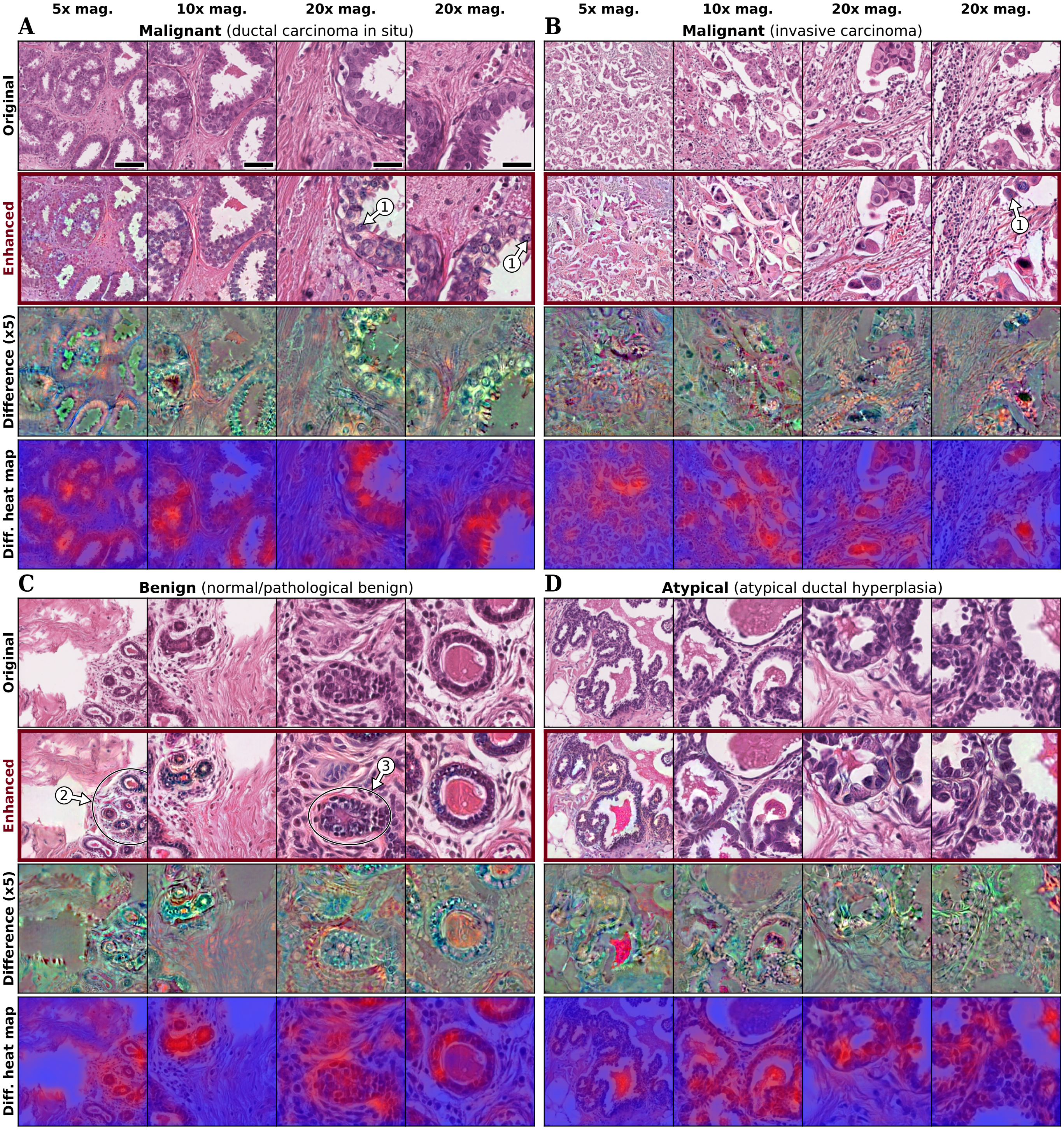

5x mag.
10x mag.
20x mag.
20x mag.
A
Malignant (ductal carcinoma in situ)
5x mag.
10x mag.
20x mag.
20x mag.
B
Malignant (invasive carcinoma)
Original
Enhanced
Difference (x5)
Diff. heat map
1
1
1
C
Benign (normal/pathological benign)
D
Atypical (atypical ductal hyperplasia)
2
3

Figure D.3: **Examples of category-specific histological feature highlighting in breast tissue images.** Arrows show examples of biologically relevant enhancements. 1: Contrast increased for prominent nucleoli. 2: Normal glandular architecture is accentuated at low magnification. 3: A lumen associated with normal glandular architecture is made more apparent by enhancement, clarifying what might otherwise resemble a solid pattern of carcinoma in situ. Scale bars are 100 μm (5x), 50 μm (10x), and 25 μm (20x).

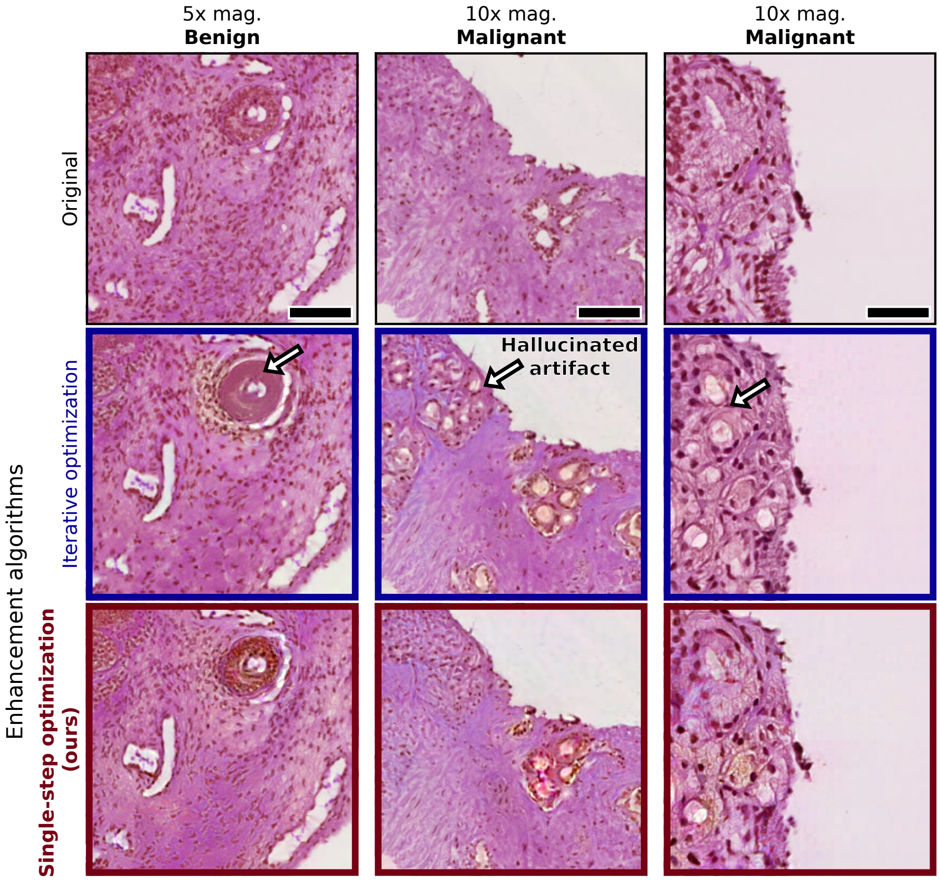


Figure D.4: **Single-step optimization mitigates hallucinatory artifacts.** The top row shows original H&E image patches from the prostate (DiagSet) dataset. The middle row shows images enhanced using iterative projected gradient ascent (40 steps, step size 0.5, pixel budget $\epsilon = 20$). Arrows indicate regions with hallucinated artifacts, such as the appearance of a structure resembling corpora amylacea replacing epithelial cells (left), and artificial formation or structural alteration of glandular lumina (middle, right). For an example of a real corpus amylaceum, please see Figure D.1 (second row from bottom, second image from right). The bottom row shows images enhanced using the proposed single-step gradient ascent (step size = 40, $\epsilon = 20$). Diagnostic features appear to be accentuated or highlighted without the appearance of spurious structures. Scale bars are 100 μm (5x) and 50 μm (10x).

State	Program Name (ACGME)
Alabama	University of Alabama Hospital (Birmingham) Program
Arkansas	University of Arkansas for Medical Sciences (UAMS) College of Medicine Program
California	Los Angeles County-Harbor-UCLA Medical Center Program
California	Stanford Health Care-Sponsored Stanford University Program
California	UCLA David Geffen School of Medicine/UCLA Medical Center Program
California	University of California (Irvine) Program
California	University of California (San Francisco) Program
California	University of California Davis Health Program
Florida	HCA Florida Healthcare/Westside/Northwest Hospital Program
Florida	Mayo Clinic College of Medicine and Science Program
Florida	University of Florida College of Medicine Jacksonville Program
Florida	University of Florida Program
Florida	University of Miami/Jackson Health System Program
Georgia	Emory University School of Medicine Program
Georgia	Medical College of Georgia Program
Illinois	Loyola University Medical Center Program
Illinois	McGaw Medical Center of Northwestern University Program
Illinois	University of Chicago Program
Illinois	University of Illinois College of Medicine at Chicago Program
Maryland	Johns Hopkins University Program
Maryland	University of Maryland Program
Massachusetts	Beth Israel Deaconess Medical Center/Harvard Medical School Program
Massachusetts	Mass General Brigham Program
Massachusetts	Tufts Medical Center Program
Massachusetts	UMass Chan - Baystate Program
Michigan	Henry Ford Health/Henry Ford Hospital Program
Minnesota	Mayo Clinic College of Medicine and Science (Rochester) Program
Mississippi	University of Mississippi Medical Center Program
Missouri	University of Missouri-Columbia Program
Missouri	University of Missouri-Kansas City School of Medicine Program
Missouri	Washington University/B-JH/SLCH Consortium Program
New Hampshire	Dartmouth-Hitchcock/Mary Hitchcock Memorial Hospital Program
New Jersey	Rutgers Health/Cooperman Barnabas Medical Center Program
New Jersey	Rutgers Health/New Jersey Medical School Program
New Mexico	University of New Mexico School of Medicine Program
New York	Icahn School of Medicine at Mount Sinai/Morningside/West Program
New York	Icahn School of Medicine at Mount Sinai/Mount Sinai Hospital Program
New York	Montefiore Medical Center/Albert Einstein College of Medicine Program
New York	New York Presbyterian Hospital (Columbia Campus) Program
New York	SUNY Downstate Health Sciences University Program
New York	University at Buffalo Program
New York	University of Rochester Medical Center Program
New York	Zucker School of Medicine at Hofstra/Northwell Program
North Carolina	ECU Health Medical Center/East Carolina University Program
North Carolina	University of North Carolina Hospitals Program
North Carolina	Wake Forest University Baptist Medical Center Program
Ohio	Ohio State University Hospital Program
Ohio	The MetroHealth System/Case Western Reserve University Program
Oregon	Oregon Health & Science University (OHSU Health) Program
Pennsylvania	Penn State Milton S Hershey Medical Center Program

Continued on next page

Table D.5 – continued from previous page

State	Program Name (ACGME)
Pennsylvania	Sidney Kimmel Medical College at Thomas Jefferson University/TJUH Program
Pennsylvania	University of Pennsylvania Health System Program
Rhode Island	Rhode Island Hospital/Brown University Health Program
Tennessee	University of Tennessee College of Medicine Program
Tennessee	University of Tennessee Medical Center at Knoxville Program
Texas	Baylor Scott & White Medical Center - Baylor College of Medicine (Temple) Program
Texas	University of Texas Health Science Center San Antonio Joe and Teresa Lozano Long School of Medicine Program
Texas	University of Texas Health Science Center at Houston Program
Texas	University of Texas Southwestern Medical Center Program
Utah	University of Utah Health Program
Wisconsin	Medical College of Wisconsin Affiliated Hospitals Program
Wisconsin	University of Wisconsin Hospitals and Clinics Program

Table D.5: **List of 62 pathology residency programs with at least one resident enrolled in the study.** Residency program names are listed as they appear in the Accreditation Council for Graduate Medical Education (ACGME) Program Search database (accessed December 2025). Three participants did not disclose their residency program. The median number of participants per institution was 2 (range: 1-6). Per-institution counts are not disclosed to prevent participant reidentification.

D.8.1 Sensitivity analysis using multiple imputation by chained equations

Participant attrition was observed at multiple stages of the longitudinal study (see Figure 4.2). The resulting data missingness was handled using the likelihood-based approach implicit within GLMMs and LMMs, which yields unbiased estimates under the missing-at-random assumption without requiring imputed values. These model-based analyses were constrained to include only participants who had completed Test 1 (see Figure 4.1) for at least one tissue type. Among the seven participants who started the task but did not complete Test 1, likelihood of dropout could conceivably have been influenced by experimental group assignment, performance during the early part of the task, and/or baseline characteristics, thereby potentially introducing attrition bias.

A sensitivity analysis using multiple imputation by chained equations (MICE) was conducted in order to (i) safeguard against possible attrition bias, (ii) assess the robustness of the study's findings to different methods of handling data missingness, and (iii) account for additional covariates that were only observed for a subset of participants. The main drawback of this approach was reliance on statistics aggregated for each participant (e.g., overall Test 1 accuracy), as it is not feasible to multiply-impute trial-by-trial response data due to high dimensionality. Multiple imputation was performed for all participants who initiated the first experimental session (n=128).

The imputation model included the following variables: age; gender (binary values were imputed due to insufficient sample size for non-binary classification); completion of

Outcome and Time Point	Total N	Main Analysis		Sensitivity Analysis (MICE; N=128)	
		Estimate (95% CI)*	**P Value**	**Estimate (95% CI)**	**P Value**
PROSTATE TASK (DiagSet)					
Test Accuracy		**Odds Ratio**		**Diff. in Means (%)**	
Test 1 (Immediate)	119	1.39 (1.05 to 1.85)	**0.01**	5.34 (1.62 to 9.06)	**0.005**
Test 2a (2 weeks)	106	1.17 (0.88 to 1.56)	0.63	2.76 (-1.92 to 7.44)	0.25
Test 3 (2 months)	103	1.31 (1.03 to 1.67)	**0.03**	4.07 (-0.45 to 8.60)	0.08
Test Response Time		**Geo. Mean Ratio**		**Geo. Mean Ratio**	
Test 1 (Immediate)	119	0.76 (0.66 to 0.88)	**<0.001**	0.76 (0.66 to 0.88)	**<0.001**
Test 2a (2 weeks)	106	0.95 (0.82 to 1.10)	0.48	0.97 (0.83 to 1.14)	0.73
Test 3 (2 months)	103	0.92 (0.78 to 1.10)	0.39	0.97 (0.83 to 1.14)	0.74
Effective Training Duration		**Geo. Mean Ratio**		**Geo. Mean Ratio**	
Training Phase 1	119	0.79 (0.68 to 0.91)	**0.001**	0.80 (0.70 to 0.92)	**0.002**
BREAST TASK (BRACS)					
Test Accuracy		**Odds Ratio**		**Diff. in Means (%)**	
Test 1 (Immediate)	119	1.13 (0.86 to 1.49)	1.00	2.40 (-1.40 to 6.20)	0.21
Test 2a (2 weeks)	106	1.36 (1.03 to 1.79)	**0.02**	5.03 (0.04 to 10.01)	**0.05**
Test 3 (2 months)	104	1.14 (0.92 to 1.42)	0.22	1.11 (-3.28 to 5.49)	0.62
Test Response Time		**Geo. Mean Ratio**		**Geo. Mean Ratio**	
Test 1 (Immediate)	119	0.95 (0.83 to 1.10)	0.51	0.97 (0.84 to 1.12)	0.69
Test 2a (2 weeks)	106	1.00 (0.86 to 1.15)	0.95	1.01 (0.87 to 1.17)	0.94
Test 3 (2 months)	104	0.95 (0.82 to 1.10)	0.50	0.97 (0.84 to 1.11)	0.64
Effective Training Duration		**Geo. Mean Ratio**		**Geo. Mean Ratio**	
Training Phase 1	119	0.94 (0.82 to 1.07)	0.36	0.96 (0.84 to 1.09)	0.50

Table D.6: **Results of sensitivity analysis via multiple imputation are largely consistent with the main analysis**. While the main analysis included only participants who completed Test 1 for at least one tissue type (breast or prostate) and used a likelihood-based approach to handle missingness due to attrition, this sensitivity analysis included all participants who started the first session (N=128), and handled missingness using multiple imputation by chained equations (MICE). Values in the "Main Analysis" columns are repeated from Table 4.1. Total N reflects the number of participants included in the analysis across both experimental groups. Odds ratios for the primary analysis are from generalized linear mixed models with logistic link functions, while mean differences ("Diff. in Means") are from linear regression coefficients. Geometric mean ratios ("Geo. Mean Ratio") are from linear models on log-transformed time values. Primary analysis has 98.75% CIs and Bonferroni-adjusted P values for the 4 primary accuracy endpoints, and 95% CIs with unadjusted P values for all others. Sensitivity analysis has all 95% CIs and unadjusted P values. P values less than 0.05 are in bold.

prior rotations in breast and/or prostate pathology; prior pathology residency abroad; experimental group assignment; primary device used (mouse vs. trackpad); and tissue type order (participants were assigned in a counterbalanced manner to complete either the breast or prostate sequence first in all sessions). Performance-based variables included: accuracy and mean log-transformed RT (correct trials only) on Tests 1, 2a, and 3 for both breast and prostate tasks; log-transformed effective training time in Session 1; delays (in days) between Test 1 and subsequent tests; and accuracy, mean log-transformed RT, and effective training time from a non-medical visual learning assessment (see Appendix D, Section D.1). The imputation procedure was restricted to not use experimental group assignment to predict pre-randomization characteristics (e.g., age, gender, prior residency/rotations, primary device expected to be used). Between-session delay variables were not predicted across tissue types to avoid multicollinearity issues (sessions completed in one sitting cause the delays between Test 1 and Test 2a/3 to be highly similar across breast vs. prostate).

MICE produced 100 imputed versions of the dataset using predictive mean matching. Each of the 100 versions was analyzed using multiple linear regression, and the results aggregated using Rubin's rules [162]. A separate model was fitted for each combination of tissue type, metric, and time point (e.g., breast, accuracy, Test 2a). RT and effective training duration outcomes were log-transformed before modeling. Covariates included in all regression models were age, gender, tissue type order, prior expertise level. Prior expertise level was a 3-level categorical variable, set to "resident" for participants with a prior residency involving anatomic pathology, "rotator" for participants who had completed a formal rotation focusing on the tissue type being analyzed, or "novice" for participants with neither a prior rotation nor a prior residency. Models for outcomes at Sessions 2 and 3 included days elapsed since Session 1 as a covariate. Accuracy outcome models included accuracy from the non-medical visual learning assessment as an additional covariate, while RT and effective training duration models included log-transformed mean RT and effective training duration from that assessment, respectively. RT and effective training duration models also included a covariate for primary device used.

The combined results of the sensitivity analyses are shown in Table D.6 alongside those from the main analysis. The sensitivity analysis results corroborate most of the study's main findings: the Intervention group significantly outperformed the Control group in the initial session for the prostate task (across test accuracy, RT, and effective training duration), and in the delayed Test 2a for the breast task in terms of accuracy. However, while Intervention had significantly higher accuracy than Control on prostate Test 3 in the GLMM analysis ($P = 0.03$), the sensitivity analysis shows a directionally consistent but non-significant difference ($P = 0.08$). This sensitivity analysis may have had lower statistical power than the GLMM/LMM analyses, due to both aggregation of accuracy/RT by participant and additional variance caused by imputing outcome measures for the seven participants who provided zero performance data.

D.9 Questionnaire results

Participants in the longitudinal study completed questionnaires immediately following Sessions 1 and 3 (see Figure 4.1A).

D.9.1 Session 1 questionnaire results

Session 1 involved a training phase with immediate feedback, followed by a test phase graded on accuracy. This sequence was repeated for two tissue types, breast and prostate (order counterbalanced among participants). Immediately following Session 1, participants completed a questionnaire regarding user experience, relevance to training, and perceptions of the AI-enhanced ("exaggerated") images.

Full results from Likert scale/multiple choice questions in Session 1 are provided in Figure D.6. A majority of participants agreed or strongly agreed that the instructions were clear (86%), the interface was intuitive (94%), the images were of adequate quality (62%), viewing four images at different magnifications was helpful (71%), the task was of appropriate duration (88% "just right") and appropriately challenging (74% "just right," 26% "too difficult," 0% "too easy"), and the content was relevant to their training (87%). While 94% found the tasks at least moderately engaging, an exploratory comparison tentatively suggests that the Control group experienced a greater degree of self-reported mental fatigue by the end of the session ($P = 0.049$, uncorrected). Participants reported modest perceived improvements during the short session (45% agreed or strongly agreed that improvement took place), with

81% agreeing or strongly agreeing that they relied primarily on existing knowledge and pattern recognition skills. Intervention group participants were split regarding the usefulness of the enhanced images; 46% agreed or strongly agreed that enhanced images contributed to their learning, while 30% disagreed or strongly disagreed. When participants were asked the extent to which they trusted the image enhancements to be accurate, the modal response was "neutral/unsure" (49%), with 26% indicating trust and 25% indicating skepticism or distrust.

In free-text responses to "Do you have any other comments, or suggestions for improving this learning tool?" in Session 1 (n=73 responses), the most prominent theme was a desire for explicit explanations or markers during the feedback phase following each trial. Keywords "explain" or "explanation" appeared in 11/73 (15%) of responses, e.g., "explanations to why the answers were wrong with appropriate labelling would greatly help" (Control group participant). Fourteen additional responses expressed similar feedback, requesting labels, arrows, or text explanations of distinguishing histological features, with one participant concerned that "[without explanations] I was the blind leading the blind and may have taught myself incorrect patterns" (Control). Participants in the Intervention group suggested clarifying the highlighting patterns in "exaggerated" images, e.g., "was curious what the colors indicated in the exaggerated view"; "I am not sure what the exaggerated images are highlighting consistently. It might be useful to pick certain diagnostically relevant entities to highlight rather than just enhance arbitrarily." A few comments requested a clearer definition of the relatively challenging "atypical" category in the breast task, and others proposed various quality-of-life enhancements such as keyboard-based diagnosis selection (instead of mouse/trackpad), end-of-session test reports, and improved image quality.

Participants were also asked: "Each trial featured one 5x, one 10x, and two 20x images. Did you use all of these magnifications equally to make decisions, or were some more useful than others?" Overall, participants predominantly favored low, 5x power (35.6%), and high, 20x power (20.8%) magnifications over medium, 10x power (8.5%) alone, although many felt they were each equally important (32.1%). Generally, participants preferred using low power to evaluate architecture and high power to observe nuclear features and finer details: "I used the 5x to help with overall architecture and the 20x to help with looking at cell morphology" (Control). Given that Session 1 occurred within roughly the first month of residency training, inexperience often led to uncertainty regarding the most useful magnification. For example, "without yet having had a dedicated rotation to these malignancy types, I felt all magnifications were equally helpful" (Intervention).

D.9.2 Session 3 questionnaire results

Session 3 was a delayed retention assessment without feedback ("Test 3" in Figure 4.1). Because of the roughly 2-month gap between Session 3 and earlier sessions including training with feedback, we note that some responses may reflect participants' experience with the testing format rather than the learning tool itself (highlighted by comments such as "I would want there to be more training lessons vs. just tests" and "lack of feedback is frustrating"). Additionally, it is conceivable that participant attrition between Session 1 (N=117) and Session 3 (N=106) questionnaires introduced non-response biases.

Multiple choice responses in Session 3 (Figure D.7) were broadly consistent with Session 1 regarding perceived usefulness and relevance. No questionnaire respondents reported "major technical issues that likely affected my performance," and 5% (5/106) reported minor issues. When participants were asked how confident they felt in distinguishing between benign, malignant, and atypical findings compared to before the study began, 42% responded "much more confident" or "slightly more confident", with 48% "about the same" and 9% slightly or much less confident. When asked whether anatomic pathology education would be improved by incorporating learning tools like the one in the study, 65% of participants agreed or strongly agreed. When participants were asked how likely they would be to recommend a learning tool like the one in the study to other pathology residents (scale from "0 (not at all likely)" to "10 (extremely likely)"), 67% responded above "5 (neutral)" (mean 6.3, standard deviation 2.5).

Although the study's enrollment form asked participants about prior breast/prostate rotations, it did not capture prior residency training. We subsequently included additional questions in the Session 3 questionnaire asking whether participants had completed a prior residency abroad; 15/106 participants responded affirmatively (10 in Control and 5 in Intervention; see Figure D.7).

Eleven participants provided free-text elaborations regarding enhanced images. Three comments explicitly stated that the enhancements were helpful, with one noting "it helped highlight subtle or easy to miss findings." Two additional comments seemed to suggest that the enhancements may have unconsciously "helped without my awareness somehow"; "It was interesting to see how the exaggeration can change the vibe of the image, but I learn the best when I can verbalize my reasoning and give concrete examples." However, 6 comments expressed skepticism or uncertainty about the diagnostic utility of the enhancements, e.g., "It made the cells look smoother and a bit more colorful, but I didn't understand what it was trying to highlight (e.g. nucleoli, myoepithelial cells, monotony, architecture, etc?). At times, it dulled the nuclear features and made cytology more difficult to assess." Three participants noted color changes that resulted from enhancement (subjectively, many of the enhanced images appear to include localized and/or widespread shifts in coloration; see Figure 4.4 and Appendix D, Figures D.2-D.4 for examples).

When asked whether certain magnifications (5x, 10x, 20x) were more useful than others for making decisions (free-response), responses were largely consistent with Session 1 and multiple-choice responses. Many participants cited low magnification for global architecture (5x mentioned positively in 40/106), and higher magnification for cellular details (20x mentioned positively in 20/106); architectural framing language emerged in 19/106.

When asked "how does the learning tool in this study compare to other learning methods you have used (e.g., lectures, slide seminars, studying textbooks)?", participants' responses (n=106) were mixed. One representative positive response noted that "this tool is a bit more interactive, repetitive and practical than lectures or textbooks" (Intervention). Fifteen participants indicated that the tool was more interactive and/or engaging than lectures or textbooks (e.g., "it's like a fun game to me" (Control)), and an apparent emergent theme was that "high volume in rapid succession" in an interactive "quiz-type format" was effective for "training the eye for quick ID." Thirteen participants viewed the learning tool as an adjunct or complement to traditional resources, e.g., "lectures and textbooks provide a lot of detailed information, but they often only have 1-2 examples of what something looks like (and usually it's the most beautiful example ever, which is almost never what we see in practice)" (Intervention). Conversely, 18 responses indicated that the tool's usefulness was limited by lack of explanations during the feedback phase (e.g., "the repetition and volume of images is helpful but without explanations I found it useless"), while 18 participants explicitly stated a preference for textbooks, lectures, slide seminars, or other resources.

When asked whether "there any other tissue types or specific diagnostic challenges where you think this type of training tool would be particularly useful," n=85 participants responded with a wide variety of tasks and tissue types. The most-suggested organ systems were gastrointestinal pathology (27 respondents; including colon, liver, pancreas, stomach, gallbladder/biliary system, anal biopsies), gynecologic pathology (14 respondents; including endometrium, cervix, and Pap smears), hematopathology (10 respondents; including peripheral blood smears, bone marrow, and lymph nodes), dermatopathology (9 respondents), and bone/soft tissue pathology (7 respondents). Several responses highlighted more specific challenges such as "endometrial phases," "artifacts (e.g., LVI vs retraction artifact)," and "what counts as a mitosis?" 11 respondents specifically suggested cytology, e.g., "This would be fun with cytology. Specifically with thyroid FNA and bronchoscopy smears. I say this because we do ROSE quite a lot with those specimens and we need to make quick decisions about adequacy in the moment. Speed and accurate identification of benign versus malignant is critical."

When asked to suggest a "single most important change" for improving the learning tool before practical usage in pathology training, a full 43% of participants (46/106) suggested adding explanations and/or additional feedback (similar to responses from Session 1). E.g., "add a brief (<10 word) explanation of the feature highlighted in the image that made it benign/malignant/atypical" (Control); "provide reasoning/arrows" (Intervention). One participant's response notably highlighted the tension between perceptual fluency and explicit reasoning, echoing the "blind leading the blind" sentiment of another participant in Session 1: "Add specific feature sets such as looking for invasive patterns in prostate or looking for monomorphic cells in the breast. It would need more structure to be useful, because if you asked me to voice my decisions I would not be able to. It was more just "feel," which I do not think makes me a better pathologist" (Intervention).

Aside from requests for explanations, 19 participants' suggestions focused on various aspects of image quality. Some of these comments emphasized "better clarity and resolution" (Intervention) and making images "closer in color to H&E; The colors were very off, and it made it more difficult" (Control). Five of the image quality-related comments focused on image and view selection, suggesting "better choice of images and selection of fields for viewing" (Control), "better view" (Intervention), "being more selective with which parts of the images have magnified" (Control), "bigger areas to evaluate" (Control), and "complete images, not just a small focus" (Control). Seven of the 106 participants suggested extending or removing the 15-second time limit on each case, e.g., "no time limit to complete the tasks...take away the 15s timer" (Intervention).

Overall, the questionnaire results from both sessions indicate that most participants found the format of the tasks valuable for learning, with many stating that it is more interactive or engaging than other resources such as textbooks and lectures. This aligns with previous work showing that pathology residents strongly favor "flash card style" histology learning activities, characterized by a high volume of short trials with immediate feedback. Participants' responses also suggest that our approach could be applicable to a broad range of histology domains, including gastrointestinal pathology, gynecologic pathology, cytology, and others.

In aggregate, participants identified three priorities for improving the task format towards eventual practical implementation:

1. **Explicit explanations:** 43% of Session 3 respondents suggested adding concise verbal explanations or visual markers (e.g., arrows) to the feedback screen, in order to combine the tool's focus on implicit fluency with explicit diagnostic reasoning.

2. **Image quality:** Suggestions included improving image resolution and standardizing the color balance to better match high-quality H&E staining.

3. **Pacing:** Some participants requested removing the 15-second timer, although others acknowledged the benefits of "high volume in rapid succession" (which is known to aid the development of perceptual fluency [209]).

The mixed reception of the image enhancements for feature highlighting underlines a discrepancy between participants' perceptions and the objective performance gains observed in the mechanistic ablation study. The ablation study showed that image enhancements significantly improved accuracy on *difficult* test cases, but only when difficult examples with enhancement were also seen during training. Because the Intervention curriculum paired enhancements with a "fading" strategy that restricted training to easier (albeit progressively more challenging) cases, in retrospect, the resident participants were never exposed to *difficult cases with enhancements*, the specific condition where the ablation study showed the greatest benefit. This suggests that future iterations should align feature highlighting via enhancements with the curriculum sequence, ensuring that image enhancement is used for challenging cases where it is most effective.

D.10 Mechanistic ablation study: methodology details and additional results

In addition to the main study with pathology residents, 466 lay participants (from the online Prolific platform) were enrolled in a single-session experiment that included the same prostate histology task sequence that residents were administered in Session 1. All participants provided informed consent, and the study followed a protocol approved by the Institutional Review Board of Boston Children's Hospital. We calibrated participant compensation to $15.00 USD per hour.

D.10.1 Experimental design for mechanistic ablation study

To select for Prolific users who were relatively attentive and amenable to visual learning, participants were recruited entirely from a pool of users who had successfully completed a qualifier task designed to assess baseline visual category learning aptitude independent of any medical knowledge. The qualifier task, developed in prior work [156], involved learning to distinguish photos of two species of sea turtles (leatherback vs. loggerhead). Participants learned to click a button labeled "A" or one labeled "B" depending on the species (labels randomly re-assigned for each participant), using a similar interface to the main study but with one image per trial instead of four. Participants completed a 16-trial training sequence with feedback followed by an 8-trial testing sequence. All participants were compensated for completing the screening task, and participants with ≥ 7 correct test phase responses were invited to participate in the main ablation study experiment.

In the ablation study experiment, participants completed a 96-trial prostate histology training sequence with feedback (curriculum differs among participants) followed by a 48-trial testing sequence without feedback (identical among all participants). The experiment compared four randomized groups: Control (same as main study; random curriculum sequence), Intervention (same as main study; combined image selection/sequencing and histology feature enhancement), Select-Only (difficulty-based image selection and sequencing, but no histology feature enhancement), and Enhance-Only (histology feature enhancement, but same random curricula as Control).

D.10.2 Data analysis for mechanistic ablation study

We quantified performance differences among the four experimental groups using three metrics: test phase accuracy, test phase response time (RT), and effective training duration. Effective training duration was defined as the cumulative time for which histology images were displayed on the screen during the training phase (i.e., not including instructions, page load times that may depend on internet connectivity, etc.), with one modification: any duration greater than 60 seconds spent on the feedback screen (which had no time limit) was clipped to 60 seconds to avoid including long periods of inactivity. Feedback screen durations exceeding 60 seconds occurred in fewer than 0.5% of trials in the resident study, and fewer than 0.1% in the ablation study.

Test phase accuracy was analyzed using a logistic generalized linear mixed model (GLMM), where each trial is one data point with a correct or incorrect response. Log-transformed RT was analyzed using a linear mixed model (LMM), keeping only trials with correct responses. Log-transformed effective training duration was analyzed using multiple linear regression.

All models included participant age as a covariate. For accuracy and RT models, the class and position in test sequence of each test case were added as covariates. Additionally, the relatively large sample size in the ablation study allowed post-hoc calculation of an empirical difficulty score for each of the 48 trials/test cases in the testing sequence. Specifically, the empirical difficulty score was defined as the error rate among n=107 Control group participants on each test case. This score was used as a covariate in the statistical models for accuracy and RT, which also included a difficulty × experimental group interaction. In addition to age, class, sequence position, and empirical difficulty, the models for RT and effective training time included additional covariates derived from the qualifier task (turtle classification). Both models included the participant's mean RT during the qualifier task's test phase as a covariate, and the effective training duration model also included the participant's total time spent on the qualifier's training phase. Accuracy and RT models both included random intercepts for participant and test case, and random participant-wise slopes for class and empirical difficulty score. Models that additionally included random participant-wise slopes for test case sequence position failed to converge.

Characteristic	**Total**	**Control**	**Intervention**	**Select-Only**	**Enhance-Only**
No. of participants	455	114	114	116	111
Age (years) — Mean ± SD	42.8 ± 12.8	42.9 ± 13.2	41.8 ± 12.5	41.8 ± 11.2	44.7 ± 14.1
Sex — No. (%)					
Female	256 (56.3%)	69 (60.5%)	60 (52.6%)	65 (56.0%)	62 (55.9%)
Male	194 (42.6%)	43 (37.7%)	54 (47.4%)	50 (43.1%)	47 (42.3%)
Unknown or prefer not to answer	5 (1.1%)	2 (1.8%)	0 (0.0%)	1 (0.9%)	2 (1.8%)
Race/Ethnicity — No. (%)					
Asian	44 (9.7%)	15 (13.2%)	8 (7.0%)	11 (9.5%)	10 (9.0%)
Black or African American	43 (9.5%)	12 (10.5%)	8 (7.0%)	16 (13.8%)	7 (6.3%)
Multiracial	31 (6.8%)	11 (9.6%)	5 (4.4%)	5 (4.3%)	10 (9.0%)
Other	12 (2.6%)	3 (2.6%)	4 (3.5%)	2 (1.7%)	3 (2.7%)
White	323 (71.0%)	73 (64.0%)	87 (76.3%)	82 (70.7%)	81 (73.0%)
Unknown or prefer not to answer	2 (0.4%)	0 (0.0%)	2 (1.8%)	0 (0.0%)	0 (0.0%)

Table D.7: **Baseline characteristics of participants in mechanistic ablation study.** Data are shown for all randomized participants who completed the study session (N=455). The analytic sample (N=446) excluded 4 participants who reported prior familiarity with histology and 5 participants flagged for excessive browser window switching during the task. Race/ethnicity data were collected using the Prolific platform's simplified reporting categories, which characterize participants as "Asian," "Black," "Mixed," "Other," or "White." This schema does not distinguish Hispanic or Latino ethnicity.

			vs. Control Group		**vs. Intervention Group**	
Outcome and Group	**N**	**Mean (SD)**	**Estimate (95% CI)**	**Adj. *P* Val.**	**Estimate (95% CI)**	**Adj. *P* Val.**
Test Accuracy (%)			**Odds Ratio**		**Odds Ratio**	
Control	107	56.9 (10.2)	—	—	—	—
Intervention	114	63.8 (9.3)	1.48 (1.24 to 1.76)	**<0.001**	—	—
Select-Only	115	63.1 (9.7)	1.40 (1.18 to 1.67)	**<0.001**	0.95 (0.80 to 1.13)	0.88
Enhance-Only	110	61.0 (10.2)	1.21 (1.02 to 1.45)	**0.02**	0.82 (0.69 to 0.98)	**0.02**
Test Response Time (sec)			**Geo. Mean Ratio**		**Geo. Mean Ratio**	
Control	107	2.52 (1.64)	—	—	—	—
Intervention	114	1.93 (0.70)	0.87 (0.76 to 1.00)	0.06	—	—
Select-Only	115	2.17 (1.18)	0.94 (0.81 to 1.08)	0.61	1.07 (0.94 to 1.23)	0.55
Enhance-Only	110	2.57 (1.18)	1.05 (0.91 to 1.21)	0.79	1.21 (1.05 to 1.39)	**0.003**
Effective Train Dur. (min)			**Geo. Mean Ratio**		**Geo. Mean Ratio**	
Control	107	7.4 (2.9)	—	—	—	—
Intervention	114	5.2 (1.4)	0.75 (0.68 to 0.83)	**<0.001**	—	—
Select-Only	115	6.0 (2.5)	0.81 (0.73 to 0.89)	**<0.001**	1.08 (0.98 to 1.19)	0.18
Enhance-Only	110	7.7 (4.2)	1.00 (0.91 to 1.11)	1	1.34 (1.22 to 1.48)	**<0.001**

Table D.8: **Mechanistic ablation study: both easy-to-hard fading and feature highlighting with enhanced images can improve performance in lay participants.** "Estimate" refers to odds ratio between conditions of odds of correct response for accuracy (modeled with generalized linear mixed model, logistic link), and to the ratio of geometric means ("Geo. Mean Ratio") for response time (linear mixed model) and effective training duration (multiple linear regression). Response time and training duration are log-transformed before modeling, but mean and standard deviation (SD) are on original scales. Response time calculations consider only trials with correct responses. All P values ("Adj. P Val.") are Tukey-adjusted for multiple comparisons among groups (separately for each outcome). All between-group comparisons not shown in the table are non-significant (i.e., those Select-Only vs. Enhance-Only), except that Select-Only had a shorter effective training duration than Enhance-Only (geometric mean ratio 0.81; 95% CI 0.73 to 0.89; $P < 0.001$).

		vs. Control Group		vs. Intervention Group	
Outcome, Difficulty, Group	**Est. Mean (95% CI)**	**Est. (95% CI)**	**Adj. P Val.**	**Est. (95% CI)**	**Adj. P Val.**
TEST ACCURACY (%)		**Odds Ratio**		**Odds Ratio**	
Easy Trials (-1 SD)					
Control	74.7 (71.0 to 78.1)	—	—	—	—
Intervention	85.3 (82.7 to 87.5)	1.96 (1.48 to 2.61)	**<0.001**	—	—
Select-Only	83.2 (80.4 to 85.7)	1.68 (1.27 to 2.23)	**<0.001**	0.86 (0.64 to 1.14)	0.49
Enhance-Only	75.2 (71.6 to 78.5)	1.03 (0.78 to 1.36)	0.99	0.52 (0.40 to 0.70)	**<0.001**
Average Trials (Mean)					
Control	58.4 (55.4 to 61.3)	—	—	—	—
Intervention	67.4 (64.7 to 70.1)	1.48 (1.24 to 1.76)	**<0.001**	—	—
Select-Only	66.3 (63.5 to 69.0)	1.40 (1.18 to 1.67)	**<0.001**	0.95 (0.80 to 1.13)	0.88
Enhance-Only	63.0 (60.1 to 65.8)	1.21 (1.02 to 1.45)	**0.02**	0.82 (0.69 to 0.98)	**0.02**
Hard Trials (+1 SD)					
Control	40.0 (36.0 to 44.1)	—	—	—	—
Intervention	42.5 (38.5 to 46.6)	1.11 (0.88 to 1.40)	0.66	—	—
Select-Only	43.9 (39.8 to 48.0)	1.17 (0.93 to 1.48)	0.3	1.06 (0.84 to 1.33)	0.93
Enhance-Only	48.8 (44.7 to 53.0)	1.43 (1.13 to 1.80)	**<0.001**	1.29 (1.02 to 1.62)	**0.02**
TEST RESP. TIME (sec)		**Geo. Mean Ratio**		**Geo. Mean Ratio**	
Easy Trials (-1 SD)					
Control	1.88 (1.72 to 2.04)	—	—	—	—
Intervention	1.62 (1.49 to 1.76)	0.86 (0.75 to 0.99)	**0.03**	—	—
Select-Only	1.74 (1.60 to 1.89)	0.93 (0.80 to 1.07)	0.49	1.07 (0.94 to 1.23)	0.54
Enhance-Only	2.02 (1.85 to 2.20)	1.08 (0.93 to 1.24)	0.56	1.25 (1.08 to 1.44)	**<0.001**
Average Trials (Mean)					
Control	1.97 (1.82 to 2.14)	—	—	—	—
Intervention	1.73 (1.59 to 1.87)	0.87 (0.76 to 1.01)	0.07	—	—
Select-Only	1.85 (1.71 to 2.00)	0.94 (0.81 to 1.08)	0.65	1.07 (0.93 to 1.23)	0.56
Enhance-Only	2.06 (1.90 to 2.23)	1.05 (0.91 to 1.21)	0.86	1.19 (1.04 to 1.38)	**0.006**
Hard Trials (+1 SD)					
Control	2.07 (1.89 to 2.27)	—	—	—	—
Intervention	1.84 (1.69 to 2.01)	0.89 (0.76 to 1.03)	0.18	—	—
Select-Only	1.97 (1.81 to 2.15)	0.95 (0.82 to 1.10)	0.82	1.07 (0.92 to 1.24)	0.63
Enhance-Only	2.11 (1.93 to 2.30)	1.02 (0.87 to 1.18)	0.99	1.14 (0.99 to 1.33)	0.09

Table D.9: **Mechanistic ablation study group comparisons broken down by task difficulty: enhancing hard images during training improves accuracy on hard images during testing.** For each trial in the 48-trial testing sequence, difficulty was estimated as the empirical error rate among n=107 Control group participants. This difficulty score was used as a covariate in the model. "Easy Trials" refers to means estimated by the model for an (hypothetical) item with difficulty one standard deviation (SD) below the mean. "Average Trials" are estimated for an item of mean difficulty, and "Hard Trials" for an item one SD above the mean. "Est. (95% CI)" is odds ratio of correct responses between conditions for accuracy (generalized linear mixed model, logistic link), and geometric mean ratio ("Geo. Mean Ratio") for response time (linear mixed model). Response times are log-transformed before modeling, but mean and SD values in the table use original scales. Response time calculations consider only trials with correct responses. All P values ("Adj. P Val.") are Tukey-adjusted for multiple comparisons (separately for each outcome). In addition to the differences shown in the table, Select-Only had higher easy-trial accuracy (odds ratio 1.63, 95% CI 1.23 to 2.16; $P < 0.001$), and faster easy-trial response time (geometric mean ratio 0.86, 95% CI 0.75 to 0.99; $P = 0.03$) than Enhance-Only.

D.10.3 Detailed results of mechanistic ablation study

Table D.7 describes the baseline demographic characteristics of the ablation study participants, and Figure D.8 shows participant recruitment and flow. The main results of the ablation study (accuracy, test response time on correct trials, and effective training duration) are shown in Table D.8 (see also Figure 4.4). Participants randomized to Select-Only, Enhance-Only, and the combined Intervention all had significantly higher test-phase accuracy than Control participants. Intervention significantly outperformed Enhance-Only overall in accuracy, response time, and effective training duration. Intervention and Select-Only both had shorter effective training durations than both Control and Enhance-Only. Section 4.4.4 and Figure 4.4 contain additional details.

In the analysis of trial-by-trial accuracy data, the GLMM model yielded a significant interaction effect between test case difficulty and experimental condition ($\chi^2_3 = 44.3$, $P < 0.001$). On easy test cases (model-estimated one SD below the mean empirical error rate), Select-Only and Intervention participants had higher accuracy than Control and Enhance-Only participants (all $P < 0.001$ corrected, see Table D.9). However, on difficult test cases (one SD above mean), Enhance-Only participants had higher accuracy than both Control (OR, 1.43; 95% CI, 1.13 to 1.80; $P < 0.001$) and Intervention (OR, 1.29; 95% CI, 1.02 to 1.62; $P = 0.02$). On moderately difficult test cases (estimated at the mean empirical error rate), results closely mirrored the overall findings, with all three AI-assisted groups outperforming Control ($P < 0.001$ for Intervention and Select-Only, $P = 0.02$ for Enhance-Only), and Intervention outperforming Enhance-Only (OR for Enhance-Only vs. Intervention, 0.82; 95% CI, 0.69 to 0.98; $P = 0.02$).

A significant interaction effect between difficulty and experimental condition was also observed for the RT endpoint ($\chi^2_3 = 13.8$, $P = 0.003$) in addition to the accuracy endpoint as described in the Main Text. On easy test cases, Intervention had significantly faster RT than Control (Geometric Mean Ratio (GMR), 0.86; 95% CI, 0.75 to 0.99; $P = 0.03$), while Select-Only had significantly faster RT than Enhance-Only (Geometric Mean Ratio (GMR), 0.86; 95% CI, 0.75 to 0.99; $P = 0.03$). On moderate test cases, Enhance-Only had significantly slower RT than Intervention (GMR, 1.19; 95% CI, 1.04 to 1.38; $P = 0.006$). Table D.9 contains the full results of the ablation study broken down by test case difficulty.

Confusion matrices for each group from the test phase of the ablation study are shown in Figure D.9. Upon visual inspection, the confusion matrices align with the primary accuracy findings: compared with the Control group, recall rates for both benign and malignant classes (diagonal elements) are generally higher among the AI-assisted groups.

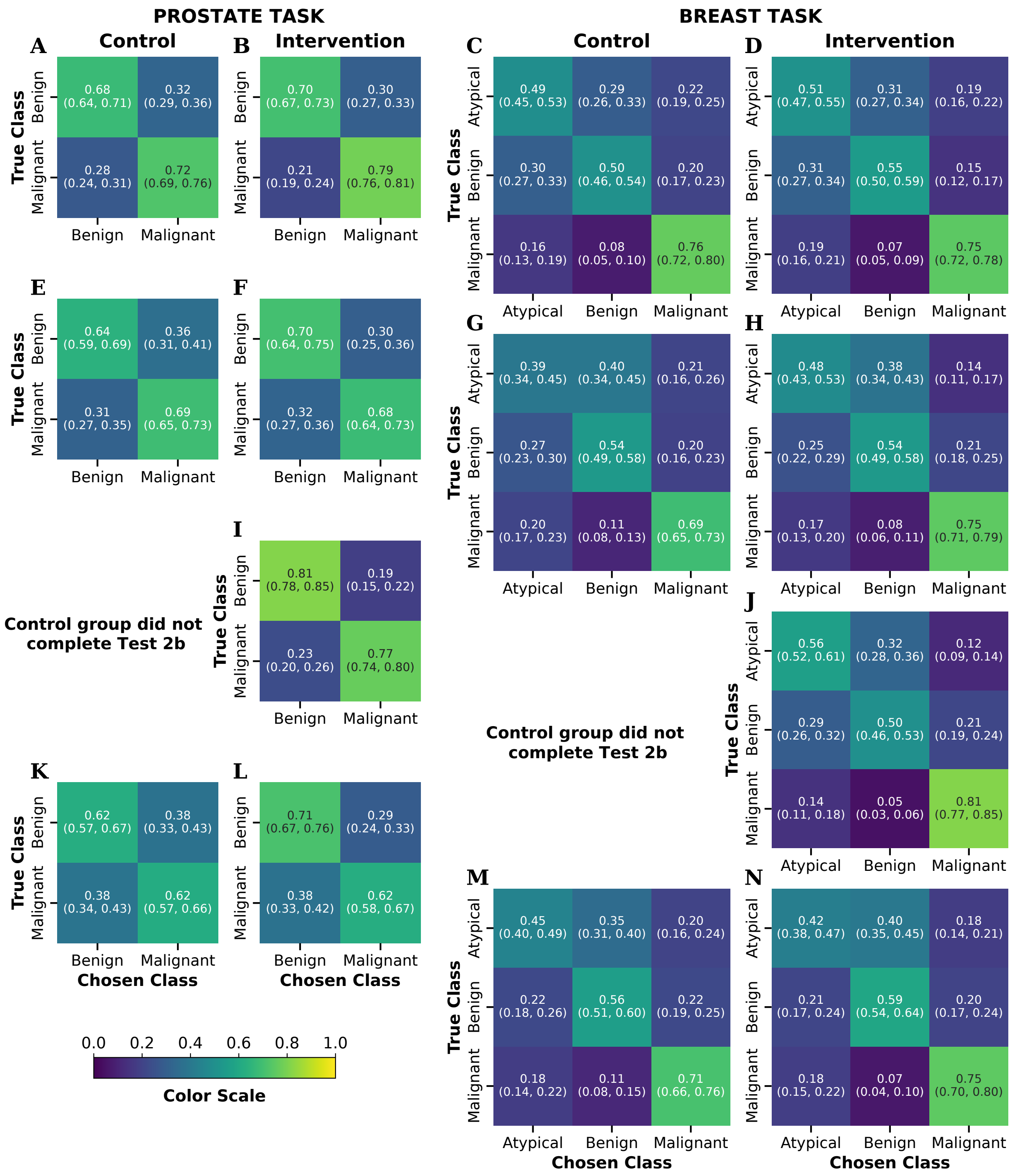


Figure D.5: **Confusion matrices of resident participants in longitudinal study during test phases, by experimental group.** Plots in the left two columns correspond to prostate task results, and plots in the right two columns correspond to breast task results (Control group on the left and Intervention group on the right within each tissue type). Each confusion matrix corresponds to one distinct test phase in the study (Figure 4.1). Panels A-D are for Test 1, E-H for Test 2a, I-J for Test 2b, K-N for Test 3. Matrices are normalized such that each row adds to 1; the value in each cell represents the proportion of times the class of its column was chosen among all true instances of the class in its row. The main values are the means among participants, with 95% confidence intervals in parentheses (10,000 bootstrap replicates, resampling among participants).

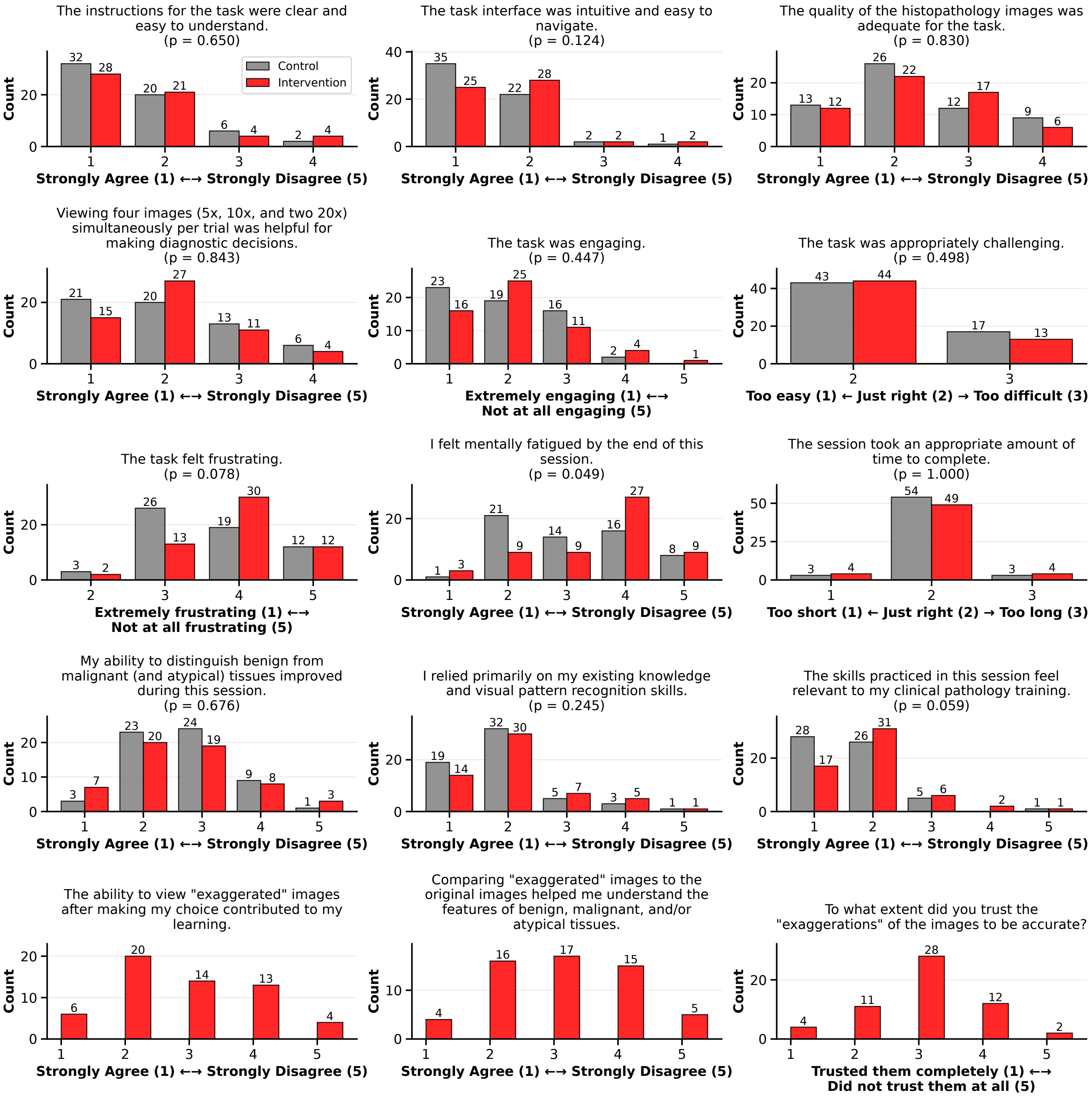
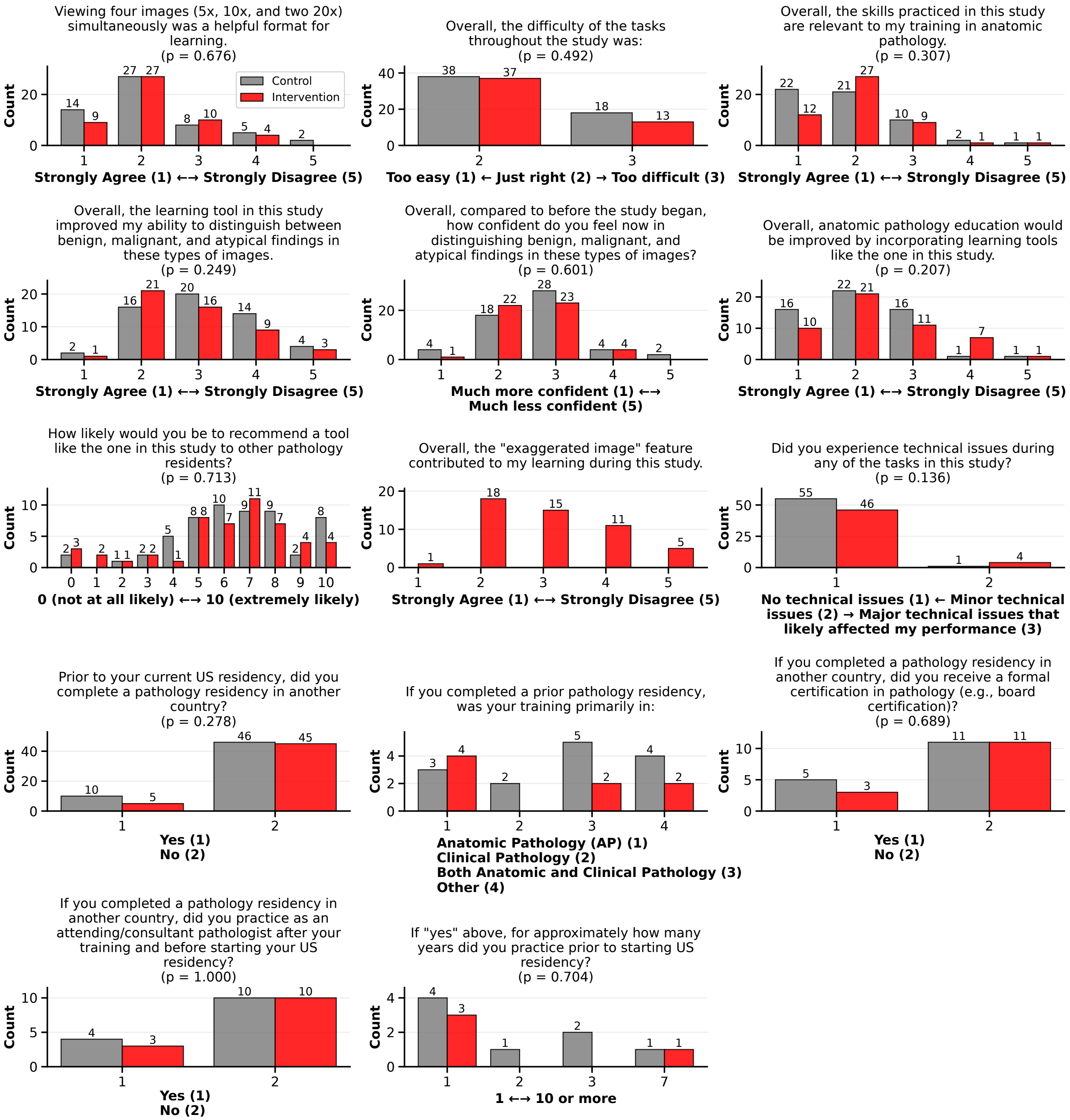


Figure D.6: **Likert scale/multiple choice question responses from the Session 1 questionnaire.** Bar plots are ordered from left to right (starting top-left) in the order they appeared to participants. Only Intervention group participants were asked the questions pertaining to enhanced ("exaggerated") images. Intermediate response options to "The task was engaging" were "Very engaging," "Moderately engaging," and "Slightly engaging"; for "The task felt frustrating" they were "Very frustrating," "Moderately frustrating," and "Slightly frustrating"; for the question about trusting enhanced images they were "Trusted them most of the time," "Was neutral / unsure," and "Was skeptical of them"; and those for all "Strongly Agree to Strongly Disagree" Likert scale questions were "Agree," "Neutral," and "Disagree." Comparisons between Control and Intervention groups were conducted on an exploratory basis using Mann-Whitney U tests. All P values are uncorrected, and are intended for descriptive purposes only.



Viewing four images (5x, 10x, and two 20x) simultaneously was a helpful format for learning. (p = 0.676)

Overall, the difficulty of the tasks throughout the study was: (p = 0.492)

Overall, the skills practiced in this study are relevant to my training in anatomic pathology. (p = 0.307)

Overall, the learning tool in this study improved my ability to distinguish between benign, malignant, and atypical findings in these types of images. (p = 0.249)

Overall, compared to before the study began, how confident do you feel now in distinguishing benign, malignant, and atypical findings in these types of images? (p = 0.601)

Overall, anatomic pathology education would be improved by incorporating learning tools like the one in this study. (p = 0.207)

How likely would you be to recommend a tool like the one in this study to other pathology residents? (p = 0.713)

Overall, the "exaggerated image" feature contributed to my learning during this study.

Did you experience technical issues during any of the tasks in this study? (p = 0.136)

Count
0
25
50
55
46
1
4
1
2
No technical issues (1) ← Minor technical issues (2) → Major technical issues that likely affected my performance (3)

Prior to your current US residency, did you complete a pathology residency in another country? (p = 0.278)

If you completed a prior pathology residency, was your training primarily in:

Count
0
2
4
3
4
2
5
2
4
2
1
2
3
4
Anatomic Pathology (AP) (1)
Clinical Pathology (2)
Both Anatomic and Clinical Pathology (3)
Other (4)

If you completed a pathology residency in another country, did you receive a formal certification in pathology (e.g., board certification)? (p = 0.689)

If you completed a pathology residency in another country, did you practice as an attending/consultant pathologist after your training and before starting your US residency? (p = 1.000)

If "yes" above, for approximately how many years did you practice prior to starting US residency? (p = 0.704)

Figure D.7: **Likert scale/multiple choice question responses from the Session 3 questionnaire.** Bar plots appear in the order of the questionnaire as it appeared to participants (left to right from top-left). Only Intervention group participants were asked about enhanced images. Intermediate response options for the confidence question (middle of second row) were "Slightly more confident," "About the same," and "Slightly less confident"; intermediate options for all "Strongly Agree to Strongly Disagree" questions were "Agree," "Neutral," and "Disagree." For the question about likelihood of peer recommendation (third row, left), response option 5 was labeled "5 (neutral)" with all other options labeled only as integers. Exploratory comparisons between Control and Intervention groups were conducted using Fisher's exact test for binary responses and Mann-Whitney U tests for non-binary ordinal responses. P values are uncorrected and are for descriptive purposes only.

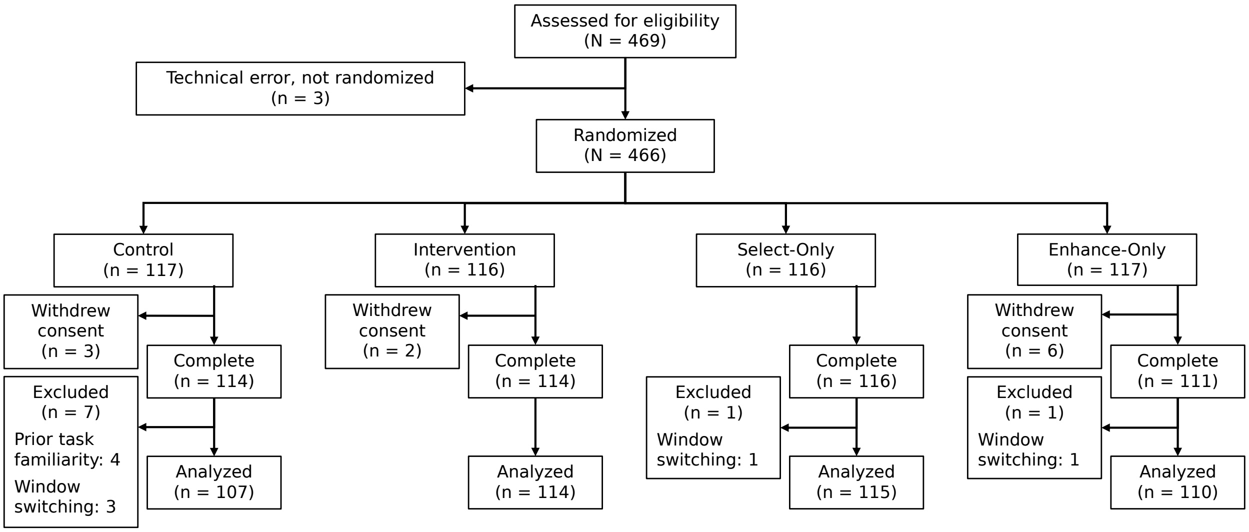


Figure D.8: **Participant flow for the ablation study (CONSORT diagram).** A total of 469 participants were recruited from the online platform Prolific. Technical errors caused 3 participants to be excluded prior to randomization. Of the 466 randomized participants, 455 (97.6%) completed the task, and 446 (95.7%) were included in the analysis. Exclusions regarding prior task familiarity and excessive window switching were adjudicated based on post-task questionnaire responses without knowledge of experimental condition.

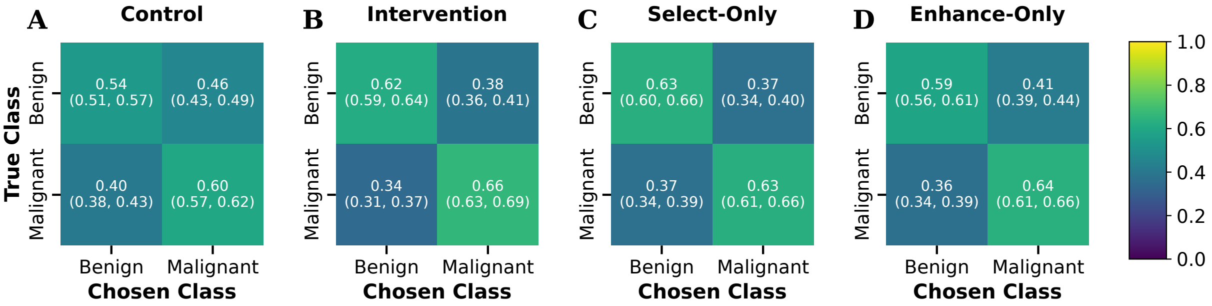


Figure D.9: **Confusion matrices of mechanistic ablation study participants in the test phase.** Panels A-D show the class confusion matrices for Control, Intervention, Select-Only, and Enhance-Only groups respectively. Matrices are normalized by row (each row adds to 1); The value in each cell is the proportion of times the class of its column was chosen among all true presentations of the class in its row. Values are means among participants, with 95% confidence intervals by bootstrap in parentheses (10,000 replicates, resampling among participants).

Appendix E

Optimizing Curricula for Human Visual Category Learning Via Image-Computable Surrogate Learners

E.1 Prompt for obtaining candidate verbal decision rules from vision-language models

The prompt on the next page was used to obtain a set of 50 descriptive sentences to be used as "verbal decision rules" in Chapter 5. This prompt was passed to the Gemini 3.0 Pro Preview model (using the online browser-based interface), along with 20 images from the MHIST dataset [131]: 10 hyperplastic polyp patches ("benign") and 10 sessile serrated adenoma patches ("malignant"). The letters A and B were randomly assigned to malignant and benign respectively. The class matching letter matching each image's class was added to each image on a white square background in the top-left corner, with the label taking up about 10% of the image's width and height (1% of the image area). The MHIST dataset was originally annotated by 7 pathologists; the image set passed to the vision-language model was randomly sampled among those where at least 5 of the 7 pathologists agreed on the diagnosis.

Prompt:

```
You are an intelligent, observant participant in a visual research study.
You have **zero** domain expertise or background knowledge about the
subject matter in these images.

Your task is to analyze the images labeled as classes ’A’ and ’B’ and
propose a set of 50 explicit hypotheses (rules) that help distinguish
them.  Your goal is to create a ’cheat sheet’ that would give another
person the best possible chance of correctly classifying new images.

**Phase 1:  Unbiased Observation**

Before writing the rules, study the images carefully.  Look for the most
consistent and distinct visual differences.  Do not look for specific
categories (like shape or color) unless they are actually the defining
features.  Let the images dictate what is important.

**Phase 2:  Rule Generation**

Generate 50 distinct rules (25 for Class A, 25 for Class B).

**Guidelines:**

1.  **Vocabulary:** Use precise, descriptive English that an adult
non-expert would understand.  Avoid domain-specific jargon (e.g.,
if these were biological images, you would not use words like
’cells’ or ’tissue’; if they were architectural, you would not use
’baroque’ or ’cantilever’).  Describe **visual appearance only**.
2.  **Comparisons:** Where possible, write rules that highlight a
contrast.  (e.g., ’The surface is rough and fuzzy, rather than smooth
and shiny’).  3.  **Metaphors:** You are encouraged to use analogies
to everyday objects to capture complex textures (e.g., ’looks like a
honeycomb,’ ’resembles a map’).  4.  **Independence:** Each rule should
be a standalone statement that a person could use to check an image.

Output the results as a formatted .csv with columns ’sentence’ (the rule)
and ’class’ (A or B).
```

E.2 Pseudocode for the verbal rule surrogate learner

Algorithm 2 Neurosymbolic learning algorithm based on verbalizable rules

```
**Initialize:** $w \leftarrow w_{init}(\mathcal{X}_{init})$, $\mathcal{B} \leftarrow \emptyset$, $\tau \leftarrow 0.5$
**Initial Rule:** $k_t \sim \text{Softmax}(w, T)$; $d_{k_t} \sim \text{Rand}(\pm 1)$; $b_{k_t} \leftarrow \text{mean}(s_{k_t}(\mathcal{X}_{init}))$   ▷ Init bias to global cosine similarity mean
**for** trial $t = 1$ to $T_{max}$ **do**
    Observe image $x_t$
    **Explicit Rule-Based Prediction:**
    $\hat{y}_{exp} \leftarrow d_{k_t} \cdot \text{sign}(s_{k_t}(x_t) - b_{k_t})$; $C_{exp} \leftarrow (1 + \exp(-\gamma |s_{k_t}(x_t) - b_{k_t}|))^{-1}$
    **Feedback & Update Given Ground Truth $y_t$:**
    $\tau \leftarrow$ (if $\hat{y}_{exp} = y_t$ then $\min(1, \tau + \alpha_{pos})$ else $\tau \times \alpha_{neg}$)
    Define Exemplar Pool $\mathcal{P} \leftarrow \mathcal{B} \cup \{(x_t, y_t)\}$
    **if** $|\mathcal{P}_A| > 0$ and $|\mathcal{P}_B| > 0$ **then**   ▷ Rational updates require data from both classes
        **if** $\tau < \tau_{min}$ **then**   ▷ Switching Logic
            **for** each rule $k$ **do**
                $D_k \leftarrow \text{mean}(s_k(x \in \mathcal{P}_A)) - \text{mean}(s_k(x \in \mathcal{P}_B))$
                $M_k \leftarrow |D_k|$; $d_k \leftarrow \text{sign}(D_k)$
            **end for**
            Define $\mathcal{K}_{top}$ as indices of top $r$ values of $(w \cdot M)$
            Select $k_{next} \sim \text{TruncatedSoftmax}(w \cdot M, \mathcal{K}_{top}, T)$
            Update Active Rule: $k_t \leftarrow k_{next}$; $d_{k_t} \leftarrow d_{k_{next}}$; $\tau \leftarrow 1.0$
            $b_{k_t} \leftarrow b^*(\mathcal{P})$   ▷ Fresh Bias
        **else**   ▷ Maintenance Logic
            $b_{k_t} \leftarrow \lambda b_{k_t} + (1 - \lambda) b^*(\mathcal{P})$   ▷ Update Bias
        **end if**
    **else**   ▷ **Insufficient Data (beginning of curriculum)**
        $b_{k_t} \leftarrow \text{mean}(s_{k_t}(x \in \mathcal{P}))$   ▷ Track mean of seen class
        *No rule switching allowed yet*
    **end if**
    $w_{k_t} \leftarrow \max(w_{k_t} + \eta(y_t \cdot \hat{y}_{exp}), 0)$; re-normalize $w$ such that $\sum_k w_k = 1$   ▷ Update global rule utility estimate
    Update $\mathcal{B}$ with $(x_t, y_t)$ (Stratified first-in-first-out)
**end for**
```

References

[1] M. M. Rogers. “Teaching-to-learn: its effects on conceptual knowledge learning in university students.” *International Journal of Innovative Teaching and Learning in Higher Education (IJITLHE)*, **2**(1), 2021, pp. 1–14.

[2] O. Henkel, H. Horne-Robinson, N. Kozhakhmetova, and A. Lee. “Effective and scalable math support: Experimental evidence on the impact of an AI-math tutor in Ghana.” In: *International Conference on Artificial Intelligence in Education*. Springer. 2024, pp. 373–381.

[3] G. Kestin, K. Miller, A. Klales, T. Milbourne, and G. Ponti. “AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting.” *Scientific Reports*, **15**(1), 2025, p. 17458.

[4] S. García-Méndez, F. de Arriba-Pérez, and M. d. C. Somoza-López. “A review on the use of large language models as virtual tutors.” *Science & Education*, **34**(2), 2025, pp. 877–892.

[5] N. Matsuda, E. Yarzebinski, V. Keiser, R. Raizada, W. W. Cohen, G. J. Stylianides, and K. R. Koedinger. “Cognitive anatomy of tutor learning: Lessons learned with SimStudent.” *Journal of Educational Psychology*, **105**(4), 2013, p. 1152.

[6] L. Ai, J. Langer, S. H. Muggleton, and U. Schmid. “Explanatory machine learning for sequential human teaching.” *Machine Learning*, **112**(10), 2023, pp. 3591–3632.

[7] A. Sen, P. Patel, M. A. Rau, B. Mason, R. Nowak, T. T. Rogers, and X. Zhu. “Machine Beats Human at Sequencing Visuals for Perceptual-Fluency Practice.” *International Educational Data Mining Society*, 2018.

[8] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. “Mastering the game of Go with deep neural networks and tree search.” *Nature*, **529**(7587), 2016, pp. 484–489.

[9] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars. “A continual learning survey: Defying forgetting in classification tasks.” *IEEE transactions on pattern analysis and machine intelligence*, **44**(7), 2021, pp. 3366–3385.

[10] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. “Intriguing properties of neural networks.” In: *International Conference on Learning Representations*. 2014.

[11] M. Schrimpf, J. Kubilius, M. J. Lee, N. A. R. Murty, R. Ajemian, and J. J. DiCarlo. "Integrative benchmarking to advance neurally mechanistic models of human intelligence." *Neuron*, **108**(3), 2020, pp. 413–423.

[12] J. Coda-Forno, M. Binz, J. X. Wang, and E. Schulz. "CogBench: a large language model walks into a psychology lab." In: *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. Proceedings of Machine Learning Research. PMLR, 2024, pp. 9076–9108. URL: https://proceedings.mlr.press/v235/coda-forno24a.html.

[13] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. "Performance-optimized hierarchical models predict neural responses in higher visual cortex." *Proceedings of the National Academy of Sciences*, **111**(23), 2014, pp. 8619–8624.

[14] E. Hosseini, C. Casto, N. Zaslavsky, C. Conwell, M. Richardson, and E. Fedorenko. "Universality of representation in biological and artificial neural networks." *bioRxiv*, 2024.

[15] M. Schrimpf, I. A. Blank, G. Tuckute, C. Kauf, E. A. Hosseini, N. Kanwisher, J. B. Tenenbaum, and E. Fedorenko. "The neural architecture of language: Integrative modeling converges on predictive processing." *Proceedings of the National Academy of Sciences*, **118**(45), 2021, e2105646118.

[16] T. Fel, I. Felipe, D. Linsley, and T. Serre. "Harmonizing the object recognition strategies of deep neural networks with humans." In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022, pp. 9432–9446.

[17] M. Zhang et al. "Can Machines Imitate Humans? Integrative Turing Tests for Vision and Language Demonstrate a Narrowing Gap." *arXiv preprint arXiv:2211.13087*, 2022.

[18] T. Bricken, X. Davies, D. Singh, D. Krotov, and G. Kreiman. "Sparse Distributed Memory is a Continual Learner." In: *International Conference on Learning Representations*. 2023.

[19] S. Chandra, S. Sharma, R. Chaudhuri, and I. Fiete. "Episodic and associative memory from spatial scaffolds in the hippocampus." *Nature*, **638**(8051), 2025, pp. 739–751.

[20] J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Russakovsky. "Human uncertainty makes classification more robust." In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 9617–9626.

[21] R. L. Haspel, J. R. Genzen, J. Wagner, C. M. Lockwood, and K. Fong. "Integration of Genomic Medicine in Pathology Resident Training: A Work in Progress." *American Journal of Clinical Pathology*, **154**(6), 2020, pp. 784–791.

[22] B. Mai, N. Aakash, J. Huddin, B. Castillo, and A. Wahed. "Pathology residency curriculum: time for a change?" *Annals of Clinical & Laboratory Science*, **51**(3), 2021, pp. 434–440.

[23] J. A. Chapman, L. M. Lee, and N. T. Swailes. "From scope to screen: the evolution of histology education." In: *Biomedical Visualisation: Volume 8*. Springer, 2020, pp. 75–107.

[24] P. Singh, Y. Li, A. Sikarwar, S. W. Lei, D. Gao, M. B. Talbot, Y. Sun, M. Z. Shou, G. Kreiman, and M. Zhang. "Learning to learn: How to continuously teach humans and machines." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 11708–11719.

[25] M. McCloskey and N. J. Cohen. "Catastrophic interference in connectionist networks: The sequential learning problem." In: *Psychology of learning and motivation*. Vol. 24. Elsevier, 1989, pp. 109–165.

[26] R. Ratcliff. "Connectionist models of recognition memory: constraints imposed by learning and forgetting functions." *Psychological review*, **97**(2), 1990, p. 285.

[27] R. M. French. "Catastrophic forgetting in connectionist networks." *Trends in cognitive sciences*, **3**(4), 1999, pp. 128–135.

[28] T. L. Hayes, K. Kafle, R. Shrestha, M. Acharya, and C. Kanan. "Remind your neural network to prevent catastrophic forgetting." In: *European Conference on Computer Vision*. Springer. 2020, pp. 466–483.

[29] R. Kemker, M. McClure, A. Abitino, T. L. Hayes, and C. Kanan. "Measuring catastrophic forgetting in neural networks." In: *Thirty-second AAAI conference on artificial intelligence*. 2018.

[30] D. Maltoni and V. Lomonaco. "Continuous learning in single-incremental-task scenarios." *Neural Networks*, 2019.

[31] I. Evron, E. Moroshko, R. Ward, N. Srebro, and D. Soudry. "How catastrophic can catastrophic forgetting be in linear regression?" In: *Conference on Learning Theory*. PMLR. 2022, pp. 4028–4079.

[32] H. Vaidya, T. Desell, and A. G. Ororbia. "Reducing Catastrophic Forgetting in Self Organizing Maps with Internally-Induced Generative Replay (Student Abstract)." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 11. 2022, pp. 13069–13070.

[33] C. Shao and Y. Feng. "Overcoming Catastrophic Forgetting beyond Continual Learning: Balanced Training for Neural Machine Translation." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2022, pp. 2023–2036.

[34] A. Prabhu, P. H. Torr, and P. K. Dokania. "Gdumb: A simple approach that questions our progress in continual learning." In: *European conference on computer vision*. Springer. 2020, pp. 524–540.

[35] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. "Overcoming catastrophic forgetting in neural networks." *Proceedings of the national academy of sciences*, **114**(13), 2017, pp. 3521–3526.

[36] C. S. Lee and A. Y. Lee. "Clinical applications of continual learning machine learning." *The Lancet Digital Health*, **2**(6), 2020, e279–e281.

[37] X. Wang, T. Ma, J. Ainooson, S. Cha, X. Wang, A. Molla, and M. Kunda. “The toybox dataset of egocentric visual object transformations.” *arXiv preprint arXiv:1806.06034*, 2018.

[38] A. Borji, S. Izadi, and L. Itti. “ilab-20m: A large-scale controlled object dataset to investigate deep learning.” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2221–2230.

[39] Z. Li and D. Hoiem. “Learning without forgetting.” *IEEE transactions on pattern analysis and machine intelligence*, **40**(12), 2017, pp. 2935–2947.

[40] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny. “Efficient lifelong learning with a-gem.” *arXiv preprint arXiv:1812.00420*, 2018.

[41] X. He and H. Jaeger. “Overcoming catastrophic interference using conceptor-aided backpropagation.” 2018.

[42] F. Zenke, B. Poole, and S. Ganguli. “Continual learning through synaptic intelligence.” In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 3987–3995.

[43] S.-W. Lee, J.-H. Kim, J. Jun, J.-W. Ha, and B.-T. Zhang. “Overcoming catastrophic forgetting by incremental moment matching.” In: *Advances in neural information processing systems*. 2017, pp. 4652–4662.

[44] Y. Kong, L. Liu, H. Chen, J. Kacprzyk, and D. Tao. “Overcoming catastrophic forgetting in continual learning by exploring eigenvalues of hessian matrix.” *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[45] J. Wen, Y. Cao, and R. Huang. “Few-Shot Self Reminder to Overcome Catastrophic Forgetting.” *arXiv preprint arXiv:1812.00543*, 2018.

[46] Z. Mai, R. Li, J. Jeong, D. Quispe, H. Kim, and S. Sanner. “Online continual learning in image classification: An empirical survey.” *Neurocomputing*, **469**, 2022, pp. 28–51.

[47] Z. Zhang, Y. Chen, and C. Zhou. “Self-growing binary activation network: A novel deep learning model with dynamic architecture.” *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[48] C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, A. A. Rusu, A. Pritzel, and D. Wierstra. “Pathnet: Evolution channels gradient descent in super neural networks.” *arXiv preprint arXiv:1701.08734*, 2017.

[49] J. Rajasegaran, M. Hayat, S. H. Khan, F. S. Khan, and L. Shao. “Random path selection for continual learning.” *Advances in Neural Information Processing Systems*, **32**, 2019.

[50] J. Serra, D. Suris, M. Miron, and A. Karatzoglou. “Overcoming catastrophic forgetting with hard attention to the task.” In: *International Conference on Machine Learning*. PMLR. 2018, pp. 4548–4557.

[51] T. Adel, H. Zhao, and R. E. Turner. “Continual learning with adaptive weights (claw).” *arXiv preprint arXiv:1911.09514*, 2019.

[52] J. Schwarz, J. Luketina, W. M. Czarnecki, A. Grabska-Barwinska, Y. W. Teh, R. Pascanu, and R. Hadsell. “Progress & compress: A scalable framework for continual learning.” *arXiv preprint arXiv:1805.06370*, 2018.

[53] S. Golkar, M. Kagan, and K. Cho. “Continual learning via neural pruning.” *arXiv preprint arXiv:1903.04476*, 2019.

[54] J. Peng, B. Tang, H. Jiang, Z. Li, Y. Lei, T. Lin, and H. Li. “Overcoming long-term catastrophic forgetting through adversarial neural pruning and synaptic consolidation.” *IEEE Transactions on Neural Networks and Learning Systems*, **33**(9), 2021, pp. 4243–4256.

[55] Q. Gao, Z. Luo, D. Klabjan, and F. Zhang. “Efficient architecture search for continual learning.” *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[56] Y. Wu, Y. Chen, L. Wang, Y. Ye, Z. Liu, Y. Guo, and Y. Fu. “Large scale incremental learning.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 374–382.

[57] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert. “ICARL: Incremental classifier and representation learning.” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 2001–2010.

[58] R. Aljundi, M. Lin, B. Goujaud, and Y. Bengio. “Gradient based sample selection for online continual learning.” *arXiv preprint arXiv:1903.08671*, 2019.

[59] C. V. Nguyen, Y. Li, T. D. Bui, and R. E. Turner. “Variational continual learning.” *arXiv preprint arXiv:1710.10628*, 2017.

[60] D. Lopez-Paz et al. “Gradient episodic memory for continual learning.” In: *Advances in Neural Information Processing Systems*. 2017, pp. 6467–6476.

[61] J. Bang, H. Kim, Y. Yoo, J.-W. Ha, and J. Choi. “Rainbow Memory: Continual Learning with a Memory of Diverse Samples.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 8218–8227.

[62] H. Shin, J. K. Lee, J. Kim, and J. Kim. “Continual learning with deep generative replay.” In: *Advances in Neural Information Processing Systems*. 2017, pp. 2990–2999.

[63] A. Robins. “Catastrophic forgetting, rehearsal and pseudorehearsal.” *Connection Science*, **7**(2), 1995, pp. 123–146.

[64] Y. Liu, Y. Su, A.-A. Liu, B. Schiele, and Q. Sun. “Mnemonics training: Multi-class incremental learning without forgetting.” In: *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 2020, pp. 12245–12254.

[65] C. Atkinson, B. McCane, L. Szymanski, and A. Robins. “Pseudo-recursal: Solving the catastrophic forgetting problem in deep neural networks.” *arXiv preprint arXiv:1802.03875*, 2018.

[66] X. Liu, C. Wu, M. Menta, L. Herranz, B. Raducanu, A. D. Bagdanov, S. Jui, and J. v. de Weijer. “Generative feature replay for class-incremental learning.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020, pp. 226–227.

[67] G. Shen, S. Zhang, X. Chen, and Z.-H. Deng. “Generative feature replay with orthogonal weight modification for continual learning.” In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2021, pp. 1–8.

[68] G. M. van de Ven, Z. Li, and A. S. Tolias. “Class-Incremental Learning with Generative Classifiers.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 3611–3620.

[69] L. Pellegrini, G. Graffieti, V. Lomonaco, and D. Maltoni. “Latent replay for real-time continual learning.” In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2020, pp. 10203–10209.

[70] B. Zhang, Y. Guo, Y. Li, Y. He, H. Wang, and Q. Dai. “Memory recall: A simple neural network training framework against catastrophic forgetting.” *IEEE Transactions on Neural Networks and Learning Systems*, **33**(5), 2021, pp. 2010–2022.

[71] H. Jegou, M. Douze, and C. Schmid. “Product quantization for nearest neighbor search.” *IEEE transactions on pattern analysis and machine intelligence*, **33**(1), 2010, pp. 117–128.

[72] J. O’Neill, B. Pleydell-Bouverie, D. Dupret, and J. Csicsvari. “Play it again: reactivation of waking experience and memory.” *Trends in neurosciences*, **33**(5), 2010, pp. 220–229.

[73] P. A. Lewis and S. J. Durrant. “Overlapping memory replay during sleep builds cognitive schemata.” *Trends in cognitive sciences*, **15**(8), 2011, pp. 343–351.

[74] J.-B. Eichenlaub, B. Jarosiewicz, J. Saab, B. Franco, J. Kelemen, E. Halgren, L. R. Hochberg, and S. S. Cash. “Replay of learned neural firing sequences during rest in human motor cortex.” *Cell Reports*, **31**(5), 2020, p. 107581.

[75] J. L. McClelland, B. L. McNaughton, and R. C. O’Reilly. “Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory.” *Psychological review*, **102**(3), 1995, p. 419.

[76] J. Peng, D. Ye, B. Tang, Y. Lei, Y. Liu, and H. Li. “Lifelong Learning With Cycle Memory Networks.” *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[77] V. Lomonaco and D. Maltoni. “CORe50: a new dataset and benchmark for continuous object recognition.” In: *Conference on Robot Learning*. PMLR. 2017, pp. 17–26.

[78] G. Pasquale, C. Ciliberto, L. Rosasco, and L. Natale. “Object identification from few examples by improving the invariance of a deep convolutional neural network.” In: *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE. 2016, pp. 4904–4911.

[79] A. Krizhevsky, G. Hinton, et al. “Learning multiple layers of features from tiny images.” 2009.

[80] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. “ImageNet: A large-scale hierarchical image database.” In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[81] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size." *arXiv preprint arXiv:1602.07360*, 2016.

[82] Y. Liu, B. Schiele, and Q. Sun. "Adaptive aggregation networks for class-incremental learning." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 2544–2553.

[83] M. De Lange and T. Tuytelaars. "Continual prototype evolution: Learning online from non-stationary data streams." *arXiv preprint arXiv:2009.00919*, 2020.

[84] S. I. Mirzadeh, M. Farajtabar, R. Pascanu, and H. Ghasemzadeh. "Understanding the role of training regimes in continual learning." *arXiv preprint arXiv:2006.06958*, 2020.

[85] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778.

[86] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. "PyTorch: An imperative style, high-performance deep learning library." *arXiv preprint arXiv:1912.01703*, 2019.

[87] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars. "Memory aware synapses: Learning what (not) to forget." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 139–154.

[88] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. "How transferable are features in deep neural networks?" *arXiv preprint arXiv:1411.1792*, 2014.

[89] Y. Chen, M. Welling, and A. Smola. "Super-samples from kernel herding." *arXiv preprint arXiv:1203.3472*, 2012.

[90] P. W. Koh and P. Liang. "Understanding black-box predictions via influence functions." In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 1885–1894.

[91] P. Prabhanjan Brahma and A. Othon. "Subset Replay Based Continual Learning for Scalable Improvement of Autonomous Systems." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2018.

[92] S. Ho, M. Liu, L. Du, L. Gao, and Y. Xiang. "Prototype-Guided Memory Replay for Continual Learning." *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[93] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." In: *International Conference on Learning Representations*. 2019.

[94] Z. Shi, Y. Sun, J. H. Lim, and M. Zhang. "On the Robustness, Generalization, and Forgetting of Shape-Texture Debiased Continual Learning." *arXiv preprint arXiv:2211.11174*, 2022.

[95] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. “Mobilenetv2: Inverted residuals and linear bottlenecks.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2018, pp. 4510–4520.

[96] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. “Attention is all you need.” *arXiv preprint arXiv:1706.03762*, 2017.

[97] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. “An image is worth 16x16 words: Transformers for image recognition at scale.” In: *International Conference on Learning Representations.* 2021.

[98] M. Schrimpf et al. “Brain-Score: Which artificial neural network for object recognition is most brain-like?” *BioRxiv*, 2018, p. 407007.

[99] A. Mądry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. “Towards deep learning models resistant to adversarial attacks.” In: *International Conference on Learning Representations.* 2018.

[100] C. Guo, M. J. Lee, G. Leclerc, J. Dapello, Y. Rao, A. Mądry, and J. J. Dicarlo. “Adversarially trained neural representations are already as robust as biological neural representations.” In: *International Conference on Machine Learning.* 2022.

[101] G. Gaziv, M. Lee, and J. J. DiCarlo. “Strong and precise modulation of human percepts via robustified ANNs.” In: *Advances in Neural Information Processing Systems.* Vol. 36. 2023.

[102] F. Croce and M. Hein. “Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks.” In: *International Conference on Machine Learning.* 2020.

[103] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum. “One shot learning of simple visual concepts.” In: *Proceedings of the Annual Meeting of the Cognitive Science Society.* Vol. 33. 2011.

[104] Z.-L. Lu and B. A. Dosher. “Current directions in visual perceptual learning.” *Nature Reviews Psychology*, **1**(11), 2022, pp. 654–668.

[105] X. Wang, Y. Chen, and W. Zhu. “A survey on curriculum learning.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**(9), 2021, pp. 4555–4576.

[106] Z. Jiang, C. Zhang, K. Talwar, and M. C. Mozer. “Characterizing Structural Regularities of Labeled Data in Overparameterized Models.” In: *Proceedings of the 38th International Conference on Machine Learning.* Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 5034–5044.

[107] R. Baldock, H. Maennel, and B. Neyshabur. “Deep learning through the lens of example difficulty.” In: *Advances in Neural Information Processing Systems.* Vol. 34. 2021, pp. 10876–10889.

[108] D. Mayo, J. Cummings, X. Lin, D. Gutfreund, B. Katz, and A. Barbu. “How hard are computer vision datasets? Calibrating dataset difficulty to viewing time.” In: *Advances in Neural Information Processing Systems.* Vol. 36. 2023, pp. 11008–11036.

[109] Y. Qi, Z. Yang, W. Sun, M. Lou, J. Lian, W. Zhao, X. Deng, and Y. Ma. "A comprehensive overview of image enhancement techniques." *Archives of Computational Methods in Engineering*, 2021, pp. 1–25.

[110] K. Zuiderveld. "Contrast limited adaptive histogram equalization." In: *Graphics Gems IV*. Academic Press, 1994, pp. 474–485.

[111] D. J. Jobson, Z.-u. Rahman, and G. 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), 1997, pp. 965–976.

[112] S. Anwar, S. Khan, and N. Barnes. "A deep journey into super-resolution: A survey." *ACM Computing Surveys (CSUR)*, **53**(3), 2020, pp. 1–34.

[113] J. Kim, J.-H. Choi, S. Jang, and J.-S. Lee. "Amicable Aid: Perturbing Images to Improve Classification Performance." In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[114] J. Tussupov, K. Kozhabai, A. Bayegizova, L. Kassenova, Z. Manbetova, N. Glazyrina, M. Bersugir, and M. Yeginbayev. "APPLYING MACHINE LEARNING TO IMPROVE A TEXTURE TYPE IMAGE." *Eastern-European Journal of Enterprise Technologies*, **122**(2), 2023.

[115] H. Salman, A. Ilyas, L. Engstrom, S. Vemprala, A. Mądry, and A. Kapoor. "Unadversarial examples: Designing objects for robust vision." In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 15270–15284.

[116] I. Frosio and J. 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*. 2023, pp. 4067–4076.

[117] M. Choi, K. Han, X. Wang, Y. Zhang, and Z. Liu. "A dual-stream neural network explains the functional segregation of dorsal and ventral visual pathways in human brains." In: *Advances in Neural Information Processing Systems*. Vol. 36. 2023, pp. 50408–50428.

[118] P. Bomatter, M. Zhang, D. Karev, S. Madan, C. Tseng, and G. Kreiman. "When pigs fly: contextual reasoning in synthetic and natural scenes." In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 255–264.

[119] S. K. Gupta, M. Zhang, C.-C. Wu, J. Wolfe, and G. Kreiman. "Visual search asymmetry: deep nets and humans share similar inherent biases." In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 34. 2021, pp. 6946–6959.

[120] R. Ganz, B. Kawar, and M. Elad. "Do perceptually aligned gradients imply robustness?" In: *International Conference on Machine Learning*. PMLR, 2023, pp. 10628–10648.

[121] X. Zhu. "Machine teaching: An inverse problem to machine learning and an approach toward optimal education." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. 1. 2015.

[122] W. Liu, B. Dai, A. Humayun, C. Tay, C. Yu, L. B. Smith, J. M. Rehg, and L. Song. “Iterative machine teaching.” In: *International Conference on Machine Learning*. PMLR, 2017, pp. 2149–2158.

[123] Z. Qiu, W. Liu, T. Z. Xiao, Z. Liu, U. Bhatt, Y. Luo, A. Weller, and B. Schölkopf. “Iterative Teaching by Data Hallucination.” In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. Ed. by F. Ruiz, J. Dy, and J.-W. van de Meent. Vol. 206. Proceedings of Machine Learning Research. PMLR, 2023, pp. 9892–9913.

[124] A. Singla, I. Bogunovic, G. Bartók, A. Karbasi, and A. Krause. “Near-optimally teaching the crowd to classify.” In: *International Conference on Machine Learning*. PMLR, 2014, pp. 154–162.

[125] E. Johns, O. Mac Aodha, and G. J. Brostow. “Becoming the expert-interactive multi-class machine teaching.” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 2616–2624.

[126] P. Wang, K. Nagrecha, and N. Vasconcelos. “Gradient-based algorithms for machine teaching.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 1387–1396.

[127] O. Mac Aodha, S. Su, Y. Chen, P. Perona, and Y. Yue. “Teaching categories to human learners with visual explanations.” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3820–3828.

[128] D. Chang, K. Pang, R. Du, Y. Tong, Y.-Z. Song, Z. Ma, and J. Guo. “Making a bird AI expert work for you and me.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**(10), 2023, pp. 12068–12084.

[129] G. Van Horn, E. Cole, S. Beery, K. Wilber, S. Belongie, and O. Mac Aodha. “Benchmarking representation learning for natural world image collections.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 12884–12893.

[130] P. Tschandl, C. Rosendahl, and H. Kittler. “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions.” *Scientific Data*, **5**(1), 2018, pp. 1–9.

[131] J. Wei 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*. Springer, 2021, pp. 11–24.

[132] A. B. Petro, C. Sbert, and J.-M. Morel. “Multiscale Retinex.” *Image Processing On Line*, 2014, pp. 71–88.

[133] Adobe Inc. *Adobe Photoshop Lightroom Classic*. Version 13.5.1. Version 13.5.1. 2024. URL: https://www.adobe.com/products/photoshop-lightroom-classic.html.

[134] A. Ali et al. “XCiT: Cross-covariance image transformers.” In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 20014–20027.

[135] J. S. Hufnagel. “Tabelle von den Nachtvögeln.” *Berlinisches Magazin*, **1**(4), 1767, p. 618. URL: http://resolver.sub.uni-goettingen.de/purl?PPN484874233_0004.

[136] H. Tsao et al. "Early detection of melanoma: reviewing the ABCDEs." *Journal of the American Academy of Dermatology*, **72**(4), 2015, pp. 717–723.

[137] E. Mettler and P. J. Kellman. "Adaptive response-time-based category sequencing in perceptual learning." *Vision Research*, **99**, 2014, pp. 111–123.

[138] R. Daneshjou et al. "Disparities in dermatology AI performance on a diverse, curated clinical image set." *Science Advances*, **8**(31), 2022, eabq6147.

[139] J. R. De Leeuw. "jsPsych: A JavaScript library for creating behavioral experiments in a Web browser." *Behavior Research Methods*, **47**, 2015, pp. 1–12.

[140] D. Kuroki. "A new jsPsych plugin for psychophysics, providing accurate display duration and stimulus onset asynchrony." *Behavior Research Methods*, **53**, 2021, pp. 301–310.

[141] J. N. Cormier, Y. Xing, M. Ding, J. E. Lee, P. F. Mansfield, J. E. Gershenwald, M. I. Ross, and X. L. Du. "Ethnic differences among patients with cutaneous melanoma." *Archives of Internal Medicine*, **166**(17), 2006, pp. 1907–1914.

[142] B. Thompson, T. Jenkins, J. P. Sánchez, M. Frederick, A. Posligua-Alban, and N. S. Barbosa. "Melanoma: Does It Present Differently in Darker Skin Tones?" *MedEdPORTAL*, **19**, 2023, p. 11311.

[143] K. Stacke, G. Eilertsen, J. Unger, and C. Lundström. "Measuring domain shift for deep learning in histopathology." *IEEE journal of biomedical and health informatics*, **25**(2), 2020, pp. 325–336.

[144] J. Futoma, M. Simons, T. Panch, F. Doshi-Velez, and L. A. Celi. "The myth of generalisability in clinical research and machine learning in health care." *The Lancet Digital Health*, **2**(9), 2020, e489–e492.

[145] A. Vaidya, R. J. Chen, D. F. Williamson, A. H. Song, G. Jaume, Y. Yang, T. Hartvigsen, E. C. Dyer, M. Y. Lu, J. Lipkova, et al. "Demographic bias in misdiagnosis by computational pathology models." *Nature Medicine*, **30**(4), 2024, pp. 1174–1190.

[146] M. A. Berbís, D. S. McClintock, A. Bychkov, J. Van der Laak, L. Pantanowitz, J. K. Lennerz, J. Y. Cheng, B. Delahunt, L. Egevad, C. Eloy, et al. "Computational pathology in 2030: a Delphi study forecasting the role of AI in pathology within the next decade." *EBioMedicine*, **88**, 2023.

[147] E. Walsh and N. M. Orsi. "The current troubled state of the global pathology workforce: a concise review." *Diagnostic Pathology*, **19**(1), 2024, p. 163.

[148] T. T. Brunyé, T. Drew, D. L. Weaver, and J. G. Elmore. "A review of eye tracking for understanding and improving diagnostic interpretation." *Cognitive research: principles and implications*, **4**(1), 2019, p. 7.

[149] T. T. Brunyé, A. Balla, T. Drew, J. G. Elmore, K. F. Kerr, H. Shucard, and D. L. Weaver. "From image to diagnosis: characterizing sources of error in histopathologic interpretation." *Modern Pathology*, **36**(7), 2023, p. 100162.

[150] T. T. Brunyé, H. Shucard, M. M. Eguchi, T. Drew, J. Marotti, L. C. Lomo, M. R. Kilgore, K. H. Allison, K. F. Kerr, J. G. Elmore, et al. "Characterizing Interpretive and Diagnostic Competency Development in Pathology Training." *Modern Pathology*, 2025, p. 100802.

[151] S. Krasne, J. D. Hillman, P. J. Kellman, and T. A. Drake. "Applying perceptual and adaptive learning techniques for teaching introductory histopathology." *Journal of pathology informatics*, **4**(1), 2013, p. 34.

[152] P. J. Kellman, V. Jacoby, C. Massey, and S. Krasne. "Perceptual learning, adaptive learning, and gamification: Educational technologies for pattern recognition, problem solving, and knowledge retention in medical learning." In: *Technologies in biomedical and life sciences education: approaches and evidence of efficacy for learning*. Springer, 2022, pp. 135–166.

[153] P. L. Mishall, W. Burton, and M. Risley. "Flashcards: the preferred online game-based study tool self-selected by students to review medical histology image content." *Biomedical Visualisation: Volume 15–Visualisation in Teaching of Biomedical and Clinical Subjects: Anatomy, Advanced Microscopy and Radiology*, 2023, pp. 209–224.

[154] H. Pashler and M. C. Mozer. "When does fading enhance perceptual category learning?" *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **39**(4), 2013, p. 1162.

[155] T. Miyatsu, R. Gouravajhala, R. M. Nosofsky, and M. A. McDaniel. "Feature highlighting enhances learning of a complex natural-science category." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **45**(1), 2019, p. 1.

[156] M. B. Talbot, G. Kreiman, J. J. DiCarlo, and G. Gaziv. "L-WISE: Boosting Human Visual Category Learning Through Model-Based Image Selection And Enhancement." In: *International Conference on Learning Representations*. 2025.

[157] R. J. Chen, T. Ding, M. Y. Lu, D. F. Williamson, G. Jaume, B. Chen, A. Zhang, D. Shao, A. H. Song, M. Shaban, et al. "Towards a General-Purpose Foundation Model for Computational Pathology." *Nature Medicine*, 2024.

[158] N. Brancati, A. M. Anniciello, P. Pati, D. Riccio, G. Scognamiglio, G. Jaume, G. De Pietro, M. Di Bonito, A. Foncubierta, G. Botti, et al. "BRACS: A dataset for breast carcinoma subtyping in h&e histology images." *Database*, **2022**, 2022, baac093.

[159] M. Koziarski, B. Cyganek, P. Niedziela, B. Olborski, Z. Antosz, M. Żydak, B. Kwolek, P. Wąsowicz, A. Bukała, J. Swadźba, et al. "DiagSet: a dataset for prostate cancer histopathological image classification." *Scientific Reports*, **14**(1), 2024, p. 6780.

[160] E. Josephs, T. Drew, and J. Wolfe. "Shuffling your way out of change blindness." *Psychonomic bulletin & review*, **23**(1), 2016, pp. 193–200.

[161] T. Drew, A. M. Aizenman, M. B. Thompson, M. D. Kovacs, M. Trambert, M. A. Reicher, and J. M. Wolfe. "Image toggling saves time in mammography." *Journal of Medical Imaging*, **3**(1), 2016, pp. 011003–011003.

[162] D. B. Rubin. *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, 1987.

[163] T. Zadvornyi, N. Lukianova, O. Mushii, A. Pavlova, O. Voronina, and V. Chekhun. "Benign and malignant prostate neoplasms show different spatial organization of collagen." *Croatian Medical Journal*, **64**(6), 2023, pp. 413–420.

[164] K. M. Curby and I. Gauthier. "The temporal advantage for individuating objects of expertise: Perceptual expertise is an early riser." *Journal of Vision*, **9**(6), 2009, pp. 7–7.

[165] R. A. Bjork and E. L. Bjork. "Desirable difficulties in theory and practice." *Journal of Applied research in Memory and Cognition*, **9**(4), 2020, p. 475.

[166] R. V. Lindsey, M. C. Mozer, W. J. Huggins, and H. Pashler. "Optimizing instructional policies." *Advances in Neural Information Processing Systems*, **26**, 2013.

[167] M. J. Lee and J. J. DiCarlo. "How well do rudimentary plasticity rules predict adult visual object learning?" *PLOS Computational Biology*, **19**(12), 2023, e1011713.

[168] N. Slater. "Eddington's Fundamental Theory." *Physics Bulletin*, **12**(6), 1961, p. 157.

[169] F. G. Ashby and V. V. Valentin. "Multiple systems of perceptual category learning: Theory and cognitive tests." In: *Handbook of Categorization in Cognitive Science*. Elsevier, 2017, pp. 157–188.

[170] C. Zhang, X. Cao, W. Liu, I. Tsang, and J. Kwok. "Nonparametric iterative machine teaching." In: *International Conference on Machine Learning*. PMLR. 2023, pp. 40851–40870.

[171] K. R. Patil, X. Zhu, Ł. Kopeć, and B. C. Love. "Optimal teaching for limited-capacity human learners." *Advances in Neural Information Processing Systems*, **27**, 2014.

[172] A. N. Rafferty, E. Brunskill, T. L. Griffiths, and P. Shafto. "Faster teaching via pomdp planning." *Cognitive Science*, **40**(6), 2016, pp. 1290–1332.

[173] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. "Caffe: Convolutional architecture for fast feature embedding." In: *Proceedings of the 22nd ACM international conference on Multimedia*. 2014, pp. 675–678.

[174] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks." In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[175] T. J. Barry, J. W. Griffith, S. De Rossi, and D. Hermans. "Meet the Fribbles: Novel stimuli for use within behavioural research." *Frontiers in Psychology*, **5**, 2014, p. 103.

[176] J. Dapello, K. Kar, M. Schrimpf, R. Geary, M. Ferguson, D. D. Cox, and J. J. DiCarlo. "Aligning model and macaque inferior temporal cortex representations improves model-to-human behavioral alignment and adversarial robustness." *BioRxiv*, 2022, pp. 2022–07.

[177] L. Tausani, P. Muratore, M. B. Talbot, G. Amerio, G. Kreiman, and D. Zoccolan. "Stretching Beyond the Obvious: A Gradient-Free Framework to Unveil the Hidden Landscape of Visual Invariance." In: *International Conference on Learning Representations*. 2026.

[178] B. Le Lan. "Robust convolutional neural networks as models of primate vision." MA thesis. EPFL Lausanne, Switzerland, 2024.

[179] D. E. Goldberg and J. Richardson. "Genetic algorithms with sharing for multimodal function optimization." In: *Genetic algorithms and their applications: Proceedings of the Second International Conference on Genetic Algorithms*. Vol. 4149. Lawrence Erlbaum, Hillsdale, NJ. 1987, pp. 414–425.

[180] A. Brindle. "Genetic algorithms for function optimization." PhD thesis. University of Alberta, 1980.

[181] D. E. Goldberg and K. Deb. "A comparative analysis of selection schemes used in genetic algorithms." In: *Foundations of Genetic Algorithms*. Vol. 1. Elsevier, 1991, pp. 69–93.

[182] D. E. Goldberg and R. Lingle. "Alleles, loci, and the traveling salesman problem." In: *Proceedings of the First International Conference on Genetic Algorithms and Their Applications*. Psychology Press. 1985, pp. 154–159.

[183] K. A. De Jong. "An analysis of the behavior of a class of genetic adaptive systems." PhD thesis. Ann Arbor, MI: University of Michigan, 1975.

[184] F. G. Ashby, L. A. Alfonso-Reese, E. M. Waldron, et al. "A neuropsychological theory of multiple systems in category learning." *Psychological Review*, **105**(3), 1998, p. 442.

[185] B. J. Knowlton, J. A. Mangels, and L. R. Squire. "A neostriatal habit learning system in humans." *Science*, **273**(5280), 1996, pp. 1399–1402.

[186] D. Shohamy, C. Myers, S. Grossman, J. Sage, M. Gluck, and R. Poldrack. "Cortico-striatal contributions to feedback-based learning: Converging data from neuroimaging and neuropsychology." *Brain*, **127**(4), 2004, pp. 851–859.

[187] J. V. Filoteo, W. T. Maddox, D. P. Salmon, and D. D. Song. "Information-integration category learning in patients with striatal dysfunction." *Neuropsychology*, **19**(2), 2005, p. 212.

[188] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. "Learning transferable visual models from natural language supervision." In: *International Conference on Machine Learning*. PmLR. 2021, pp. 8748–8763.

[189] S. J. Luck and E. K. Vogel. "The capacity of visual working memory for features and conjunctions." *Nature*, **390**(6657), 1997, pp. 279–281.

[190] S. J. Luck and E. K. Vogel. "Visual working memory capacity: from psychophysics and neurobiology to individual differences." *Trends in Cognitive Sciences*, **17**(8), 2013, pp. 391–400.

[191] E. M. Waldron and F. G. Ashby. "The effects of concurrent task interference on category learning: Evidence for multiple category learning systems." *Psychonomic Bulletin & Review*, **8**(1), 2001, pp. 168–176.

[192] F. G. Ashby, S. W. Ell, and E. M. Waldron. "Procedural learning in perceptual categorization." *Memory & Cognition*, **31**(7), 2003, pp. 1114–1125.

[193] J. P. Minda, C. L. Roark, P. Kalra, and A. Cruz. "Single and multiple systems in categorization and category learning." *Nature Reviews Psychology*, **3**(8), 2024, pp. 536–551.

[194] M. S. DeCaro, R. D. Thomas, and S. L. Beilock. "Individual differences in category learning: Sometimes less working memory capacity is better than more." *Cognition*, **107**(1), 2008, pp. 284–294.

[195] B. J. Spiering and F. G. Ashby. "Initial training with difficult items facilitates information integration, but not rule-based category learning." *Psychological science*, **19**(11), 2008, pp. 1169–1177.

[196] A. S. Benjamin and J. Tullis. "What makes distributed practice effective?" *Cognitive Psychology*, **61**(3), 2010, pp. 228–247.

[197] C. L. Roark and B. Chandrasekaran. "Stable, flexible, common, and distinct behaviors support rule-based and information-integration category learning." *npj Science of Learning*, **8**(1), 2023, p. 14.

[198] D. L. Yamins and J. J. DiCarlo. "Using goal-driven deep learning models to understand sensory cortex." *Nature Neuroscience*, **19**(3), 2016, pp. 356–365.

[199] G. Van Horn and O. Mac Aodha. *iNat Challenge 2021 - FGVC8*. 2021. URL: https://kaggle.com/competitions/inaturalist-2021.

[200] L. Engstrom, A. Ilyas, H. Salman, S. Santurkar, and D. Tsipras. *Robustness (Python Library)*. 2019. URL: https://github.com/MadryLab/robustness.

[201] J. A. Chapman and J. E. C. Flux. "Introduction to the Lagomorpha." In: *Lagomorph Biology: Evolution, Ecology, and Conservation*. Springer, 2008, pp. 1–9.

[202] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo. "CutMix: Regularization strategy to train strong classifiers with localizable features." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 6023–6032.

[203] E. Debenedetti, V. Sehwag, and P. Mittal. "A light recipe to train robust vision transformers." In: *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, 2023, pp. 225–253.

[204] A. Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." In: *International Conference on Learning Representations*. 2020.

[205] R. Ganz and M. Elad. "CLIPAG: Towards generator-free text-to-image generation." In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024, pp. 3843–3853.

[206] S. Zhao, Z. Liu, J. Lin, J.-Y. Zhu, and S. Han. "Differentiable augmentation for data-efficient GAN training." In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 7559–7570.

[207] O. Russakovsky et al. "ImageNet large scale visual recognition challenge." *International Journal of Computer Vision*, **115**, 2015, pp. 211–252.

[208] B. D. Rowley. “AMA—Fellowship and Residency Electronic Interactive Database Access (AMA-FREIDA): A Computerized Residency Selection Tool.” *JAMA*, **260**(8), 1988, pp. 1059–1059.

[209] P. J. Kellman and C. M. Massey. “Chapter Four - Perceptual Learning, Cognition, and Expertise.” In: *Psychology of Learning and Motivation*. Ed. by B. H. Ross. Vol. 58. Psychology of Learning and Motivation. Academic Press, 2013, pp. 117–165. DOI: https://doi.org/10.1016/B978-0-12-407237-4.00004-9. URL: https://www.sciencedirect.com/science/article/pii/B9780124072374000049.