A Discussion

Our benchmark brings a new perspective to classification, as we not only seek models that predict accurately, but also predict for the right reasons. Assessing model performance using accuracy alone can obscure key misconceptions held by models, which may only become apparent when models are deployed to new domains at test time. Moreover, design decisions such as training strategy and architecture may affect the degree to which spurious features are relied upon, as observed in [26]; this dataset and accompanying benchmark can reveal these model differences. Finally, we emphasize the need to understand model behavior under "bad" data; that is, images where the object of interest is not centered or large, unlike most cases. With models becoming increasingly data hungry, it is inevitable that some portion of the data will not capture objects in ideal conditions. Further, certain objects simply are not well suited to be captured prominently (i.e. large and centered) in square photos. Figuring out how to learn to recognize objects from these suboptimal data conditions will be an important challenge to extend the impressive performance of deep classifiers from standard datasets to many more realistic settings.

With Hard ImageNet, the community can evaluate the capacity of any ImageNet trained model to faithfully learn challenging objects, and also explore how going beyond single class label annotations can lead to improved image classifiers. While segmentation masks are expensive to collect, procedures that are much more automated already exist [35], and we envision newer ones are likely to emerge with time. Also, the procedure with which we ranked images was largely automated, indicating that these types of annotations are by no means prohibitively expensive. We hope Hard ImageNet can lead to new perspectives on both training and evaluation paradigms for image classification.

B Distinguishing Hard ImageNet from Related Challenge Datasets

Our work is inspired by other challenge datasets that focus on improving deep classifiers by aggregating edge cases where usually strong performance falters. We highlight two datasets in particular: ObjectNet [3] and ImageNet-A [16]. Both of these datasets include samples where spurious correlations are broken, leading to dramatically lower accuracy. Further, these datasets consist of clean images, as opposed to other challenge sets that make synthetic changes [40, 30].

We now outline some key distinctions between Hard ImageNet and these datasets. First, ObjectNet and ImageNet-A only contain test sets. We include a training set in Hard ImageNet because the central goal of our work is for the community to develop new algorithms that can learn to recognize objects without relying on spurious cues, even when the spurious signals are very strong in the training data. To this end, we also introduce two new forms of annotation (object segmentation and image ranking), with hopes of challenging the community to explore training paradigms beyond single-label supervision. Lastly, model performance on ObjectNet and ImageNet-A is evaluated using *accuracy*. In contrast, we present three alternative evaluation metrics leveraging the richer annotations of Hard ImageNet. In spirit, our evaluation is orthogonal to the traditional metric of accuracy, as we shift the focus from *what* models predict to *how* they predict. We believe that the reliability and trustworthiness of deep models hinges on their use of appropriate reasoning structures. That is, if a model predicts correctly but for the wrong reasons, the model may act erratically when deployed.

We greatly value the inspiring work of these earlier challenge datasets and recognize the similarities of their work to our contribution, though we believe that Hard ImageNet may open the door to understanding deep classifier performance, specifically with respect to spurious feature reliance, in a new light.

C Improving Models with Hard ImageNet Annotations

In this section, we begin the exploration of harnessing Hard ImageNet's annotations for improved model classification. Namely, we leverage object segmentations and image rankings to reduce model reliance on spurious features while performing Hard ImageNet classification. We focus our study on finetuned models pretrained on ImageNet, using ResNet50 and DeiT (Small) as in Section 4. We keep features fixed during finetuning, only optimizing the parameters of a new final layer for the 15-way Hard ImageNet classification.

We employ two approaches for mitigating spurious feature reliance. [35] propose **Core Risk Minimization** (CoRM) as an alternative to ERM when segmentations of core (i.e. not spurious) regions are available; Hard ImageNet's object segmentations fulfill this prerequisite. Specifically, the objective of CoRM is to minimize classification error over the distribution of images *with noise applied to non-core regions*, so that the optimal classifier predicts correctly even when spurious features are corrupted. In that work, *random noising*, where small amounts of Gaussian noise are added to non-core regions with probability $p = 0.5$, and *saliency regularization*, where the ℓ_2 norm of the gradient on non-core pixels is added to the classification loss, were applied in tandem to improve relative core sensitivity (an analagous metric to RFS). [20] propose **Deep Feature Reweighting** (DFR), in which retraining a final linear layer using a *balanced dataset* reduces spurious feature reliance. The balanced dataset consists of a subset of the training data containing an equal portion of samples with and samples without spurious features, essentially breaking spurious correlations that impede generalization to minority groups. Using Hard ImageNet's image rankings, we extract the top and bottom 100 images for each class to form the spurious-balanced subset.

Method		Ablation Accuracies (↓)				RFS (↑)		Saliency (↑)
CoRM	DFR	None (↑)	Gray	Gray BBox	Tile	$\sigma = 0.25$	$\sigma = 0.5$	IoU
		Finetuned DeiT (Small)						
✗	✗	**96.79**	84.22	80.48	81.15	−0.19	−0.35	20.90
✓	✗	96.39	**81.02**	78.74	80.75	**0.02**	**−0.19**	21.57
✗	✓	96.66	81.28	**77.01**	77.94	−0.20	−0.33	21.63
✓	✓	96.52	82.35	**77.01**	**77.81**	−0.10	−0.29	**21.99**
		Finetuned ResNet50						
✗	✗	94.25	75.94	69.39	67.38	−0.18	**−0.27**	18.44
✓	✗	92.91	76.20	69.12	68.32	**−0.08**	**−0.27**	**20.43**
✗	✓	**94.39**	73.53	67.51	66.71	−0.27	−0.35	18.39
✓	✓	91.31	**72.59**	**63.64**	**63.90**	−0.23	−0.31	20.35

Table 1: Final layer retraining improves faithful learning on Hard ImageNet. Results shown for entire benchmark under two different training approaches: i) Core Risk Minimization (CoRM) via *random background noising* and *saliency regularization*, and ii) deep feature reweighting (DFR) using a *spurious-balanced training subset*. We also report results for the combination of the two approaches and ordinary finetuning (as a baseline) under two architectures. Relative Foreground Sensitivity (RFS) is evaluated under two ℓ_∞ noise levels, indicated by σ. Saliency refers to *saliency alignment* as measured by intersection over union (IoU).

Table 1 shows that these two methods can considerably reduce model reliance on spurious features, improving numbers across all metrics in our benchmark. Between the two approaches, CoRM appears to lead to more improvement in saliency alignment and RFS, while DFR yields beter results for ablation. Combining CoRM and DFR leads to even better performance with respect to accuracies under ablation. While improvements are at times small, we note that in these experiments, the vast majority of model parameters are left unchanged, as we only train a new final layer. We leave the door open to new approaches for improving the *faithful* learning of Hard ImageNet objects, including training models from scratch.

D Evaluation of Additional Pretrained Models

In addition to the transformer and convolutional neural networks (DeiT and ResNet50) explored in the main text, we evaluate four other deep classifiers. Namely, we investigate Swin Transformer[24], ConViT[6], DenseNet161[18], and VGG16 [33]. As seen in table 2, our results on new models corroborate the findings of the main text. Specifically, we see that across models, classifying Hard ImageNet objects leads to higher accuracy under ablation, lower RFS scores, and lower saliency alignment, compared to classifying RIVAL20 objects. This implies that certain properties inherent to the data in Hard ImageNet makes it far more challenging to learn to classify without heavily relying on spurious cues.

Model	Ablation Accuracies (↓)				RFS (↑)	Saliency (↑)
	None (↑)	Gray	Gray BBox	Tile	$\sigma = 0.25$	IoU
Hard ImageNet						
Swin	80.59	61.19	59.97	59.30	-0.01	4.28
Convit	79.92	60.11	59.97	55.93	-0.12	22.37
Densenet161	57.68	37.06	30.05	29.65	-0.26	18.10
Vgg16	71.83	46.63	41.37	42.86	-0.53	16.80
RIVAL20						
Swin	86.96	30.13	25.18	18.50	0.44	6.64
Convit	85.74	21.64	25.28	16.68	0.31	35.89
Densenet161	78.67	7.58	3.03	2.73	0.40	43.95
Vgg16	76.74	6.88	3.03	3.44	0.80	30.08
Hard ImageNet - RIVAL20						
Swin	-6.36	31.05	34.80	40.80	-0.45	-2.36
Convit	-5.82	38.47	34.69	39.25	-0.43	-13.53
Densenet161	-20.98	29.48	27.02	26.92	-0.67	-25.85
Vgg16	-4.91	39.76	38.34	39.42	-1.32	-13.27

Table 2: Evaluation of additional pretrained models (no finetuning). All models have higher accuracies under ablation, lower RFS scores, and lower saliency alignment on Hard ImageNet than RIVAL20.

E Overview of Salient ImageNet

We refer readers to [35] for all details related to the Salient ImageNet-1M data and collection procedure. For completeness, we offer brief discussion of the methods relevant to this paper. Namely, we elaborate on the way in which class-feature pairs were annotated as core or spurious (i.e. a neural feature was annotated as detecting input regions that were spurious with respect to the given class label). Recall that the motivation for closer inspection of Hard ImageNet classes was that all class-feature pairs for Hard ImageNet were annotated as spurious.

Salient ImageNet annotations first correspond to labeling 5 neural features as core or spurious for each of the 1000 ImageNet classes, resulting in 5000 class-feature pair binary annotations (core or spurious). Neural feature refers to the nodes in the penultimate layer of a deep classifier. Specifically, the neural features of an ℓ_2 adversarially trained ResNet50 were inspected because adversarially robust models have been observed to be more interpretable. For each class, the five neural features annotated were those that contribute the most to the logit of the given class. The average contribution of a neural feature to a class can easily be computed as the product of the average feature activation and the weight of the linear layer connecting the feature to the class logit.

Of the 5000 class-feature pairs annotated, 4370 (87.4%) were deemed to be core, signifying that in most cases, the model effectively learned to use the appropriate features in classification. For 342 classes, at least one feature was annotated as spurious. However, only a small minority (the 15 classes comprising Hard ImageNet) had all five features annotated as spurious. This motivated our hypothesis that there were inherent properties of the data in Hard ImageNet that leads standard supervised classification training algorithms to result in models that rely on spurious cues. We do not claim that the classes in Hard ImageNet have the strongest spurious cues, nor do we claim that models do not rely on spurious cues for classes outside of Hard ImageNet; only that strong spurious

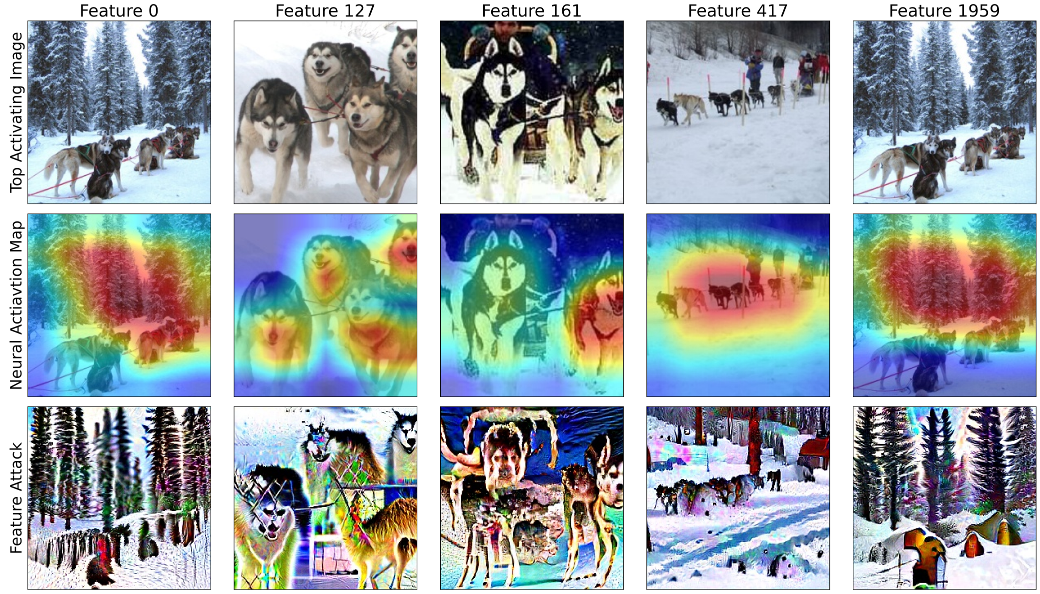


Figure 11: Example visualizations of the five most important neural features for the class **Dog Sled** used in Salient ImageNet annotation. From left to right, the features may be described as focusing on *trees, dogs, dogs, dogs in snow,* and *trees*.

cues exist in Hard ImageNet, and studying this data can yields insights related to causes and solutions for image classifier reliance on spurious features.

There are three key visualization techniques applied in order to reveal the function a neural feature serves over images from some class: natural images that highly activate the feature, ii. neural activation maps which highlight the input region responsible for the neural feature activation, and iii. feature attacks that optimize the input image to amplify feature activation. These visualizations are shown for the top five activating images per class-feature pair to five human annotators, who each vote to describe the focus of the feature as either on the main object (core) or a separate object or the background (spurious). The final annotation of the class feature pair is determined by majority vote.

We show the top activating image, its neural activation map, and a feature attack performed on it for each of the five features annotated for the Dog Sled and Patio classes in Figures 11 and 12 respectively. Visualizations for all Hard ImageNet classes (as well as the rest of ImageNet) can be viewed here: `www.salient-imagenet.cs.umd.edu`.

Lastly, we note that Salient ImageNet-1M also consists of soft segmentations masks for the objects for all images, *except* for those belonging to Hard ImageNet classes. This discrepancy is because the soft segmentation masks are constructed from the neural activation maps of core features. Thus, since Hard ImageNet classes have no annotated core features, Salient ImageNet-1M lacks segmentations for those classes. Therefore, the segmentations collected for the Hard ImageNet dataset effectively complete Salient ImageNet-1M. These masks can be potentially leveraged to train more reliable models, though this is an open research problem with little existing work, since annotations of this kind (segmentations) have not been prevalent for classification at scale until these recent works.

F Inspecting Prediction Confidence on Ablated Images

One may argue that it is unreasonable to fault a model for classifying an ablated image to its original class, particularly when it is not an option to predict some other more suitable class. After all, the model simply returns probabilities that an image belongs to each class, and chooses the class that is most likely. A similar metric would be to inspect prediction confidence instead of accuracy. This

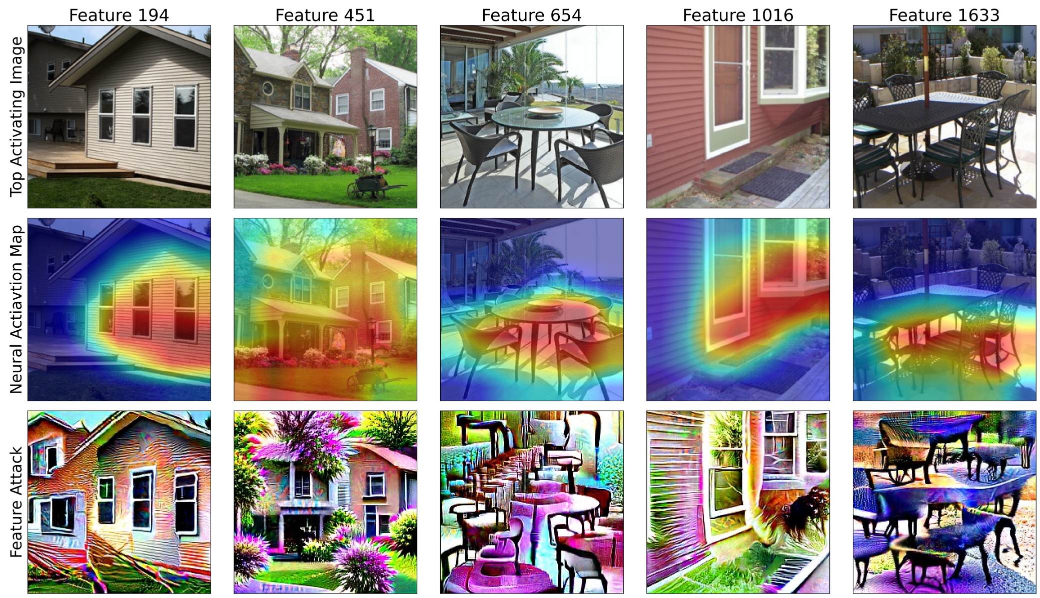


Figure 12: Example visualizations of the five most important neural features for the class **Patio** used in Salient ImageNet annotation. From left to right, the features may be described as focusing on *windows, house, furniture, window jambs,* and *patio chairs*.

way, we no longer directly fault a model for still predicting the original class, but still reward the calibration of a model. That is, it may be more reasonable to hope a model at least predicts an ablated image to the true class with far less confidence.

We explore this related metric in this section so to validate our ablation analyses using the more canonical (though potentially slightly more problematic) accuracy measure. Quite simply, we aggregate prediction confidences of the true class (not the predicted class) for each ablated image, and view the average confidence. As figure 13 shows, we very closely corroborate the findings obtained when inspecting accuracy under ablation.

While using accuracy under ablation directly may be imperfect, we find that it is an intuitive measure that may be more easily interpretable than our noise or saliency based metrics. Furthermore, accuracy is the standard evaluation metric for classification, and is highly correlated with true class prediction confidence, which as detailed above, reveals analogous findings and is less affected by the fact that desired classification behavior on ablated images is unclear. Thus, we present accuracy in the main text, though we argue that either accuracy or true class prediction confidence under ablation can be used in practice. We provide implementations for both metrics.

G Datasheet

We now share more detail on our dataset, following the *Datasheets for Datasets* protocol [10]. Access all code and data at the following link: `mmoayeri.github.io/HardImageNet`.

G.1 Motivation

Hard ImageNet was created to assess and improve image classifier capacity to learn to objects that commonly occur with strong spurious cues. We hypothesized that despite high classification accuracy, models were incorrectly learning the objects corresponding to Hard ImageNet classes. Going beyond single class-label annotations allowed for quantitative demonstration of this undesirable (and otherwise undetectable) behavior, as well as opening the door to new ways of improving models on these challenging objects. The dataset was created by academics (namely from the University of

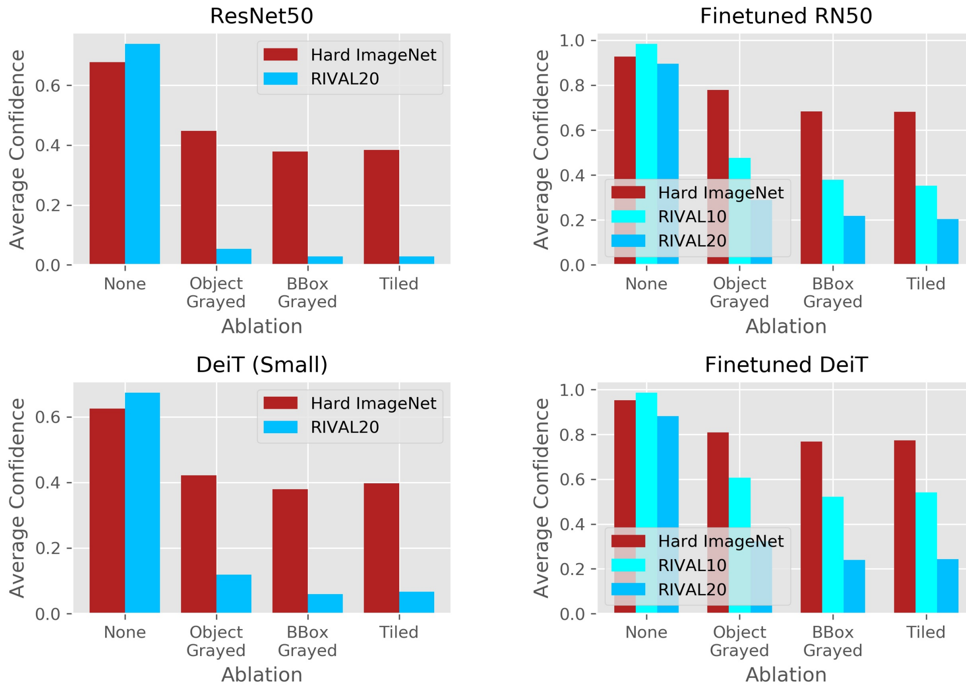


Figure 13: Probability (confidence) of true class under ablation. Confidence drops much less when Hard ImageNet objects are ablated than when RIVAL10 objects are ablated, exactly as observed for accuracy under ablation.

Maryland) for academic purposes, leveraging crowd annotations through Amazon's Mechanical Turk platform. Data collection was funded by an AWS Machine Learning Research Award.

G.2 Composition

Each instance consists of an image with a label and a binary segmentation mask corresponding to the class object. Instances either fall in the training or validation split, which is consistent with ImageNet's split. Images in the training split additionally are ranked within their respective class by the strength of spurious cues present, as determined in an automated procedure leveraging the neural feature annotations of Salient ImageNet-1M [35]. Instances in the validation split are unlikely to be noisy, as they consolidate five separate repetitions of annotations, while training set segmentations may be noisier, though quality is generally ensured via qualification exams and attention checks. Generally, the dataset does not relate to people, though many images do contain people (in fact, they are a common spuriuos cue). It is unlikely that the data can be used to identify any individuals are subpopulations, and we note that these pitfalls are inherited from the standard benchmark dataset ImageNet, from where Hard ImageNet images are drawn. Nonetheless, we advise caution in using and sharing images containing faces or otherwise prominently displaying people; we attempted to avoid including such images in our figures to the best of our ability.

G.3 Collection Process

Images were drawn directly from ImageNet. Segmentations were collected via Amazon Mechanical Turk. Image rankings were computed by inspecting the activations of particular neural features in a ℓ_2 adversarially trained ResNet50 using an attack budget of $\epsilon = 3$ (see [37] for more detail). Segmentations were validated in the sense that quality was monitored via attention checks, where annotators consistently achieved high IoUs with ground truth segmentations (average IoU of 0.76). Validation set segmentations were validated across one another, by having five separate annotation rounds and taking a pixel-wise majority to obtain final segmentations. The people involved in data collection were the first author and roughly 50 crowdworkers. The crowdworkers were paid 0.2 per segmentation, amounting to about $\$12$ to $\$16$ per hour. Workers were also eligible for bonuses of $\$1, \$3$ or $\$7$ for submitting $100, 250,$ or 500 segmentations respectively. Workers could collect a maximum of two bonuses: one for training images and one for validation images. We did not obtain IRB approval as our data annotation does not constitute human subject research as defined in federal regulation 45 CFR 46.102. Specifically, because we do not ask the human annotators information about themselves, they are not technically human subjects. We consulted our institution's IRB to confirm that our study is exempt from approval. Nonetheless, we closely followed principles learned from the history of human subject research, such as providing informed consent (see 15), ensuring the rights of the participants, anonymizing responses, keeping work entirely transparent, voluntary, and justly compensated. We strived to uphold the highest ethical standards in our procedures and actively mainatin a healthy environment for our workers. Annotations were collected in a two week span.

G.4 Preprocessing/cleaning/labeling

The only data cleaning performed was the consolidation of multiple rounds of annotations for validation set segmentations via pixel-wise majority vote.

G.5 Uses

The dataset has not been used yet. We have developed a suite of evaluation metrics and demonstrated the utility of the dataset to improve model performance. We hope new training methods can be developed leveraging the richer annotations of our dataset (relative to standard single label classification datasets). The dataset is not intended to replace large scale diverse datasets (e.g. ImageNet), but instead focuses on the specific subproblem of faithful learning despite strong spurious cues. Data and evaluation code will be publicly available and accessible at `mmoayeri.github.io/HardImageNet`.

G.6 Distribution

The dataset will be publicly distributed immediately upon submission. There are no limitations on use of this data.

G.7 Maintenance

The authors will maintain the dataset website and answer any question regarding usage. We encourage questions to be asked via GitHub, though the authors can be contacted directly. The primary author can be emailed at mmoayeri@umd.edu. There are no current plans to release new versions to this dataset, though if that does occur, old versions will remain archived.

H Mechanical Turk Forms

We now provide screenshots for all Amazon Mechanical Turk Forms used to facilitate data collection in our study. For full transparency, we leave copies these forms up on the free analog of Mechanicla Turk so that any interested parties can view and familiarize themselves with the annotation platform. The forms are listed at the following link: `https://workersandbox.mturk.com/requesters/ATCCTSC7WNN97/projects`. Figures 14, 15, and 16 showw the forms for the qualification exam, information/consent phase, and full data collection respectively.

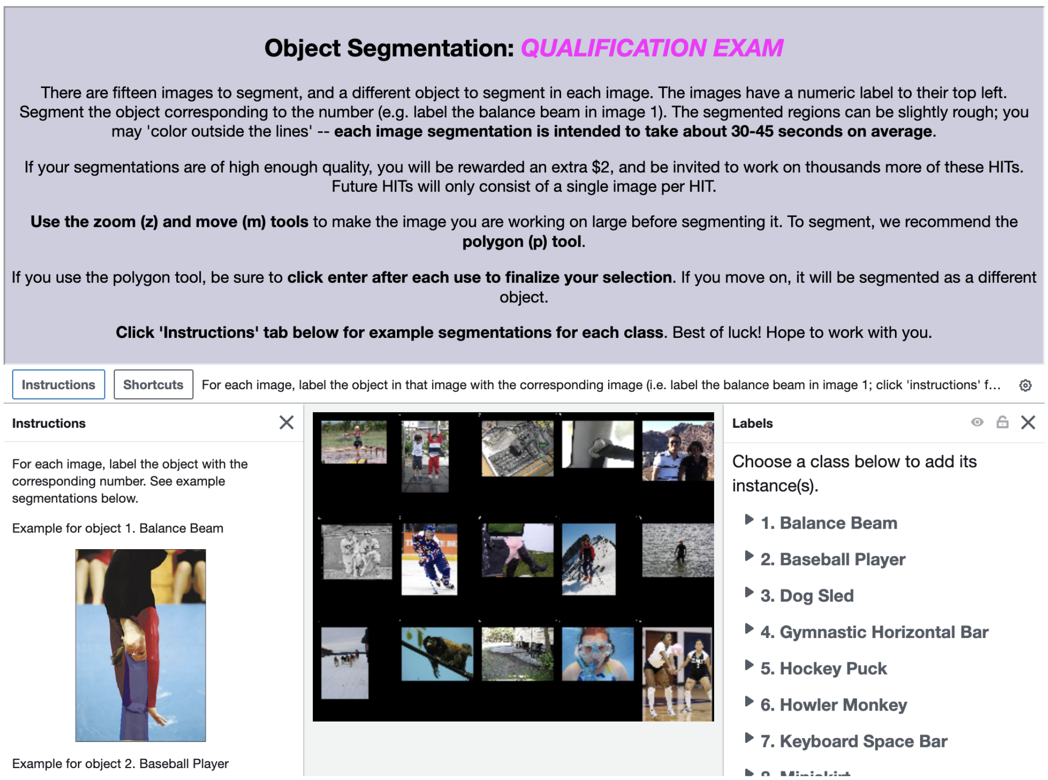


Figure 14: Qualification exam.

Consent form for Object Segmentation

Hello workers! Congratulations on qualifying. We are AI researchers from the University of Maryland. We are developing a dataset to improve robustness in machine learning models. Specifically, we are gathering segmentations for objects that models often use spurious cues to detect. For example, a model may look for car windows to detect a seatbelt, which can lead to mistakes when the model is deployed for a convertible.

We present the consent form below to confirm your voluntary participation in our study. The work will be similar to the qualifiation test, except at a much larger volume, and with a HIT corresponding to just one image (as opposed to the 15 in the qualification test).

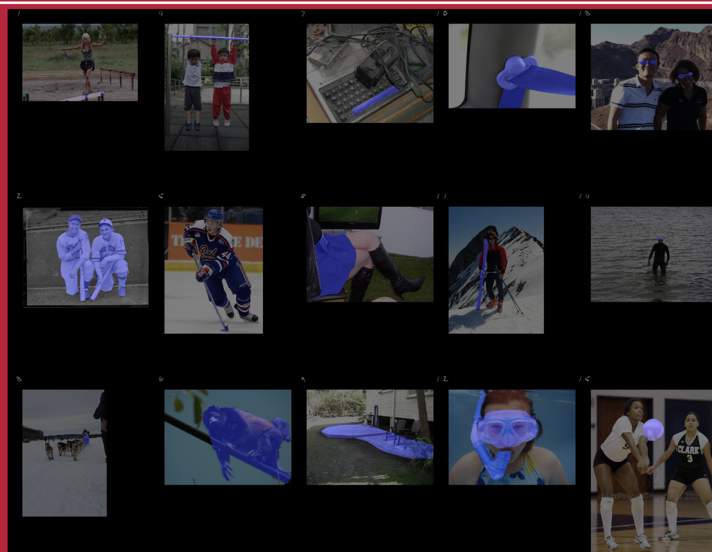
To continue receiving HITs from us (and also collect a $2 reward), mark 'I agree', hit submit, and await HITs to be released imminently (likely within a day!).

IMPORTANT NOTES

We will continue to monitor the quality of your work with randomly placed known segmentations. We will reject HITs from workers who perform poorly on these attention checks. We do not wish to reject HITs, but quality control is imperative, so we will be checking.

Only complete this consent form once***. You will not receive payment for additional submissions, and you may be removed from the study.

Please look over the correct segmentations for the qualification exam below and take note of any mistakes that may have occurred in your submission. Common errors were: including the vertical stands for the horizontal gymnastics bar, including the entire backyard when segmenting patio (**do not include grass when possible**).



Segmentations do not need to be as precise as the above picture (we do not want to overwork you); recall, you can color slightly outside the lines so long as you are segmenting the correct object and not including much else.

Instructions | Shortcuts | Do you consent to the listed terms and agree to voluntarily take part in this study?

Project Title	*Hard ImageNet: Faithful Object recognition in the presence of Strong Spurious Cues*
Purpose of the Study	*This research is being conducted by Professor Soheil Feizi at the University of Maryland, College Park. We are inviting you to participate in this research project as a worker on Mechanical Turk. The purpose of this data collection is to gather object segmentations for challenging classes towards use in machine learning models.*
Procedures	*The data collection involves coloring in regions of images where certain objects are present. The object to be annotated for each image is provided on a HIT by HIT basis.*
Confidentiality	*Your responses to HITs in this study are anonymous and will not contain information that may personally identify you.* *Anonymous responses may be shared with other scientists for research purposes or communicated via a research paper or report.*
Potential Risks and Discomforts	*The authors declare there is minimal risk for harm to participants. Images involved in this study are deemed to be benign and non-offensive.*
Compensation	*You will receive around $0.10-0.25 for completing each HIT, consisting of a single image and single attribute. We expect each task to take no more than 30-60 seconds on average. All funds will be paid through Mechanical Turk. All efforts will be made to make payment and approvals in a timely fashion.*
Qualification Test and Attention Checks	*You all have partaken in a qualification test in order to verify the quality of your work. Your work will continue to be monitored via attention checks, which are images with known annotations randomly inserted into the pool of HITs for you to complete. If the quality of your work on these attention checks is subpar, we reserve the right to reject some of the HITs you complete. We do not wish to do so, so please try your best for all the HITs, as we will be checking.*
Right to Withdraw and Questions	*Your participation in this research is completely voluntary. You may choose not to take part at all. If you decide to participate in this research, you may stop participating at any time. If you decide not to participate in this study or if you stop participating at any time, you will not be penalized or lose any benefits to which you otherwise qualify.* *If you decide to stop taking part in the study, if you have questions, concerns, or complaints, or if you need to report an injury related to the research, please contact the investigator's point person:* ***Mazda Moayeri*** ***4120 Brendan Iribe Center*** ***8125 Paint Branch Dr, College Park, MD 20740*** ***mmoayeri@umd.edu***
Statement of Consent	*Your signature indicates that you are at least 18 years of age; you have read this consent form or have had it read to you; your questions have been answered to your satisfaction and you voluntarily agree to participate in this research study. You may print a copy of this signed consent form.* *If you agree to participate, please click "I agree/consent", and then hit submit.*

Figure 15: Consent Form. Workers who passed the qualification exam then moved on to sign a consent form, where the purpose of their work was explained, common mistakes were corrected, and an extra payment was awarded. This phase is intended to create an active dialogue between data annotators and collectors. We received many inquiries and enthusiastic bits of feedback.

Figure 16: Example task for full data collection phase, only accessible to workers who passed the qualification exam and signed the consent form.