# *Perspective*: Leveraging Human Understanding for Identifying and Characterizing Image Atypicality

SHAHIN SHARIFI NOORIAN, Delft University of Technology, The Netherlands
SIHANG QIU, Hunan Institute of Advanced Technology, China
BURCU SAYIN, University of Trento, Italy
AGATHE BALAYN, Delft University of Technology, The Netherlands
UJWAL GADIRAJU, Delft University of Technology, The Netherlands
JIE YANG*, Delft University of Technology, The Netherlands
ALESSANDRO BOZZON, Delft University of Technology, The Netherlands

High-quality data plays a vital role in developing reliable image classification models. Despite that, what makes an image difficult to classify remains an unstudied topic. This paper provides a first-of-its-kind, model-agnostic characterization of image *atypicality* based on human understanding. We consider the setting of image classification "in the wild", where a large number of unlabeled images are accessible, and introduce a scalable and effective human computation approach for proactive identification and characterization of atypical images. Our approach consists of *i)* an image atypicality identification and characterization task that presents to the human worker both a local view of visually similar images and a global view of images from the class of interest and *ii)* an automatic image sampling method that selects a diverse set of atypical images based on both visual and semantic features. We demonstrate the effectiveness and cost-efficiency of our approach through controlled crowdsourcing experiments and provide a characterization of image atypicality based on human annotations of 10K images. We showcase the utility of the identified atypical images by testing state-of-the-art image classification services against such images and provide an in-depth comparative analysis of the alignment between human- and machine-perceived image atypicality. Our findings have important implications for developing and deploying reliable image classification systems.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; • **Information systems → Crowdsourcing**; • **Computing methodologies → Computer vision**.

Additional Key Words and Phrases: Image classification, image atypicality, machine learning in the wild, humans in the loop

**ACM Reference Format:**
Shahin Sharifi Noorian, Sihang Qiu, Burcu Sayin, Agathe Balayn, Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2023. ***Perspective***: Leveraging Human Understanding for Identifying and Characterizing Image Atypicality. In *28th International Conference on Intelligent User Interfaces (IUI '23), March 27–31, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3581641.3584096

## 1 INTRODUCTION

Data quality is a key factor in the success of image classification systems. Despite their impressive performance, image classification models remain largely unreliable, especially in situations slightly different from those captured in their

*Corresponding author.

training phase [1, 13]. As an implication, lack of reliability can lead to negative and sometimes damaging effects, particularly in critical domains such as transport, finance, or medicine. Among image recognition errors, a specific type known as unknown unknowns has gained particular interest [2]. Unknown unknowns refer to the images for which a model is highly confident about its predictions but is wrong. Unknown unknowns are often discovered after deployment since identifying such errors is challenging due to the overconfidence of the model. Thus, high-quality test data has become vital for understanding and proactively uncovering vulnerabilities in image classification models, as partly demonstrated by recent efforts from both academia and industry [29, 37], e.g., the Dynabench platform by Facebook[1] and the CATS4ML data challenge by Google[2].

A promise of these efforts is the creation of a feedback loop in the lifecycle of an image classification model, thereby enabling a never-ending learning scenario where model performance can continuously improve. Existing methods generally consider both a model-and-human-in-the-loop approach, where human workers identify adversarial instances that are challenging for certain specific image classification models [2, 20, 24]. Those methods, mainly contributed by Human Computation studies, are concordant with findings from Computer Vision showing that human visual systems are more robust than machines [11, 32, 41]. Little work, however, has addressed deeper questions pertaining to *i)* the characteristics of images that lead to difficulty in their classification from a human perspective and *ii)* whether such human understanding is aligned with the distribution of data that the machine perception is built upon [56].

An instrument that can allow us to gain insights into the difficulty of both human and machine classification of images is the notion of atypicality [25, 34], defined as "the strength of association between observable properties and concepts". From the human perspective, the difficulty in classification has been explained through the difficulty in recognizing components of the image [3]. When such components deviate from the norm (either due to their unusual representation, attributes that deviate from our mental models, the unfamiliar context they are presented, or partial or complete occlusion), we experience difficulty in the image classification task. From the machine learning point of view, models that fail in image classification generally learn incorrect or spurious associations (or correlations) between an image class and the components, arising from incompleteness, imbalances, or undesired biases in the data. This has mainly been found to be due to the under-representation of atypical images in the data [20, 42].

To utilize human understanding for identifying and characterizing atypical images in the context of image classification, a straightforward design for such a task would be gathering responses about the atypicality of a given image from human annotators on a subjective rating scale. However, in the context of image classification, the quality of the resulting insights would depend not only on the cognitive capability of the human annotators (and their open world view) but also on their ability to envision the perceived atypical images with respect to the distribution of data. In other words, the perceived atypical images from the human point of view may not necessarily represent a rare concept that the classification model has not encountered during the training phase.

Moreover, cost-efficiency represents another important challenge in real-world settings of image classification in the wild, where stakeholders (e.g., developers and users) have access to a large number of images without knowing the model performance on such images. In such a setting, reducing the number of images for human annotation is of critical importance to save human effort and hence cost. This paper, therefore, seeks to answer the following research questions:

**RQ:** How to support humans to effectively identify and characterize image atypicality in a cost-efficient manner?

[1]https://dynabench.org
[2]https://cats4ml.humancomputation.com

Given the research question, we developed ***Perspective***, an annotation tool that supports effective and scalable human computation for proactive identification and characterization of atypical images. Given an image for annotation, ***Perspective*** presents users with both a global view of images in the class of interest – including both random and visually diverse samples) and a local view of visually similar images in the dataset (possibly from multiple classes) to support human annotation of atypically. ***Perspective*** employs a data sampling method that accounts for both the atypicality and the redundancy of visual and semantic information in the sampled images, thereby narrowing down the most likely atypical images that can be passed along for human annotation. Through controlled crowdsourcing experiments, we demonstrate that our annotation tool can significantly improve worker performance in terms of accuracy and speed in atypicality annotation and that the sampling method is effective in filtering atypical images for annotation.

Through several iterations of annotation (including crowd workers) on 10K images, we present a coding scheme of 20 distinct characterizations of image atypicality, ranging from atypicality with respect to the semantic content (e.g., unusual objects or objects presented in an unusual context), the visibility of objects (e.g., occlusion), to the image quality (e.g., resolution and lighting) and formation (e.g., vantage point, out of focus). The coding scheme reveals the diversity of image atypicality characteristics; particularly, atypical semantic content constitutes the largest category of image atypicality, indicating the heavy skewness of image atypicality due to the unusual content.

To demonstrate the utility of image atypicality annotation, we test the performance of several vision APIs from the industry against our identified atypical images. Results show that atypical images present a strong challenge to state-of-the-art image classification services. To gain a deeper understanding of the alignment between human and machine perception of atypicality, we further fine-tune several image classification models locally, and manually compare the model rationales (using interpretable machine learning techniques) with human annotations. Our analysis shows that model rationales match human-annotated image atypicality to a large extent. This highlights the potential of ***Perspective*** for not only collecting atypical images to expose model errors but also for identifying reasons which have important implications for developing and deploying reliable image classification models. For example, the identified atypical images can be used to augment the training data in order to improve model performance; the reasons for atypicality can also be used to defer atypical images where models are more likely to fail for human takeover in a hybrid human-AI setting [38, 39].

In summary, we make the following key contributions:

- We introduce a scalable human-in-the-loop framework that orchestrates automatic and human computation components for efficient and effective identification and characterization of image atypicality.
- We identify 10K atypical images and provide a set of structured characterizations (code schemes) of atypicality across four atypicality categories: semantic content, object visibility, image quality, and formation.
- We demonstrate the utility of the collected atypical images by testing against state-of-the-art computer vision models and exposing their weaknesses through atypicality characteristics.
- We provide a set of insights into the need for support for data exploration and navigation in human annotation and the alignment between human and machine perception of image atypicality.

## 2 RELATED LITERATURE

We discuss related works pertaining to data quality issues and their implications, and others presenting methods that tackle concomitant problems from both algorithmic and human computation angles.

The term "data quality" in the context of machine learning usually refers to the coverage or representativeness of data distribution in terms of relevant attributes, e.g., demographics [47] and location [40], or to the correctness of the label [51]. In image classification, it has been found that state-of-the-art models fail when the objects are in strange positions [1] or even exhibit slight changes in position [36] not captured in the training phase. Such a problem remains even with big training data. For example, studies have shown that image classification models trained on the ImageNet dataset exhibit misclassifications consistent with racial stereotypes [47], biases towards textures [10], and limited generalizability to under-representative geo-locations [40]. Those problems are mainly attributed to the inequality of representation in the images within concepts, hence atypicality [52]. Technically, the coverage or distributional representativeness issue in the training data can lead to incomplete models that are prone to generate high-confidence errors, referred to as unknown unknowns [2, 9, 42]. Due to the high confidence, such errors are hard to detect and consequently, implying an ever big challenge in high-stakes domains with safety, trust, or ethical requirements.

The problem of data quality has been addressed from different perspectives. A large body of work has focused on reducing undesired bias through data preprocessing or posing additional regularization in model training or inference [14, 18]. Work can also be found on calibrating prediction confidence such that model confidence can become a reliable signal of error risks [27, 48]. Those ideas, while helping to alleviate the issue, are suboptimal by ignoring underrepresented instances or by trading off accuracy for fairness or confidence. We are, instead, more interested in methods that augment the data with adversarial instances.

A closely related line of research in machine learning is adversarial training, referring to the class of methods that automatically generate adversarial instances [6, 49]. The observation mainly drives the idea that *human-imperceptible* differences in the processed data can lead to prediction failures. Adversarial training methods are, therefore, often designed to generate instances similar to existing training instances (with imperceptible differences) while coming with different ground truth labels. As an implication, those methods cannot "naturally" generate images with significant deviation from the training data (e.g., recognizable by humans, often due to the different objects or contexts), rendering them strongly limited in the types of adversarial instances can be generated.

Human-in-the-loop methods have been developed mainly to address model errors. Unlike machines that fully rely on knowledge explicitly encoded in predefined training data, humans excel at leveraging broad, tacit, and contextual knowledge in decision-making and justification. Human computation has, therefore, emerged as a new, promising approach to detecting model errors. A seminal work by Attenberg *et al.* (2011) proposed to ask humans to gather publicly accessible instances that are potentially difficult for the model to handle. Lakkaraju *et al.* (2017) introduce a data partitioning technique that first organizes the data into multiple partitions based on feature similarity and then uses an explore-exploit strategy to search for difficult instances across these partitions. An important finding in human computation studies reveals that model errors often come with internal consistency, making them particularly suitable to be described by human language building on top of concepts and properties [24].

The potential of human computation is also verified by studies in Computer Vision, where findings have shown that human visual systems are more robust than machines, especially to distributional shift [11, 32, 41], making humans a promising computational means to identifying challenging images for machines. Existing human-in-the-loop methods, however, are specific to address errors of individual models and are hard to generalize to different tasks. Most importantly, we lack a general understanding of the characteristics of an image which leads to the difficulty in classifying it from the human perspective. We set out to fill this gap through our work in this paper.

In terms of interface design, several studies show the effectiveness of visual analytics tools in discovering model errors. For instance, DriftVis by Yang et al. [53] addresses concept drift in data streams, combining a drift detection

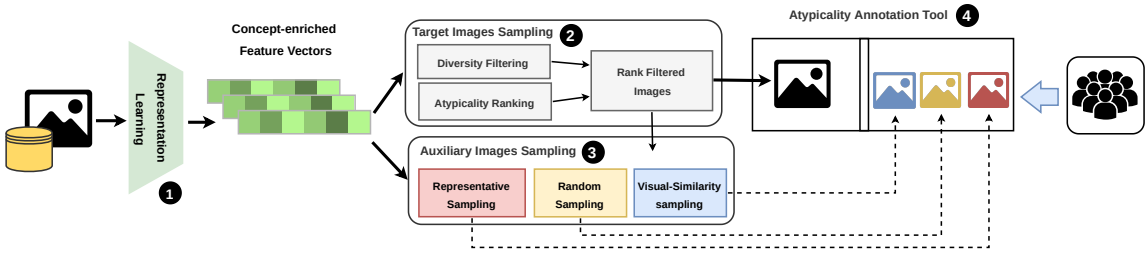



Fig. 1. Workflow of ***Perspective***. Images from a given dataset are fed to the 1) Representation Learning module to obtain image representations, which are then fed to both the 2) Target Images Sampling module to sample images for annotation and the 3) Auxiliary Images Sampling module to sample three types of auxiliary images, i.e., representative images of a class, random images of a class, and visually similar images, to assist the annotation. The selected target image, together with the auxiliary images are sent to the 4) Atypicality Annotation Interface for human annotation. Note that within Target Images Sampling, images are first filtered by Diversity Filtering, and then ranked by Atyplicaty Ranking. The top-ranking images are sent to the Auxiliary Images Sampling module for sampling images visually similar to the target image.

method and a streaming scatterplot visualization. ConceptExplorer by Wang et al. [50] detects and analyzes concept drift in multi-sourced time-series data, with visual detection based on prediction models, drift level index, and consistency judgment. Yeshchenko et al. [54] presents a system for processing drift detection and visualization in business process event logs. Chen et al. [7] introduces a visual approach for identifying and explaining out-of-distribution samples that cause degradation in predictive models. It uses an improved ensemble detection method and a grid-based visualization with a novel kNN-based layout algorithm for better context analysis. Our work complements this work by introducing a human annotation tool for identifying and characterizing image atypicality.

## 3 *PERSPECTIVE*: AN ANNOTATION TOOL FOR IMAGE ATYPICALITY

This section describes ***Perspective***, our proposed tool for annotation. We first present an overview of the tool and then describe in more detail its components.

*3.0.1 Overview.* ***Perspective*** takes as input a set of images, samples a subset of images, and feeds them to an annotation interface for human workers to annotate, concerning atypicality rating and rationale. Figure 1 presents the overall workflow of the tool. It contains four components:

(1) *Image Representation Learning*, to obtain a low-dimensional vector representation of every image in the input dataset for the following sampling components;
(2) *Target Images Sampling*, to sample a diverse set of potentially atypical images for atypicality annotation;
(3) *Auxiliary Images Sampling*, to sample visually similar images as well as images representative of the class of interest for a given target image and class;
(4) *Atypicality Annotation*, to engage human workers for rating the atypicality and providing rationales for the sampled target images, by referring to the auxiliary images.

For ease of understanding, we introduce the components in backward order.[3]

[3]Implementation details of all methods in the four components are provided in the companion page: https://sites.google.com/view/iui23perspective.

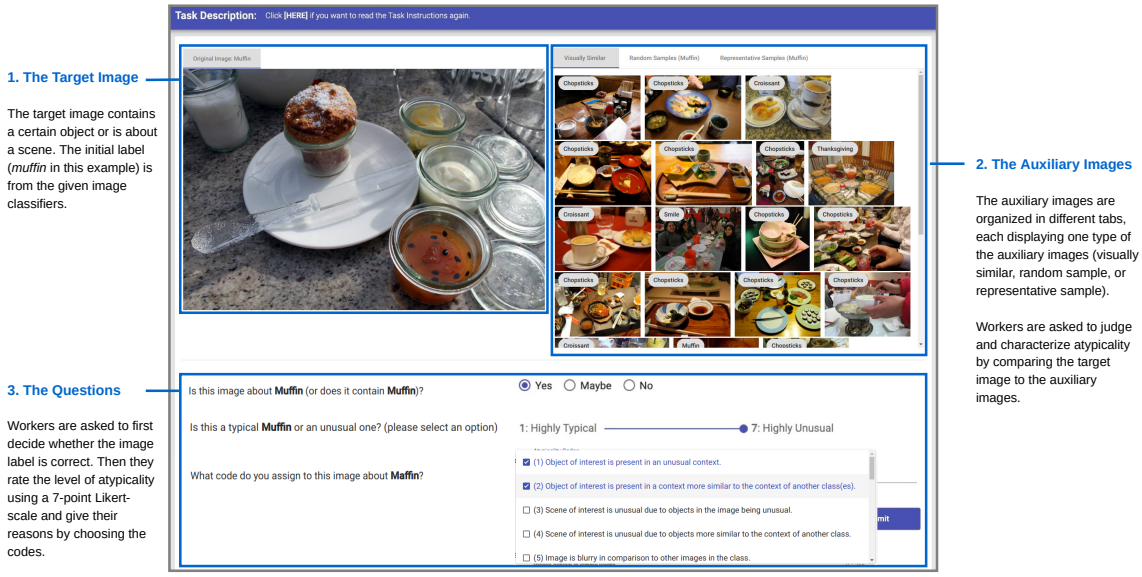



Fig. 2. A screenshot of the worker interface while using the ***Perspective*** annotation tool.

### 3.1 Image Atypicality Annotation

At the front end of our approach is the atypicality annotation task, which is used to both develop the codes of atypicality by trained annotators and to annotate images at scale by crowd workers.

*3.1.1 Task Definition & Design.* We consider two types of target image classes, namely object and scene images. Object images are those that contain a given object, e.g., "bird", "muffin"; scene images, on the other hand, are those that contain multiple objects that together describe a theme, often being an activity or event, e.g., "graduation", "thanksgiving". Atypicality generally means that to the human perception, the object of interest shows an unusual appearance, in an unusual context, or the scene of interest contains unusual objects. In image classification, we further emphasize atypicality with a relative meaning, that is, we consider an object image to be atypical if the object of interest is present in a context *more similar* to the context of any other classes, or a scene of interest is unusual due to the contained objects being *more similar* to the context of any other classes.

In atypicality rating, we consider two types of errors that annotators can potentially make, namely wrong recognition of typical images to be atypical, i.e., type I error, and the other way around, i.e., type II error. To reduce type I error, human workers need to have insight into a set of typical images representative of the entire class. To reduce type II errors, it is useful to show to the worker which classes visually similar images belong to. We note that both types of auxiliary images are selected from a given dataset that, while coming with its own limit in terms of coverage, is often available at a large size (e.g., publicly available training set or test data in the wild). Methods for sampling those images will be introduced in the next subsections and validated in our experiments.

*3.1.2 Task Interface.* Figure 2 shows the task interface. It contains three parts: 1) the target image (left), 2) the auxiliary images (right), and 3) the questions that workers answer (bottom).

The auxiliary images are organized in different tabs, each displaying one type of the auxiliary image. In addition to the visually similar images and representative images, we show a random set of images from a given class to help workers gain an idea of the general distribution of visual information in a class.

The task starts by confirming if the target image contains a certain object or is about a scene (initial labels can be obtained from any given image classifiers). Workers are asked to judge and characterize image atypicality for the given image and the associated label by analyzing the target image and comparing that to the different types of auxiliary images. They are asked to rate the level of atypicality using a 7-point Likert-scale (from Highly Typical to Highly Atypical) and when the judgment is atypical (rating bigger than the threshold 4), workers are asked to enter their rationales by selecting from a drop-down list our developed codes of image atypicality (described in Section 4).

### 3.2 Sampling Target Images

We now describe our method for sampling the target images for annotation, in order to reach a high cost-efficiency for annotation. To design the sampling method, we consider two requirements of the sampled images: 1) atypicality, i.e., the set should contain as many as possible the indeed atypical ones; and 2) diversity, i.e., the images show a variety of the atypicality characteristics. To this end, we introduce a two-stage method that first filters a subset of images with high visual diversity, and then ranks them according to an atypicality measure we derived from visual features.

*3.2.1 Diversity Filtering.* To sample a subset of diverse images, we use the recently proposed adversarial filtering method AFLite [21]. The goal of AFLite is to remove "spurious artifacts in data beyond what humans can intuitively recognize, but those which are exploited by powerful models." For that purpose, the method is designed to reduce the bias in the training data by selecting only a subset of data samples that are the most diverse possible (to avoid spurious artifacts). Consequently, filtered samples by AFLite contain a rather equal distribution of both highly typical samples (if it did not contain any, the model would not be able to learn the typical representations of the target classes of the model) –that we need to exclude through annotation–, and rarer samples –the ones that are indeed atypical.

AFLite works in an iterative process consisting of model training and evaluation. At each iteration, the available dataset is randomly partitioned into two subsets for training and test sets, respectively. The partition is performed $m$ times, and a linear classifier is trained and evaluated independently on each partition. Note that the linear classifier uses the image representation vector as features, which we introduce in the next subsection "Representation Learning" –this allows the sampling to consider visually meaningful features as compared to the low-level, pixel-based features. The evaluations on the $m$ test sets are aggregated into a *predictability score* per sample in the dataset, representing the ratio of the number of times the sample received a correct prediction over the total number of predictions. In the case of no ground truth labels available, we approximate the predictability score with the agreement among predictions from the linear classifiers. The top $k$ samples with the highest predictability scores are then removed from the dataset, and then we proceed to the next iteration. The stopping criteria is defined over the number of samples remaining in the dataset, and over the number of samples that have a predictability score higher than a pre-defined threshold.

*3.2.2 Atypicality Ranking.* Atypical images are a type of outliers, and can be detected through a relevant distribution of image representations: images that are deviated from the mean/medium of a relevant distribution are considered atypical. A general approach for outlier detection in non-parametric distributions is item ranking. In our scenario, we consider leveraging the distribution of images by model-based image representations [28], i.e., the activation of the neurons in a given layer of a neural network model. Using the model-based image representation is favored over the original image representation as it accounts for the varying contribution of different visual features in classification.

Specifically, for a given image $i$ our goal is to find the rank of the image in a subset of images $\mathcal{V}$ randomly sampled from the large dataset (such that the subset keeps the same distribution as the large image set in the wild). To do so, for each image represented by the feature vector (introduced in the next subsection), we run it through an independent multi-layer perceptron model for image classification, record the activation values of neurons in the last layer before the classification layer, and use that as a new representation of the image. Images in $\mathcal{V}$ are then ordered based on the activation values – multiple orderings corresponding to multiple neurons are aggregated into one order. We then obtain a similar representation of image $i$ and find its ranking position in the ordered list of $\mathcal{V}$.

The ranking effectiveness is, to a large extent, dependent on the neural network model. We can start with a given deployed model when available as the initial model. To best leverage human annotations, we progressively train the model following the active learning process [8]. Active learning is a way of training a machine learning model using an optimal subset of the training data, by selecting the most informative instances from the given dataset in multiple iterations. In each iteration, the model is retrained with the newly selected instances combined with the existing ones. Informativeness has various forms that are modeled in different sampling strategies, e.g., uncertainty sampling measures the informativeness of an instance by the uncertainty of model prediction [22]. In our scenario, we replace the informativeness criteria with our atypicality (ranking) measure for sampling.

### 3.3 Sampling Auxiliary Images

Sampling for the auxiliary images is straightforward for random and visually similar images: when the image representations are available, visually similar images are found through a nearest neighbor search. For representative image sampling, we develop an optimization-based approach to ensure that we present the whole spectrum of the visual appearance of a given class to human annotators. We convert the problem into a data partitioning problem, where the goal is to split the images of a given class into partitions such that images of the same partition are visually similar and those from different partitions are not; representative images can then be sampled from each of the partitions. Given a budget $\mathcal{B}$ (the number of representative images that can be sampled), we solve the following objective function for sampling:

$$\begin{aligned} & min \sum_{i \in C} \sum_{j \in C} D(Z_i, Z_j) X_{ij} \\ & s.t. \sum_{j \in C} X_{ij} = 1 \\ & X_{ij} \leqslant Y_j \\ & \sum_{j \in C} Y_j = \mathcal{B} \end{aligned} \tag{1}$$

, where $D$ represents the cosine distance between the feature vectors of two images, i.e., $Z_i, Z_j$; $X_{ij}$ indicates the decision of whether image $i$ is assigned to partition $j$; $Y_j$ indicates if the image $j$ is selected as a representative sample (note that the index $j$ is overloaded to represent both the partition and the representative image of the partition). Due to the large number of possible solutions that are associated with the problem of finding an optimal set of representative samples, it is very challenging to provide a deterministic solution. We employ a meta-heuristic approach based on a genetic algorithm which has proven to be effective in finding an optimal solution for such partitioning problems [30].

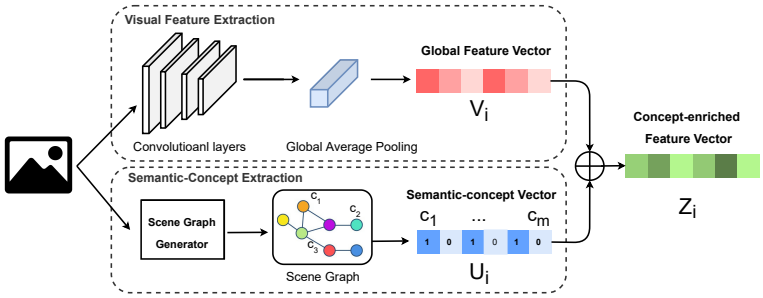



Fig. 3. The image representation learning model.

### 3.4 Representation Learning

We now present our approach for generating the image representation that supports all the previously introduced components. Considering the fact that the atypicality definition is especially relevant about the *content* of an image, i.e., objects and contexts, we aim to generate image representations that not only capture the visual features but also the semantic concepts in the image. Our representation learning approach is depicted in Figure 3 that extracts both the visual and semantic features, and concatenates them as the image representation.

*3.4.1 Visual Feature Extraction.* We use the convolutional network ResNet-152 [15] to generate a feature vector of the input image. To be specific, we feed the image to a pre-trained model until the final max-pooling layer (prior to the fully-connected layers), and extract the activations at that layer. Then we flatten the output of the max-pooling layer to obtain a feature vector $V_i : \mathbb{R}^{1\times 2048}$ for each input image $i$.

*3.4.2 Semantic Feature Extraction.* To extract semantic concepts, we use scene graph, a structured representation of objects and their relationships in an image. It consists of a set of relationships, which are represented as $\langle o_1, r, o_2 \rangle$, where $o_1$ and $o_2$ refer to two objects in the image, and $r$ represents their relation. We generate scene graphs using the state-of-the-art scene graph generation method Neural Motifs [55]. After obtaining the scene graphs for a given set of $N$ images, we extract a set of unique objects and relations. Then, for each input image $i$, we construct a fixed-length concept vector $U_i : \mathbb{R}^{1\times M}$, where $M$ corresponds to the number of unique concepts. Given $u_i^c \in U_i$ as the $c$-th concept in the concept vector, we set $u_i^c$ to 1 if the concept $c$ appears in image $i$, otherwise 0. Finally, for each input image $i$, we concatenate $U_i$ to the visual feature vector $V_i$, resulting an enriched image representation $Z_i$.

## 4 ANNOTATION, EVALUATION, AND EXPERIMENTATION SETUP

We conduct our annotation and experiments on Open Images [19], a dataset of 9.2M images with 30.1M image-level labels for 19.8K concepts. Following the CATS4ML data challenge, we use a subset consisting of 117K images of 23 classes including Canoe, Lipstick, Bird, Firefighter, Graduation, etc.[4]

We apply diversity filtering and pick the top 10K most atypical images through our atypicality ranking method. Six authors of this paper act as trusted annotators manually annotate the 10K images based on their perceived degree of

[4]see https://cats4ml.humancomputation.com for the full list.

atypicality. Each author independently annotated 1K images and the remaining 4K images were annotated by crowd workers using the ***Perspective*** interface. As a result of this process, 1925 images were identified as being atypical.[5]

### 4.1 Developing a Coding Scheme

To characterize image atypicality, we follow the open coding method rooted in grounded theory [12] and use thematic analysis to develop insights from the images [5]. As a first step, six authors independently assessed a random subset of 46 atypical images (sampled from the set of 1925 atypical images, two for each class) using the interface shown in Figure 2. In this round authors provided detailed explanations for characterizing given images as being atypical based on both their understanding of the image class in general, and the distribution of images in the dataset. Authors then iteratively identified different rationales from their explanations for characterizing image atypicality and assigned codes to represent them. Next, to refine the coding scheme and resolve any disagreement, all authors discussed each of the 46 atypical images and iteratively identified and assigned codes to characterize image atypicality.[6] For the sake of completeness and to ensure that the resulting coding scheme can be used by the community for further research, complementary codes which went beyond what was observed in this sample were added.

### 4.2 Evaluating the *Perspective* Annotation Tool

*4.2.1 Auxiliary Images Sampling.* We conduct a controlled crowdsourcing experiment to evaluate the effectiveness of the different types of auxiliary images for annotation. We design a between-subject study across four experimental conditions of the auxiliary images: 1) **Random** samples from the dataset; 2) **Representative** samples; 3) **Visually Similar** samples; and 4) **Combined**, which combines all the three above types of auxiliary images.

We randomly selected equal splits of typical and atypical images annotated by the authors, resulting in 300 images in total and at least 10 sample per each class label. We ensure that 180 common images are used for all four conditions.

We recruited 50 workers on Prolific crowdsourcing platform for each condition.[7] Only workers whose approval rate were greater than 90% were considered as qualified. During the task execution, each worker is asked to annotate six images, three atypical images and three typical images randomly selected. To avoid learning bias, each worker is allowed to perform only a singe task throughout the entire experiment. All the tasks across four conditions were published and completed within the same four-hour period on Prolific, to reduce the bias of worker availability. Each worker was paid 0.90 USD (0.65 GBP) for participating in our study. According to Prolific, the actual average hourly reward of our experiment that workers received was 11.75 USD (8.59 GBP).

We measure worker performance in terms of both annotation accuracy and speed. Specifically, accuracy is measured using the metrics precision and recall with respect to both typical and atypical images (indicated by author annotations). We measure the speed of worker annotation by the average time spent on identifying each atypical and typical images.

*4.2.2 Target Images Sampling.* We evaluate the effectiveness of our target image sampling techniques, i.e., diversity filtering and atypicality ranking. To do so, we compare the precision of the sampling by diversity filtering alone, atypicality ranking alone, and combined. Precision is measured by the fraction of truly atypical images in the sampled ones. To evaluate diversity filtering, we pick 1K images out of 12K obtained filtered images and measure the precision. We repeat the experiment for ten times and report the average precision. For a fair comparison, we rank the whole

[5]Our annotated dataset will be released in the companion page.
[6]We do not report inter-rater reliability, as the disagreement between the researchers was resolved through detailed discussions and critical reflections through multiple rounds of iterative coding [26].
[7]Note this group of workers are recruited only for evaluating our task design; workers are annotating the 4K images are recruited separately.

data set using atypicality ranking techniques and select top 1K images from the ranked list. Finally, we combine both techniques by ranking the 12K images obtained by diversity filtering, using the ranking score obtained from atypicality ranking. Then, we pick the top 1K instances to calculate the precision.

### 4.3 Human vs. Machine Perception

We apply our identified atypical images to evaluate the performance of state-of-the-art image classifiers on those images. We first test three industrial APIs: Google Vision API[8], Amazon Rekognition API[9], Microsoft Azure Vision API[10]. To gain a deeper understanding of the alignment between human and machine perceived atypicality, we locally fine-tune three models pre-trained on ImageNet, namely InceptionV3 [48], a VGG19 [45], and a DenseNet121 [17], onto a subset of the images in the Open Images dataset, corresponding to 13 classes containing the largest number of images.[11] To be able to compare the rationales of model predictions to human characterization of atypicality, we extract saliency maps from the three models using SmoothGrad [46], and manually interpret the visual elements the models highlight.

## 5 RESULTS

### 5.1 Characterizing Atypical Images

The resulting coding scheme from the six authors is presented in Table 1. We group the codes into four categories, namely, *Semantic Content*, *Image Medium Quality*, *Object Visibility*, and *Formation*. *Semantic Content* contains codes that describe the atypicality of an object in an unusual context (for object images) or a context with unusual objects presented in (for scene images). This category is different from *Image Medium Quality* that describes the atypicality in terms of image resolution, lighting, and color scheme, and from *Formation* that describes atypicality on how the image was formed (e.g., photographed) in terms of the type of medium, vantage point, and focus point. *Semantic Content* is also different from *Object Visibility* in that the former describes the atypicality of the object or scene itself, e.g., an unusual type of Pizza, whereas the latter concerns the appearance of typical object or scene, e.g., a normal Pizza partly occluded in the image. As a remark, we note that many of the codes represent the human perspective of an image while considering the perceived distribution of other images in the class, as well as other classes in the dataset.

*5.1.1 Atypicality Distribution.* To understand the distribution of atypical images with varying characteristics in our dataset (i.e., corresponding to different codes), we considered a uniformly random sample of 220 atypical images and coded them using the coding scheme. Figure 4 presents the distribution of codes that were observed as a result. *Semantic Content* is the largest atypicality category: 60% of the images were assigned `Code#1`, suggesting that the most frequent characterization of an atypical image in our sample corresponded to the object of interest being present in an unusual context; 29% images were assigned `Code#2`, i.e., atypical images where the object of interest is presented in an atypical context, in a relative sense (i.e., the context being more similar to that of other classes). *Object Visibility* is another category of atypicality many images are assigned, especially `Code#17` that describes image atypicality from object appearance in terms of shape or other attributes. Interestingly, `Code#18` was assigned to 36% of the atypical images, indicating a different medium of representing the object of interest in comparison to other images of the class. As mentioned earlier, certain complementary codes were added to the coding scheme for the sake of completeness. Either

[8]https://cloud.google.com/vision
[9]https://aws.amazon.com/rekognition/
[10]https://azure.microsoft.com/services/cognitive-services/computer-vision/
[11]Fine-tuning details are provided in the companion page.

| Type | Code | Description |
|---|---|---|
| Semantic Content | **1** | Object of interest is present in an unusual context in comparison to other images in the class. |
| | **2** | Object of interest is present in a context more similar to the context of one or more other classes. |
| | **3** | Scene of interest is unusual due to objects in the image being unusual. |
| | **4** | Scene of interest is unusual due to objects more similar to the context of another class. |
| Image Medium Quality | **5** | Image is blurry in comparison to other images in the class. |
| | **6** | Image is blurry, making it more similar to images of another class. |
| | **7** | Lighting in (a portion of) the image is too dark in comparison to other images in the class. |
| | **8** | Lighting in (a portion of) the image is too dark, making it more similar to images of another class. |
| | **9** | Lighting in (a portion of) the image is too bright in comparison to other images in the class. |
| | **10** | Lighting in (a portion of) the image is too bright, making it more similar to images of another class. |
| | **11** | Color scheme of the image is inconsistent with other images in the class. |
| | **12** | Color scheme of the image is more similar to that of images in other class(es). |
| Object Visibility | **13** | Aspect ratio of the object of interest is smaller than other images in the class. |
| | **14** | Aspect ratio of the object of interest is larger than other images in the class. |
| | **15** | The majority of the object(s) of interest in comparison to other images in the class is(are) occluded. |
| | **16** | Dominant object in the image belongs to another class(es). |
| | **17** | The shape or other attributes of the object of interest look unusual with respect to other images in the class. |
| Formation | **18** | The type of medium for representing the object of interest is inconsistent with other images in the class. |
| | **19** | The vantage point of the image is inconsistent with other images in the class. |
| | **20** | The object of interest is out of focus in comparison to other images in the class. |

Table 1. Coding scheme to characterize atypical images (in a given dataset). Multiple codes can be assigned to a single image.

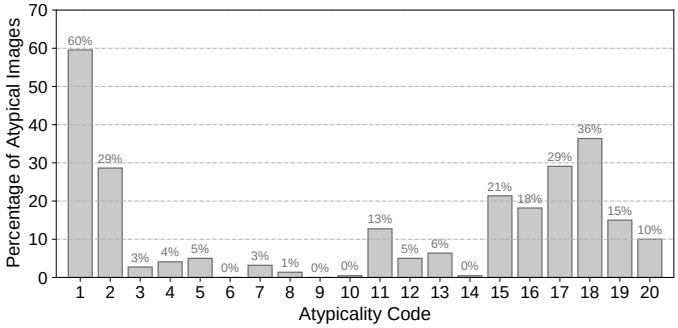



Fig. 4. Distribution of (a random sample of 220) atypical images as characterized using the coding scheme for image atypicality.

a small fraction of atypical images or none were found to correspond to such complementary characterizations (e.g. `Code#6`, `Code#8`, `Code#9`, `Code#10`).

(a) Codes: #1,#2 (b) Codes: #9,#10,#17,#19 (c) Codes: #1,#15,#16,#18,#20

Fig. 5. Example characterization of atypical images (a) **Muffin**, (b) **Pizza**, and (c) **Bird** and the corresponding codes assigned to them.

*5.1.2 Examples of Atypical Images.* Figure 5 presents examples atypical images and the corresponding codes assigned to them. 5(a) shows an image with the class label **Muffin**, characterized as being atypical due to `Codes#1` the unusual context of the muffin, i.e., presented in a glass, as well as `Codes#2` since the surrounding context is most similar to the class of **Chopsticks** where there are multiple dips nearby the main plate. Figure 5(b) shows an image with the class label **Pizza**, characterized as being atypical due to `Codes#9&10` the bright lighting, and importantly, the presence of the pizza in an `Codes#17` unusual shape due to the close-up angle, which is also related to `Codes#19` the inconsistent vantage point. Figure 5(c) shows a particularly interesting example of an atypical image corresponding to the class label of **Bird**, that is characterized by a range of codes. We can see that a **Bird** in the image is occluded by two children on the photograph held in a person's hands, thereby characterizing the atypicality of this image on several different fronts.

### 5.2 Effectiveness of *Perspective*

*5.2.1 Auxiliary Images Sampling.* Figure 6 shows estimation plots of worker performance (precision, recall, and annotating speed for atypical and typical images) [16]. In this figure, jitter plots show all the measures, and how they distribute, across the four experimental conditions. The estimation plots also show the effect size by displaying the resampling distributions of the mean difference. We found that resulting data pertaining to all the measures do not follow normal distributions (Shapiro-Wilk tests).

*Precision & Recall.* In terms of precision, the Combined condition outperforms the other three conditions on both atypical images and typical images (Figures 6 (a) and (d)). We can observe comparatively large effect sizes of the differences between the Combined condition and the Visually Similar condition. In terms of recall, the Combined condition also achieves higher performance on atypical images than the other conditions, with relatively large effect sizes. For typical images, the mean recall of Combined, Random, and Representative conditions are almost equal, while that of the Visually Similar condition is relatively lower. Note that we found no significant difference in worker precision or recall ($p > 0.05$, Kruskal–Wallis tests), which is likely due to the low number of images (six) annotated by each worker.

*Annotation Speed.* We observe that workers in Representative and Visually Similar conditions annotated atypical images faster as shown in Figure 6 (c). In annotation of typical images, workers across all the conditions exhibited comparable annotation speeds.

***Summary.*** The Combined condition is most effective for accurate annotation (precision and recall), especially for identifying atypical images. The result signifies the advantage of displaying different types of auxiliary images for annotation accuracy. This, however, comes with the trade-off of longer annotation times on average. The Representative

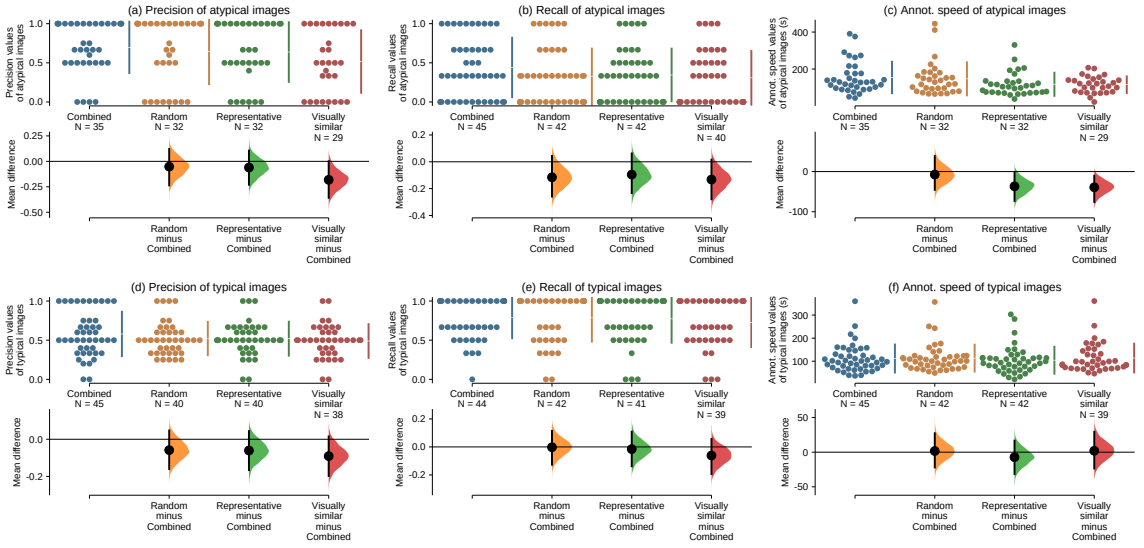



Fig. 6. Estimation plots of worker precision, recall, and annotating speed on atypical images and typical images respectively, where in the jitter plots each point represents the performance value (precision, recall, or speed) of a worker.

| Method | Diversity Filtering | Atypicality Ranking | Combined |
|---|---|---|---|
| Precision | 25.06 | 21.98 | 29.1 |

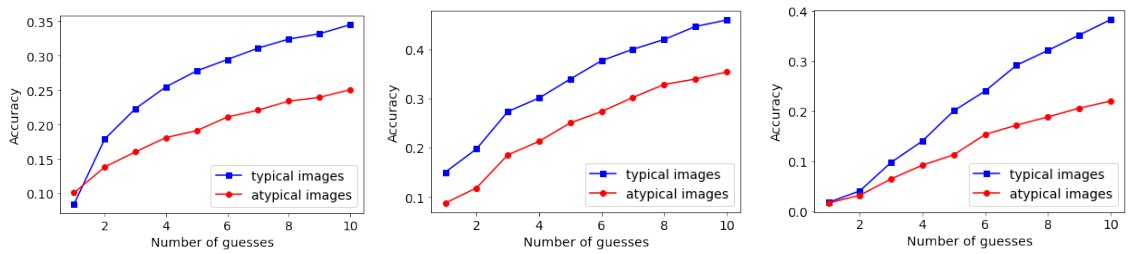
Table 2. Precision (%) of different target images sampling methods in identifying atypical images.

condition results in relatively high-quality annotations while enabling fast annotation, especially on typical images, as compared to other conditions. Visually similar images, when presented alone, do not allow workers to deliver high quality annotations; this is possibly due to the lack of a global view of the image class. By analyzing worker activity logs in the Combined condition, we noticed that all the workers who had switched the tabs (32 out of 50) clicked on visually similar images in the annotation. This suggests the perceived utility of visually similar images in informing atypicality identification in the Combined condition.

*5.2.2 Target Images Sampling.* Table 2 reports the precision of our sampling methods in identifying target atypical images. When diversity filtering and atypicality ranking are used together, we observe a significant improvement in precision.

### 5.3 Human vs. Machine Perception

*5.3.1 Industrial APIs.* Figure 7 shows the performance of the three image classification APIs on our identified typical and atypical images. These APIs classify an image with multiple labels along with their confidence scores; we, therefore, evaluate the accuracy with respect to the number of guesses allowed – classification is considered correct if one of the guesses is correct. Note that the comparison between different APIs is *not fair* due to the different sizes and vocabularies of image classes they cover. We resample the typical images according to the distribution of atypical images, such that the results on typical and atypical images are comparable. We observe from the figures that the APIs consistently show




(a) Google Vision API (b) Amazon Rekognition API (c) Microsoft Azure Vision API

Fig. 7. Performance of industrial vision APIs on the typical and atypical images.

| Atypicality | DenseNet121 | VGG19 | InceptionV3 |
|---|---|---|---|
| Typical | 63.90 | 55.69 | 57.65 |
| Atypical | 42.77 | 22.96 | 39.04 |

Table 3. Percentages (%) of correct predictions of the fine-tuned models on typical and atypical images.

higher accuracy on typical images than on atypical ones. In particular, when the number of guesses is five, the average accuracy on atypical images is 18%, as compared to 27% on typical ones.

*5.3.2 Local Models.* Locally fine-tuned models also consistently show a higher rate of correct predictions on the typical images than on the atypical ones (see Table 3). Due to the high bar of our atypicality judgment, specific samples annotated as typical with incorrect predictions might actually be atypical for more lenient characterizations of atypicality, which can further reinforce the above observations. Those results indicate that, statistically, challenges in model predictions are generally aligned with human judgments of atypicality. To gain a deeper understanding of the alignment of atypicality perceived by humans and machines, we look into the saliency maps as the rationales of model classifications and compare those to image atypicality characterized by humans.

*Typical Images Correctly Classified.* The models do not always use a correct rationale for correctly predicting the labels of typical images. Typicality does not mean that the models can easily learn the correct reasoning. Potential spurious biases across the typical images of a training dataset can lead the model to pick up on simpler, incorrect reasons. For instance, for the class **Canoe** (see Table 4 (1)), the DenseNet121 model has learned to look at the presence of water when canoes are in the water, but at the presence of a canoe itself when no water is present in the image.

*Atypical Images Incorrectly Classified.* The rationale of the models is more frequently aligned with human judgments for atypical images that the model predicts incorrectly. It is especially the case when the atypicality code relates to another class. For instance, an image of an Athlete surfing on the water is predicted as **Canoe** by the model due to the presence of water (see Table 4 (4)), which follows `Code#2` indicating that the background of this image is more similar to the ones of **Canoe** images in the dataset.

*Typical Images Incorrectly Classified.* Only a few typical images receive a wrong prediction from the models. Such misalignment between the model and human reasoning is the most complex to interpret the images and saliency maps. The most obvious cases are when the image contains two rather dominant visual cues hinting at two different classes: one referring to the expected but wrongly predicted class and one potentially referring to another class. For instance,

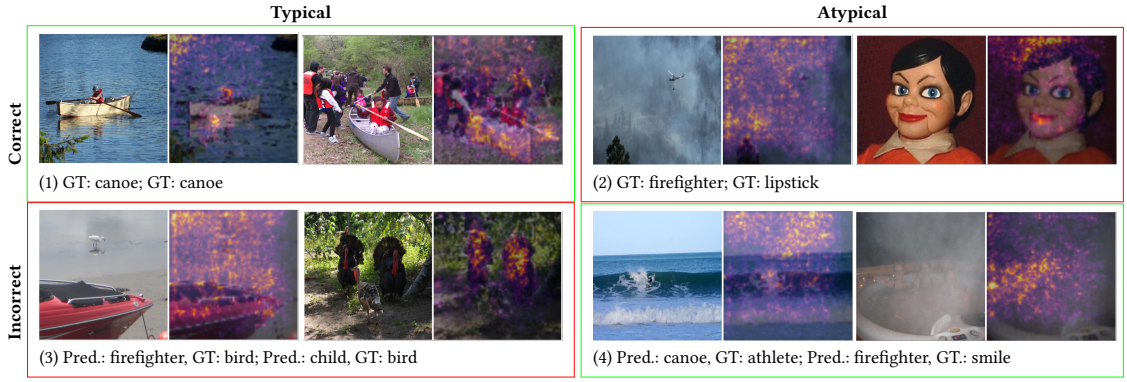



Table 4. Example images that received correct or incorrect predictions from the DenseNet121 model and judged as typical or atypical by humans (in green: alignment between human and machine reasoning, in red otherwise). The images are associated with their saliency maps on the right and the predicted (Pred.) and ground truth (GT) labels underneath.

an image of a **Bird** on the beach with a red boat is associated by the model to **Firefighter** probably due to the red color (see Table 4 (3)).

*Atypical Images Correctly Classified.* Part of the images which received correct predictions while marked atypical merit their atypicality judgment to be reviewed once the human judges have further understood the model rationale by analyzing multiple images and saliency maps. As an example, a **Firefighter** image showing a small helicopter in a background of smoke (see Table 4 (2)) is marked atypical as firefighter images instead usually contain a firetruck or individual firefighters. Yet, the model learned to use the smoke to predict the label **Firefighter** (e.g., see the image in Table 4 (4)). Another image presents a plastic doll with red lips (see Table 4 (2)), that was coded as atypical due to the medium of representation (doll with simple facial features) that is unusual for **Lipstick**, yet the model still correctly focuses on the lipstick to make its predictions. These cases show that it is not always sufficient to use the atypicality codes to estimate whether a model prediction will be correct. Still, it also requires an understanding of how important a given atypicality characterization is concerning other potentially more typical characteristics of the image that are less obvious from an open-world human perspective (e.g., the smoke for the firefighter instead of the individual firefighter). This hints at new opportunities in the coding process: a sequential coding procedure could potentially first allow the judges to build an understanding of model reasoning by visualizing saliency maps, ground truth, and predictions and then ask them to characterize image atypicality based on such understanding of the reasoning.

## 6 DISCUSSION

### 6.1 Importance of Context Expansion

Results from our controlled crowdsourcing experiments show that workers, when presented with representative images of a given class and with visually similar images to the target image, perform significantly better in terms of both annotation quality and speed. These results verify our initial assumption that human perspectives that rely on annotator experiences can be limited in envisioning image atypicality; for that, being able to see image distributions in the dataset (images in the same class but also the other classes). This is confirmed further by our results from the coding exercise, where many of the codes represent not only the human perspective but also such perspective conditioned on the distributions of the images. These results therefore, pose new research questions on what impacts human perspective,

and especially how new experiences gained from human interactions with new environments (objects, scenes) shape the development of human perspectives. Such questions are related to the literature on cognitive science and creativity especially. In this literature, it has been shown that collecting and navigating through information is an important phase in the creative process, which expands the current context of the topic (and fosters associative and inspirational learning) [4, 43]. While partly answering the research questions, more research is needed to cross-check the exact influence of human experience on the perception of atypicality.

### 6.2 Need for Collaboration and Interaction Tools

From the tooling perspective, the results imply that providing adequate support for human annotation is an important and perhaps indispensable part of human annotation. In our work, we have mainly explored methods and interfaces for sampling and visualizing images from certain distributions, while much is left for future studies. An important aspect to be considered in developing new tools would be to consider the cooperation among human workers. In our specific task of image atypicality identification and characterization, being able to communicate with other workers allow further expanding the current context of an individual worker as constrained by what they observe and their own memory, making it possible to connect to the new contexts other workers are experiencing. When developing support for context expansion from either extra information or communication, an essential type of atypicality that needs to be accounted for is the semantic content atypicality, namely unusual content and context. This type of atypicality makes the majority and is perhaps the most complex type given the diversity of objects and scenes. Future work in this direction can benefit from cognitive science but also more technical domains such as knowledge management, to link the annotation interface to knowledge bases in the backend that can offer in real-time new concepts related to the running context. This also calls for new research on interaction techniques, namely, how to display the increasing amount of information to workers while not significantly increasing their cognitive load.

### 6.3 Response and Data Sampling Biases

One common issue in most crowdsourced image annotation tasks is "response bias", where annotators may tend to complete tasks quickly to earn rewards, leading them to choose the simplest answer without considering answer quality. In our task, workers might have the tendency to label images as atypical which does not require characterizing atypicality. To reduce such a bias, we have employed several approaches such as using “gold standard” images to filter out unreliable annotations and soliciting multiple annotations per image. Another potential source of bias in our approach can be the data sampling bias. This has been a major consideration in the design of our approach, which takes into account image diversity in addition to atypicality. ***Perspective*** however, can potentially be further improved by integrating alternative data sampling methods, e.g., combined with out-of-distribution data detection [7].

### 6.4 Implications for Machine Learning and Interdisciplinary Research

As for the application domain, findings from our study have several implications on machine learning (computer vision specifically). An important one is the notion of atypicality as the proxy of data quality. We showed how easy it is for state-of-the-art machine learning systems to fail in dealing with atypical images. A direct implication would be the need to consider atypical images not only in model development but also in model deployment: it is nearly impossible to collect the perfect set of images covering all possible scenarios in one shot, yet this can be compensated by the incremental discovery of atypical images in model deployment, from which the data quality can be gradually improved by integrating such images. This implication aligns with the current discussions on data-centric AI, which stresses

more the importance of developing higher-quality datasets than models [33]. Our contribution in this sense is showing how human perspectives can be leveraged in the process of image atypicality identification and characterization. A further implication from this study is, therefore, the need to better bridge machine learning research and data science, with the IUI and HCI communities.

There are multiple ways of using ***Perspective*** to improve the reliability of image classification. One approach is augmenting the training data with the identified atypical images to retrain the model, in an active learning setting where model performance can be continuously improved with new images [23, 31, 35, 44]. Another approach one can also consider to use ***Perspective*** for reliable image classification is a hybrid human-AI setting, where humans can take over decision-making when model decisions are unreliable. In such a scenario, the human-identified reasons for image atypicality can be used to build a decision deferral mechanism that filters images for decision handover [38, 39].

## 7 CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a study on image atypicality identification and characterization through human annotation. We introduced ***Perspective***, an annotation tool that increases annotation accuracy and speed by presenting several distinct sets of auxiliary images, and that increases cost-efficiency by carefully designed sampling techniques. Iterative coding resulted in a coding scheme for image atypicality, comprising 20 distinct characterizations of image atypicality. Trusted and crowdsourced annotation resulted in 10K images with atypicality judgments. Experiments show that the identified atypical images present a strong challenge to state-of-the-art image classification services and models, and that atypical characteristics can well explain model rationales in instances of incorrect classification.

In the imminent future, we will improve the annotation tool to account for model behavior, explore the integration of model interpretability methods, and study further the utility of atypical images for improving system performance by, e.g., augmenting the training data.

## REFERENCES


[1] Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. 2019. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4845–4854.

[2] Josh M Attenberg, Pagagiotis G Ipeirotis, and Foster Provost. 2011. Beat the machine: Challenging workers to find the unknown unknowns. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

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

[4] Thomas Binder, Giorgio De Michelis, Pelle Ehn, Giulio Jacucci, and Per Linde. 2011. *Design things*. MIT press.

[5] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis*. American Psychological Association.

[6] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. 2018. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069* (2018).

[7] Changjian Chen, Jun Yuan, Yafeng Lu, Yang Liu, Hang Su, Songtao Yuan, and Shixia Liu. 2021. OoDAnalyzer: Interactive Analysis of Out-of-Distribution Samples. *IEEE Transactions on Visualization and Computer Graphics* 27, 7 (2021), 3335–3349. https://doi.org/10.1109/tvcg.2020.2973258

[8] David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. 1996. Active Learning with Statistical Models. *Journal of Artificial Intelligence Research* 4, 1 (March 1996), 129–145.

[9] Thomas G Dietterich. 2017. Steps toward robust artificial intelligence. *AI Magazine* 38, 3 (2017), 3–24.

[10] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. 2018. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231* (2018).

[11] Robert Geirhos, Carlos R Medina Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. 2018. Generalisation in humans and deep neural networks. *arXiv preprint arXiv:1808.08750* (2018).

[12] Barney G Glaser and Anselm L Strauss. 2017. *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

[13] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324* (2018).

[14] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413* (2016).

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[16] Joses Ho, Tayfun Tumkaya, Sameer Aryal, Hyungwon Choi, and Adam Claridge-Chang. 2019. Moving beyond P values: data analysis with estimation graphics. *Nature methods* 16, 7 (2019), 565–566.

[17] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2261–2269. https://doi.org/10.1109/CVPR.2017.243

[18] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*. IEEE, 1–6.

[19] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2018. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982* (2018).

[20] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. 2017. Identifying unknown unknowns in the open world: representations and policies for guided exploration. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. 2124–2132.

[21] Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *International Conference on Machine Learning*. PMLR, 1078–1088.

[22] David D. Lewis and William A. Gale. 1994. A Sequential Algorithm for Training Text Classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland). 3–12.

[23] Lin Li and Michael Spratling. 2023. Data Augmentation Alone Can Improve Adversarial Training. *arXiv preprint arXiv:2301.09879* (2023).

[24] Anthony Liu, Santiago Guerra, Isaac Fung, Gabriel Matute, Ece Kamar, and Walter Lasecki. 2020. Towards Hybrid Human-AI Workflows for Unknown Unknown Detection. In *Proceedings of The Web Conference 2020*. 2432–2442.

[25] Eric Margolis and Stephen Laurence. 2007. The ontology of concepts-abstract objects or mental representations? *Noûs* 41, 4 (2007), 561–593.

[26] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[27] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help?. In *Advances in Neural Information Processing Systems*. 4694–4703.

[28] Robert Munro. 2021. *Human-in-the-Loop Machine Learning*. Manning Publications.

[29] Andrew NG. 2021. MLOps: From Model-centric to Data-centric AI. https://www.deeplearning.ai/wp-content/uploads/2021/06/MLOps-From-Model-centric-to-Data-centric-AI.pdf. Deeplearning.ai [Online; posted: June-2021].

[30] Shahin Sharifi Noorian, Achilleas Psyllidis, and Alessandro Bozzon. 2018. A time-varying p-median model for location-allocation analysis. In *Proceedings of the 21st Conference on Geo-Information Science (AGILE 2018)*. AGILE.

[31] Sanglee Park and Jungmin So. 2020. On the effectiveness of adversarial training in defending against adversarial example attacks for image classification. *Applied Sciences* 10, 22 (2020), 8079.

[32] Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9617–9626.

[33] Neoklis Polyzotis and Matei Zaharia. 2021. What can Data-Centric AI Learn from Data and ML Engineering? *arXiv preprint arXiv:2112.06439* (2021).

[34] Bing Ran and P Robert Duimering. 2010. Conceptual combination: Models, theories and controversies. *International Journal of Cognitive Linguistics* 1, 1 (2010), 65–90.

[35] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. 2021. Data Augmentation Can Improve Robustness. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=kgVJBBThdSZ

[36] Amir Rosenfeld, Richard Zemel, and John K Tsotsos. 2018. The elephant in the room. *arXiv preprint arXiv:1808.03305* (2018).

[37] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[38] Burcu Sayin, Fabio Casati, Andrea Passerini, Jie Yang, and Xinyue Chen. 2022. Rethinking and Recomputing the Value of ML Models. *arXiv preprint arXiv:2209.15157* (2022).

[39] Burcu Sayin, Jie Yang, Andrea Passerini, and Fabio Casati. 2021. The science of rejection: a research area for human computation. *arXiv preprint arXiv:2111.06736* (2021).

[40] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536* (2017).

[41] Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. 2020. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning*. PMLR, 8634–8644.

[42] Shahin Sharifi Noorian, Sihang Qiu, Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2022. What Should You Know? A Human-In-the-Loop Approach to Unknown Unknowns Characterization in Image Recognition. In *Proceedings of the ACM Web Conference 2022*. 882–892.

[43] Ben Shneiderman. 2002. Creativity support tools. *Commun. ACM* 45, 10 (2002), 116–120.

[44] Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data* 6, 1 (2019), 1–48.

[45] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1409.1556

[46] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825* (2017).

[47] Pierre Stock and Moustapha Cisse. 2018. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 498–512.

[48] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[49] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*.

[50] Xumeng Wang, Wei Chen, Jiazhi Xia, Zexian Chen, Dongshi Xu, Xiangyang Wu, Mingliang Xu, and Tobias Schreck. 2020. ConceptExplorer: Visual analysis of concept drifts in multi-source time-series data. In *2020 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 1–11.

[51] Jie Yang, Alisa Smirnova, Dingqi Yang, Gianluca Demartini, Yuan Lu, and Philippe Cudré-Mauroux. 2019. Scalpel-cd: leveraging crowdsourcing and deep probabilistic modeling for debugging noisy training data. In *The World Wide Web Conference*. 2158–2168.

[52] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 547–558.

[53] Weikai Yang, Zhen Li, Mengchen Liu, Yafeng Lu, Kelei Cao, Ross Maciejewski, and Shixia Liu. 2020. Diagnosing concept drift with visual analytics. In *2020 IEEE conference on visual analytics science and technology (VAST)*. IEEE, 12–23.

[54] Anton Yeshchenko, Claudio Di Ciccio, Jan Mendling, and Artem Polyvyanyy. 2021. Visual drift detection for event sequence data of business processes. *IEEE Transactions on Visualization and Computer Graphics* 28, 8 (2021), 3050–3068.

[55] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural Motifs: Scene Graph Parsing with Global Context. In *Conference on Computer Vision and Pattern Recognition*.

[56] Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju, and Avishek Anand. 2019. Dissonance between human and machine understanding. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.