Rapid Image Labeling via Neuro-Symbolic Learning

Yifeng Wang*† yifeng.wang@connect.polyu.hk The Hong Kong Polytechnic University Hong Kong, China

Zhi Tu* tu85@purdue.edu Purdue University West Lafayette, USA

Yiwen Xiang† 20183749@cqu.edu.cn Chongqing University Chongqing, China

Shiyuan Zhou† shiyuan.zhou@mail.utoronto.ca University of Toronto Toronto, Canada

Xiyuan Chen† garethcxy.chen@mail.utoronto.ca University of Toronto Toronta, Canada

Bingxuan Li li3393@purdue.edu Purdue University West Lafayette, USA

Tianyi Zhang tianyi@purdue.edu Purdue University West Lafayette, USA



Figure 1: Two image labeling tasks from highly specialized domains and two from common domains with example labeling rules.

ABSTRACT

The success of Computer Vision (CV) relies heavily on manually annotated data. However, it is prohibitively expensive to annotate images in key domains such as healthcare, where data labeling requires significant domain expertise and cannot be easily delegated to crowd workers. To address this challenge, we propose a neuro-symbolic approach called Rapid, which infers image labeling rules from a small amount of labeled data provided by domain experts and automatically labels unannotated data using the rules. Specifically, Rapid combines pre-trained CV models and inductive logic learning to infer the logic-based labeling rules. Rapid achieves a labeling accuracy of 83.33% to 88.33% on four image labeling tasks with only 12 to 39 labeled samples. In particular, Rapid significantly outperforms finetuned CV models in two highly specialized tasks. These results demonstrate the effectiveness of Rapid in learning from small data and its capability to generalize among different tasks. Code and our dataset are publicly available at https://github.com/Neural-Symbolic-Image-Labeling/Rapid/

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Computer vision**; **Knowledge representation and reasoning**.

KEYWORDS

Image Labeling, Neuro-symbolic Learning, Active Learning, Inductive Logic Learning

ACM Reference Format:
Yifeng Wang, Zhi Tu, Yiwen Xiang, Shiyuan Zhou, Xiyuan Chen, Bingxuan Li, and Tianyi Zhang. 2023. Rapid Image Labeling via Neuro-Symbolic Learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3580305.3599485

*Both authors contributed equally to this research.
†This work was done when these students were research interns at Purdue University.

This work is licensed under a Creative Commons Attribution International 4.0 License.

KDD '23, August 6–10, 2023, Long Beach, CA, USA

© 2023 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0103-0/23/08.
https://doi.org/10.1145/3580305.3599485

1 INTRODUCTION

Deep learning methods have shown great power in challenging computer vision tasks, such as traffic scene detection [5, 19] and disease diagnosis [41, 52]. These methods often require a vast amount of labeled image data to achieve good performance. Labeling these data is laborious and expensive. This challenge is exacerbated in highly specialized domains such as healthcare, where data labeling requires significant domain expertise.

Many data labeling methods have been proposed to address this challenge with a small amount of labeled data (e.g., less than 100 labeled samples). One mainstream line of research is to develop models for automated data labeling [8, 24, 29, 39]. Existing approaches in this category often learn class prototypes from the training data samples and infer the class of unlabeled data by assigning the class of its nearest class prototype. To adopt these approaches in a low resource setting, the distance between data samples is often designed to depend on task-specific information such as meta-data or other task-specific insights. However, the task-specific nature of these approaches restricts their generalizability to other tasks. Besides, the need for designing specific models for a specific task requires extensive human efforts, which is against the motivation of saving human efforts for data labeling in the first place.

To address the aforementioned limitations, a new data labeling paradigm called data programming [32] has been proposed for rapid data labeling. For example, Snorkel [31] asks domain experts to create labeling functions and uses a generative model to combine those labeling functions to provide probabilistic labels. However, those labeling functions must be written in a programming language such as Python. This requirement not only incurs an overhead of manually composing labeling functions but also incurs a steep learning curve for domain experts and end users who typically do not have any programming experience.

In this paper, we propose a new neuro-symbolic approach called Rapid for image labeling in low-resource settings (e.g., less than 100 labeled images). The novelty of this framework lies in synergizing the strength of neural models (i.e., handling rich, complex image data) and the strength of inductive logic learning (i.e., learning from small datasets) to handle the image labeling challenge. Specifically, Rapid automatically infers logic rules from a small amount of labeled data, applies the inferred rules to label the unlabeled data, and solicits user feedback to refine the rules iteratively.

Unlike Snorkel, Rapid infers labeling rules automatically rather than requiring users to manually construct these rules. Rapid leverages the First-Order Inductive Learner (FOIL) algorithm to infer logic rules based on the low-level visual attributes extracted by pre-trained models. This way, our approach disentangles the perception and the learning process, making it more transparent and explainable to human labelers. Furthermore, to maximize the efficiency of data usage, we develop a multi-criteria active learning method to iteratively elicit human feedback to refine the labeling rules.

We conduct extensive experiments on datasets from two highly specialized domains and two common domains. Our method achieves significantly higher labeling accuracy on the two highly specialized domains (85.52% on disease diagnosis and 86.11% on bird species labeling, respectively) compared to the baseline models. We demonstrate that by actively refining labeling rules with rapid, incremental human feedback, Rapid can effectively embed expert knowledge and achieve high image labeling accuracy with a limited amount of training data.

Overall, this work makes the following contributions:

- We proposed a new neuro-symbolic learning framework that synergizes pre-trained computer vision models with inductive logic learning for rapid image labeling with a limited amount of training data;
- We designed a new conflict-based informativeness metric for data selection in active learning;
- We conducted comprehensive evaluations on four labeling tasks from different domains with user simulation and multiple baselines.

2 RELATED WORK

2.1 Image Labeling

Reducing the human effort in image labeling has become increasingly important in recent years with the advent of deep learning, which requires a massive amount of labeled data to train a model. There has been a continuous effort to develop automated solutions for image labeling. The basic idea is to train a model with some labeled data and automatically assign labels to new data samples without further human involvement. Among these existing methods, fully-automated methods have attracted significant attention and achieved promising performance. Some methods exploit the similarity between unlabeled and label images [8]. Some methods resolve the problem of data scarcity by creating representations of images using auxiliary information such as corresponding captions [29], meta-data [24], or pseudo-labels generated by other models [39]. Despite the promising performance of these methods, they often lack generalizability to domains where data acquisition is challenging, as the models usually require a substantial amount of data to learn the knowledge. Thus, some semi-automated labeling approaches use human efforts in the training or inference process to provide the information needed for training [37] or a coarse initial label [13]. Compared with existing work, our approach uses inductive logic learning to infer logic rules from a small amount of human-annotated data to label images.

2.2 Active Learning

Active learning is widely adopted to get humans involved in the labeling process iteratively while minimizing the amount of data to be labeled by humans. Active learning methods select the data samples that can benefit the model most in each iteration and request humans to label them to push the usage of human efforts to the minimum. To determine which data sample to be labeled by humans, some approaches use probability models and prioritize the data samples with high prediction inconsistency [12, 14, 26, 53], some depend on the vectorized representation from deep learning models [16, 34, 54], and some calculate the low-rank matrix representation for both labeled and unlabeled data to calculate the informativeness [46].

Though the active learning strategy can optimize data selection to reduce human effort, it often requires many iterations to achieve a reasonable labeling accuracy. Thus, it remains too expensive for labeling tasks where time is precious for domain experts such

as clinicians. Snorkel [31] enables users to explicitly embed their domain knowledge by creating labeling functions. However, the labeling functions are either written in programming languages or special declarative functions defined by the author of Snorkel, causing huge overhead effort in learning the labeling function grammar. Our work combines interactive learning with an inductive logic learner, generating logic rules to classify images. The generated rules are represented with simple logic and descriptive predicates, which are easy for users to read and edit.

2.3 Neuro-Symbolic Learning

There has been a growing interest in combining neural networks with symbolic methods [1, 17, 23, 25, 50]. Here, the term *neuro* refers to artificial neural networks or connectionist systems, while the term *symbolic* refers to AI approaches that perform explicit symbol manipulation, such as term rewriting, graph algorithms, and formal logic. There are different ways of combining neural network modules with symbolic learning. Following the categorization in a recent survey [33], our method belongs to a cascading neuro-symbolic paradigm that extracts latent patterns from input data using a neural system and then feeds them into a symbolic reasoner for final prediction.

Existing approaches in this category include NS-VQA [51], NS-CL [23], and FO-SL [25]. NS-VQA [51] firstly parses an image to a structural scene representation with Mask R-CNN and ResNet-34 and converts a natural language question into a query program with an LSTM model. Then it uses a symbolic executor to run the program on the scene representation to obtain the answer to the given question. NS-CL [23] adopts a similar approach as NL-VQA but learns the feature vector representation of an object from question-answer pairs, instead of extracting them directly with pre-trained models. FO-SL [25] represents images in first-order logic and uses an SAT solver to solve visual discrimination puzzles.

Unlike existing neuro-symbolic learning approaches in this category, we are the first to use inductive logic learning as the symbolic method for rule inference. In this way, we can explicitly model the logic of labeling rules. Furthermore, our approach is also the first to apply neuro-symbolic learning to image labeling.

2.4 Human Feedback in Inductive Logic Learning

To the best of our knowledge, our work is the first that integrates active learning with Inductive Logic Learning (ILL). Existing research in ILL has focused on improving the learning algorithm for better efficiency and scalability. There are only a few that investigates the interactivity of ILL in the 1990s [4, 9, 10]. Specifically, De Raedt et al. [9, 10] propose an interactive paradigm in which the inductive learner asks a yes/no question about the correctness of a learned rule. If a user answers no, the learner will backtrack and learn a new rule. Bergadano et al. [4] propose to prompt users for new counter-examples to refute an incorrect logic program, but users need to manually design counter-examples from scratch. More recently, Sivaraman et al. [35] present an interactive inductive logic programming approach to infer rule-based code patterns for code search. However, this approach only allows users to mark some search results as correct or incorrect for pattern refinement. None of these four approaches have an active learning component (i.e., a data selection algorithm) to carefully rank and select data samples for user inspection and correction.

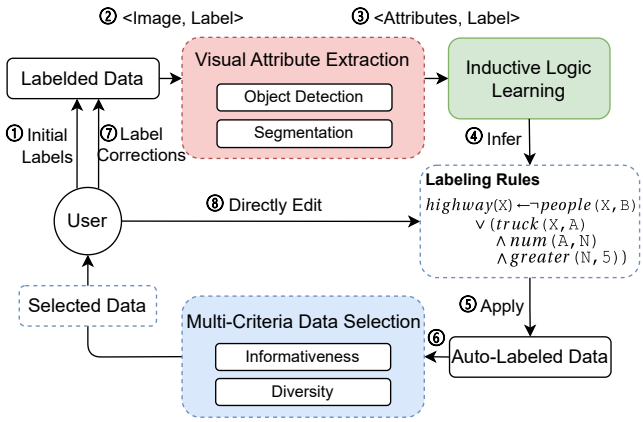


Figure 2: The overview of our image labeling approach.

3 METHOD

Figure 2 gives an overview of our approach called Rapid. Rapid consists of three parts—(1) a pre-trained visual attribute extractor to extract basic, low-level visual attributes from images, (2) an inductive learner to infer logic labeling rules from the relationship between the visual attributes and the target classes, and (3) a multi-criteria data selection module to select a small set of informative and diverse automatically labeled data for users to inspect and fix. Rapid works iteratively. It starts with a small initial set of training images created by human users. In each iteration, Rapid first extracts the visual attributes of each labeled image. Then, it takes as input the visual attributes and the corresponding class label of the images to infer a set of labeling rules for each class. The inferred rules are then applied to automatically generate labels for unannotated images. In case of contradicting labels, Rapid selects the corresponding label of the rule with the highest Clause Satisfaction Ratio (detailed in Section 3.3.2, Equation 7). Next, Rapid adopts a multi-criteria data selection strategy to compute the informativeness score for each unannotated data and select a diverse set of data samples for users to inspect. Users can fix the incorrect labels, which will be used to refine the labeling rules by Rapid.

3.1 Visual Attribute Extraction with Pre-trained Models

A visual attribute extractor processes a given image and extracts the basic, low-level attributes of the image that are useful and relevant to the labeling task. The visual attributes can be object types in an image, the relationships between the objects, and an object's properties (e.g., size, number). This work mainly uses pre-trained perception models as the visual attribute extractors, detailed in Section 4.3. But one can also use a traditional feature extractor such as SIFT [21] to extract visual attributes. The visual attribute extractors are designed as pluggable components in our approach. For different labeling tasks, we use different pre-trained models to extract

visual attributes related to the labeling task. This design increases the flexibility of our approach to be reused for new labeling tasks.

3.2 Rule Inference via Inductive Logic Learning

Given its capability to learn from a small amount of data, we use First Order Inductive Learner (FOIL) [30] to infer labeling rules. Furthermore, the declarative nature of logic rules makes them easy to be understood and refined by human labelers based on their domain knowledge. FOIL is initially designed to learn a logic rule with pre-defined predicates to distinguish a set of positive and negative examples. In our design, each predicate represents one trait of a visual attribute. The original FOIL algorithm can only infer logic rules with variables, which lacks the expressiveness for logic rules with constant values. Therefore, we extend FOIL to support the inference of constant values. As a consequence, this increases the search space exponentially. To address this challenge, we design several inductive biases, such as a TF-IDF-based heuristic, to improve search efficiency (detailed in Section 1 of Supplementary Material [45]).

Table 1: Logic Predicates for Expressing Visual Attributes

Predicate	Description
object(X,A)	Object A exists in image X
overlap(A,B)	Object A and B overlap the image
color(A,Y)	The color of object A is Y
num(A,N)	There are N object A in the image
area(A,N)	Object A has the area of N in the image
greater(N,α)	N is greater than α
smaller(N,α)	N is smaller than α

In our approach, a labeling rule is defined in a disjunctive normal form with k clauses, as shown below.

$$L \leftarrow C_1 \vee C_2 \vee ... \vee C_m \tag{1}$$

$C_1, ...C_m$ denote clauses and L denotes the label. If at least 1 clause is satisfied, an image is labeled as class L. A clause is defined as,

$$C \leftarrow p_1 \wedge p_2 \wedge ... \wedge p_k \tag{2}$$

where $p_1, ...p_k$ are logic predicates for visual attributes. A clause is a conjunctive normal form with k predicates. Hence, a clause is satisfied if and only if all the predicates are satisfied. In this work, we design a set of primitive predicates for different kinds of visual attributes, as shown in Table 1. For example, in the traffic scenario labeling task, a target image class, "highway", can be inferred based on the types of objects on the road, e.g., "trucks". An example labeling rule for "highway" images can be:

$$\begin{aligned} highway(\mathtt{X}) \leftarrow \neg people(\mathtt{X},\mathtt{B}) \vee \\ (truck(\mathtt{X},\mathtt{A}) \wedge num(\mathtt{A},\mathtt{N}) \wedge greater(\mathtt{N},5)) \end{aligned} \tag{3}$$

where X is an input image, A and B are objects detected by a pre-trained object detection model. This rule means that if there are no pedestrians or there exist more than five trucks, the image is classified as "highway". In practice, users can redesign the predicates, e.g., by removing irrelevant predicates and adding domain-specific predicates based on the characteristics of each labeling task to improve the efficiency of inferring logic rules.

Algorithm 1 Inductive Learning for Labeling Rules

Input: positive examples (T^+) and negative examples (T^-) for a label, clauses that must be included (I) and must be excluded (E)
Output: rule R

```
1:  R ← ∅
2:  R.APPEND(I)
3:  while T+ != ∅ do
4:      clause ← ∅
5:      T_i^- ← T^-
6:      S ← INITIALIZE(T+)
7:      while T_i^- != ∅ do
8:          clause.APPEND(MAX_GAIN(S))
9:          T_i^- ← REMOVE(T^-, clause)
10:     end while
11:     if clause ∉ E then
12:         T+ ← REMOVE (T+, clause)
13:         R.APPEND(clause)
14:     end if
15: end while
16: return R
```

Algorithm 1 describes how to infer labeling rules with inductive learning. For each image label, our algorithm takes the set of positive examples T^+ and the set of negative examples T^- as input and infers a logic rule. It also allows human labelers to specify which clauses must be included (I) or excluded (E) based on their domain knowledge. It first adds must-include clauses into the rule R (Line 2). Then, it keeps searching for possible clauses until T^+ is empty (Lines 3 to 15). If the clause does not need to be excluded (Line 11), the algorithm removes the positive examples which contain all visual attributes in the clause from T^+ (Line 12), and then adds the clause to the rule (Line 13). The algorithm initializes a negative set T_i^- as T^- (Line 5) and a set containing all possible predicates from the set of positive examples S (Line 6) before the first iteration of the inner loop. To find a possible clause (Line 7 to 10), the algorithm constantly selects predicates with the maximum information gain from S and adds into the clause (Line 8) until T_i^- is empty (Line 7). The information gain of each predicate is defined below:

$$Gain(S_i) = T_i^{++} \times (log_2(\frac{T_{i+1}^+}{T_{i+1}^+ + T_{i+1}^-}) - log_2(\frac{T_i^+}{T_i^+ + T_i^-})) \tag{4}$$

where T_i^+ and T_i^- denote the set of positive examples and the set of negative examples *before* adding the new predicate S_i. T_{i+1}^+ and T_{i+1}^- denote the set of positive examples and the set of negative examples *after* adding S_i. Then in each iteration in the inner loop, T_i^- is redefined to a set that removes the negative examples which contain all visual attributes in the clause from T^-(Line 9). The loop continues until it finds a labeling rule that matches all labeled images in a given class (i.e., positive examples) while not matching labeled images in other classes (i.e., negative examples).

3.3 Labeling Rule Refinement via Active Learning

Due to the ambiguity and incompleteness of the small amount of training data, the inductive learning module may not learn the best labeling rules in one pass. We propose to use active learning

to improve the performance of inductive learning by iteratively soliciting more human labels.

In each iteration of active learning, when selecting data samples, our goal is to choose the data with the most information and variety to reduce data usage and improve performance. We propose a multi-criteria data selection strategy to achieve this goal.

Algorithm 2 Multi-criteria Data Selection

Input: Unlabeled data U, the size of the intermediate set M, the number of data samples to select N
Output: The set of selected data samples S
1: Calculate the informative score for U
2: $U_{ranked} \leftarrow$ sort U by informativeness score
3: $S_{intermediate} \leftarrow$ pick top M data instances
4: $S \leftarrow$ K-Means($S_{intermediate}$)
5: **return** S

3.3.1 Multi-Criteria Data Selection. Algorithm 2 describes the multi-criteria data selection process. First, it calculates the informativeness score of each unlabeled data instance and ranks them based on the scores. Then, it selects the M most informative samples to form an intermediate set. These samples are then clustered based on similarity, and our algorithm selects the final set of N samples that are both informative and diverse. The informativeness metric and the clustering algorithm are detailed in the following subsections.

3.3.2 Informativeness. Existing informativeness metrics in the literature of active learning are typically calculated based on prediction probabilities, e.g., entropy-based uncertainty [18]. Thus, they are only applicable to statistical models that output prediction probabilities. Since our approach uses logical rules, it does not produce a probability for an inferred image label.

To bridge the gap, we propose a novel informativeness metric based on the extent of labeling conflicts among labeling rules. This design is based on the insight that multiple labeling rules may generate conflicting labels for the same image, which can be leveraged to measure the uncertainty of image labeling. The more conflicts there are in our labeling rules about the image labeling result of an image, the more information the image can bring to our model.

The informativeness is largely measured by the number of inconsistent labels. To break the tie among images with the same number of inconsistent labels, we further consider the extent of conflict in the unsatisfied labels (U in Equation 5).

$$Score(i) = \begin{cases} 0, & \#\text{label} = 1 \\ \lambda(\#\text{label}) + U(i), & \text{Otherwise} \end{cases} \tag{5}$$

$$U(i) = 1 - \frac{\sum_{r \in \text{Unsatisfied}} CSR(i, r)}{|\text{Unsatisfied}|} \tag{6}$$

In the equation above, i denotes an image; R denotes a labeling rule composed of a disjunction of clauses ($L \leftarrow C_1 \vee C_2 \vee ... \vee C_m$). Here, each rule exclusively defines a unique label. Unsatisfied denotes the set of rules the image i does not satisfy. λ is a hyper-parameter, which is empirically set to 0.6 in our experiments. The *CSR*—Clause Satisfactory Ratio—measures the degree of satisfaction of a single rule, as defined in Equation (7). U measures the average *CSR* for the rules that the image does not satisfy.

$$CSR(i, r) = \max_{i=1..m} \frac{\sum_{j=1}^{k} Sat(i, p_{ij})}{k} \tag{7}$$

In Equation (7), p_{ij} denotes the j-th predicate of the i-th clause in rule R. $Sat(i, p)$ represents whether a predicate is satisfied:

$$Sat(i, p) = \begin{cases} 1, & p \text{ is satisfied by } i \\ 0, & \text{Otherwise} \end{cases} \tag{8}$$

3.3.3 Diversity. The robustness of image labeling models largely depends on the variety of labeled data. Therefore, our data selection algorithm also accounts for the diversity of images to maximize the variety of a group of data samples when selecting the data to label. We propose to cluster all the unlabeled images into k clusters and choose one sample from each cluster to form the group of data to label. We use k-Means to cluster the data samples and choose the final centroids to generate a diverse group of samples. In k-Means, each image is represented with a vector, and the similarity between every two images is measured by the cosine similarity of the two vectors. The feature vector for an image has d dimensions, which represent the total d types of objects we consider for this task (or dataset). The value in each dimension is the number of the corresponding objects detected in the image.

4 EXPERIMENTS

We design experiments to answer the following research questions.

- RQ1 How effective is Rapid on image labeling tasks from different domains?
- RQ2 How effective are the two kinds of human feedback solicited by Rapid?
- RQ3 How effective is inductive logic learning compared to statistical and neural network models?
- RQ4 How sensitive is our active learning algorithm to different data selection strategies?

4.1 Experiment Design & Setup

To answer RQ1, we measure the image labeling accuracy and efficiency of Rapid on four different image labeling tasks—two from highly specialized domains and two from common domains. Section 4.3 describes the four tasks and datasets. To represent the condition of learning from limited training data, for each task, we bootstrap Rapid with only 3 randomly sampled data instances with labels to learn the initial set of rules. In each following iteration, the active learning module selects 3 images and corrects their labels if wrong to refine the learned labeling rules. The choice of 3 is to simulate the rapid, incremental feedback from users. This process continues until the 20th iteration. Thus, for each task, Rapid is trained with in total 60 images. We compare Rapid with 6 baselines. Section 4.4 describes these baselines and their training procedures.

To answer RQ2, we measure the degradation of image labeling accuracy and efficiency when abating the human feedback mechanisms in Rapid. Rapid supports two kinds of human feedback—(1) directly editing the labeling rules generated by Rapid and (2) fixing incorrect labels inferred by Rapid and supplementing new labels. Thus, we create three variants of Rapid—Rapid without rule editing

(RAPID$_{no\text{-}edit}$), RAPID without any labeling correction (RAPID$_{no\text{-}al}$), and RAPID without any kinds of feedback (RAPID$_{no\text{-}feedback}$).

RQ3 aims to measure the effectiveness of adopting inductive logic learning for image labeling. To answer RQ3, we create four variants of RAPID by replacing the inductive logic learning module in RAPID with three statistical models—SVM, random forest, and XGBoost [7]—and a neural network. The design of these variants is inspired by existing frameworks such as Snorkel [31] and Concept Bottle Network [20], which use statistical models to make the final prediction based on symbolic representations extracted from raw input data. For the neural network variant, we adopt the design of the fully connected layers in ResNet-18 [15].

To answer RQ4, we compare RAPID with three alternative data selection strategies—random selection, selection with only the informativeness criterion, and selection with only the diversity criterion. We use image labeling accuracy, as well as the hit rate of misclassified data, as the evaluation metrics. Specifically, the hit rate is defined as the percentage of selected data samples that RAPID mislabels in the current iteration and thus is worth fixing. A higher hit rate indicates better effectiveness of data selection.

4.2 User Simulation

Since RAPID is designed as a human-in-the-loop approach, RAPID needs to keep soliciting feedback from human experts to refine the labeling rules. It is expensive to recruit human participants to provide feedback, especially in the two highly specialized labeling tasks that require domain experts such as ophthalmologists and ornithologists. Therefore, we develop an automated script to simulate human feedback based on the ground truth data. To simulate label corrections, our script compares the ground truth label of each image with the labels inferred by RAPID and automatically fixes the incorrect labels. To simulate human edits to labeling rules, the authors first manually constructed a set of high-quality labeling rules based on their own knowledge and the information shared on professional websites. In each iteration of the training process, our script compares the labeling rule inferred by RAPID with the corresponding manually curated rule. Our script then replaces the first inconsistent clause with the clause in the manually curated rule. In all experiments, we restrict the simulation script to only edit one clause per iteration to simulate the incremental editing process of human labelers.

4.3 Imagle Labeling Tasks and Datasets

RAPID is designed for image labeling tasks in highly specialized domains. Therefore, we first select two datasets—Glaucoma Diagnosis and Bird Species Labeling—from highly specialized domains. To test the generalizability of RAPID, we construct the two datasets on general domains, including traffic scene labeling and occupation labeling, by searching on Google and Flickr.

Glaucoma Diagnosis. Given a color fundus image, this task requires labeling the eye in the image to be either normal or diseased. We combine the color fundus images from three datasets, Drishti-GS [36], RIM-ONE_r3 [3], and REFUGE [28]. We have 116 images of glaucomatous eyes and 189 images of normal eyes with both glaucoma diagnosis and structure segmentation. We use a pre-trained model called BEAL [43] as the visual attribute extractor to obtain the segmentation of eye fundus structures in the images. The visual attributes designed for this task are the diameter, area, and cup-to-disk ratio calculated based on the segmentation results.

Bird Species Labeling. Given a bird image, this task requires labeling the bird species. We use the Caltech-UCSD Birds-200-2011 (CUB 200-2011) dataset [40], containing 11,788 bird images annotated with 200 bird species and 312 attributes that describe each body part of a bird, e.g., wing color, tail shape, etc. Following the experiment settings in Koh et al. [20], we use 112 out of the 312 attributes, and randomly choose three bird species to label in our experiments. We use the pre-trained concept models from Koh et al. [20] to extract the visual attributes.

Occupation Labeling. Given an image of a person, this task requires labeling the occupation of the person. For this task, we build a dataset containing 300 images of three occupations—chef, farmer, and teacher. Each occupation has 100 labeled images. We use a pre-trained object detection model [2] to detect objects in the images and use the type (glasses, long hair, kitchen, etc.), color, and overlapping relationship between them as visual attributes.

Traffic Scene Labeling. Given a road image, this task requires labeling the traffic scene of the image. For this task, we build a dataset containing 420 images of three traffic scenes—mountain road, highway, and downtown. Each traffic scene has 140 labeled images. We use a pre-trained object detection model called DETR [6] to detect the objects in the images and use the position, color, and type (e.g., pedestrian, truck, car, etc.) of the objects and the overlapping relationship between them as visual attributes.

4.4 Comparison Baselines

For RQ1, we compare RAPID with four image classification neural network baselines—ResNet-18 [15], ResNet-34, ResNeXt-32 [47] and Inception-V3 [38])—and an active learning baseline called CEAL [42]. We further compare RAPID with GARDNet [22] on the Glaucoma diagnosis task since GARDNet is specially designed for this task. To represent the condition of learning from limited training data, in each task, we randomly sample 30 training data per class label to finetune the baseline models. That is a total of 60 training samples for the Glaucoma Diagnosis task since it only has two class labels and 90 for the other three tasks since they have three class labels. For the first five baselines, we first pre-train them on ImageNet [11] and then fine-tune them on the four datasets, with training sets in the same size as training RAPID. For GARDNet, we obtained the trained model from its original paper and then fine-tuned it on our Glaucoma diagnosis dataset.

For RQ2, we build three variants of RAPID—RAPID without editing rules by users (RAPID$_{no\text{-}edit}$), RAPID without labeling correction or new labels (RAPID$_{no\text{-}al}$), and RAPID$_{no\text{-}feedback}$, RAPID without feedback when selecting training samples. Similar to the training setting of RAPID, both RAPID$_{no\text{-}edit}$ and RAPID$_{no\text{-}al}$ are initially trained with 3 randomly selected images. In the following interactions, RAPID$_{no\text{-}edit}$ only fixes incorrect labels in the 3 images selected by the multi-criteria active learning algorithm per iteration but does not apply any direct edits to the inferred rules. By contrast, RAPID$_{no\text{-}al}$ only makes direct edits to one clause of the inferred rules per iteration but does not fix any incorrect labels. RAPID$_{no\text{-}feedback}$ does not use any active feedback from users. It is

trained with randomly sampled images in various numbers (e.g., 3, 6, 9, etc.) ahead of time without soliciting further human feedback.

For RQ3, we create four variants of Rapid by replacing the inductive logic learner with other machine learning methods, including SVM, random forest, gradient boosting, and neural network. For these variants, we use a feature vector generated with the extracted visual attributes to represent each image. The feature vector is constructed in the same way as calculating the diversity criterion in Section 3.3.3. We repeat each training three times and compute the average performance of each baseline.

5 RESULTS

5.1 RQ1. Effectiveness on Different Labeling Tasks

Table 2 shows the image labeling accuracy of Rapid in the four labeling tasks from different domains in comparison to the fine-tuned models. Rapid outperforms all baselines on the two highly specialized domains (Glaucoma Diagnosis and Bird Species Labeling) by 11.75% to 12.03%. Specifically, Rapid achieves a high accuracy of 85.52% and 86.11% in these two tasks, respectively. The result shows Rapid can effectively infer accurate labeling rules in highly specialized domains with a small amount of training data. For more details about the labeling rules, such as examples and statistics of the optimal rules, and how the labeling rules change over iterations, please refer to Section 2 of Supplementary Material [45].

Table 2: Comparison of accuracy (%) between Rapid and image labeling baseline models in the four tasks.

	Highly Specialized Domains		Common Domains	
	Glaucoma	Bird Species	Occupation	Traffic
ResNet-18	62.84	74.08	87.78	63.09
ResNet-34	72.13	65.74	**97.78**	74.21
ResNext-32	55.74	57.41	93.33	54.37
Inception-V3	50.82	58.33	94.44	54.76
CEAL	73.77	66.67	89.99	**92.86**
GARDNet	54.65	-	-	-
Rapid	**85.52**	**86.11**	88.33	83.33

For the two common domains, Rapid achieves comparable or worse accuracy. Specifically, Rapid achieves an accuracy of 83.33% in the traffic scene labeling task, while the best baseline model achieves 92.86% accuracy. For the occupation labeling task, the accuracy of Rapid is 88.33% while the best baseline model achieves 97.78% accuracy. This result is not surprising since the baseline models are pre-trained on ImageNet, which includes a considerable number of images similar to the ones in these two common tasks. Thus, the baseline models have already learned from many similar cases during the pre-training process.

Figure 3 shows the image labeling accuracy of Rapid during the training process. At the 4th iteration, with only 12 training samples, Rapid has already achieved a reasonable accuracy—85% in Glaucoma diagnosis, 70% in occupation labeling, 65% in traffic scene labeling, and 64% in bird species labeling. Within the 13th iteration, Rapid has achieved the peak accuracy on all four tasks. Besides, during the training process, the performance of Rapid is stably improving. The result shows Rapid can effectively learn and refine labeling rules within a small number of iterations.

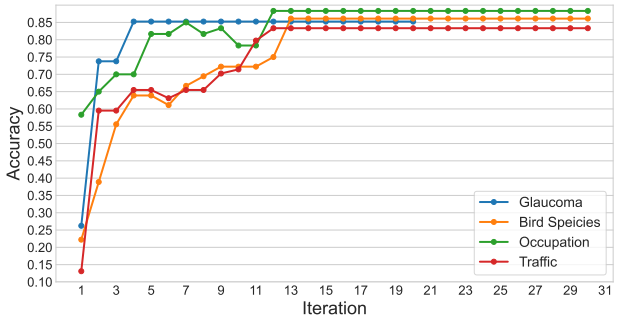


Figure 3: Image labeling accuracy of Rapid on four tasks during the training process.

5.2 RQ2. Effectiveness of Human Feedback

Table 3 shows the image labeling accuracy and the number of iterations to achieve the optimal accuracy of Rapid in comparison to its invariants after ablating each feedback mechanism. Overall, Rapid always achieves the highest accuracy with the smallest number of iterations (10.25 on average). $\text{Rapid}_{no\text{-}feedback}$ has the worst performance with the largest number of iterations (25.25 on average). Without direct rule editing, $\text{Rapid}_{no\text{-}edit}$ takes significantly 4X more iterations to achieve a comparable accuracy on the Glaucoma diagnosis task and has significantly lower accuracy on the other three tasks (11.87% accuracy decrease on average).

Table 3: Comparison of accuracy (%) and the number of iterations consumed (Iter.) between Rapid and three variants ablating two types of user feedback in the four tasks.

	Highly Specialized Domains				Common Domains			
	Glaucoma		Bird Species		Occupation		Traffic	
	Acc.	Iter.	Acc.	Iter.	Acc.	Iter.	Acc.	Iter.
$\text{Rapid}_{no\text{-}edit}$	**85.25**	21	66.67	13	81.67	27	73.81	10
$\text{Rapid}_{no\text{-}al}$	**85.25**	4	**86.11**	19	**88.33**	11	**83.33**	8
$\text{Rapid}_{no\text{-}feedback}$	83.61	14	55.56	28	80.00	29	67.86	30
Rapid	**85.25**	4	**86.11**	13	**88.33**	12	**83.33**	12

Though $\text{Rapid}_{no\text{-}al}$ can achieve the same final accuracy as Rapid, it takes 6 more iterations in the bird species labeling task. It is interesting to observe that $\text{Rapid}_{no\text{-}al}$ takes the same or even fewer iterations in three tasks. This is because the user simulation script always makes the right edit based on the ground-truth labeling rule each time. When using active learning together with rule editing, Rapid will regenerate the label after receiving new labels in each iteration. These newly generated rules may deviate from the ground-truth rules, therefore leading to more iterations.

Compared with the three variants, Rapid achieves the largest performance gain in the bird species labeling task. This performance gain can be largely attributed to the inherent learning challenge of the dataset. For example, birds of the same species are of great

variety (e.g., a Kentucky Warbler can exhibit eight distinct wing color variations.). Thus, this increases the difficulty of learning a proper labeling rule to define what a certain bird species look like from such a small training dataset. On the other hand, by editing the rules, experts can directly embed their expert knowledge into the labeling rules, leading to a huge performance improvement.

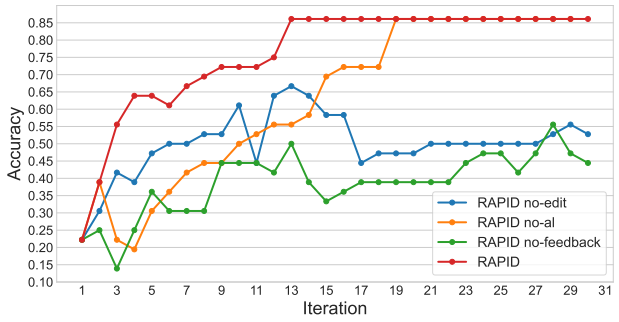


Figure 4: Comparison of accuracy in the training process between RAPID and three variants ablating two types of user feedback in the bird species labeling task.

Figure 4 shows the image labeling accuracy of RAPID and its variants over iterations in the bird species labeling task. For ease of comparison, we show the accuracy of $\text{RAPID}_{no\text{-}feedback}$ over the number of randomly sampled training samples it uses. In other words, for $\text{RAPID}_{no\text{-}feedback}$ in Figure 4, 1 in the x-axis means training with 3 samples, 2 means training with 6 samples, 3 means training with 9 samples, so on and so forth. Overall, RAPID achieves the highest image labeling accuracy (86.11%) with only 12 iterations and 36 images in total. While $\text{RAPID}_{no\text{-}al}$ achieves the same accuracy, it takes 7 more iterations and thus 21 more images to reach this accuracy. In this task, both $\text{RAPID}_{no\text{-}edit}$ and $\text{RAPID}_{no\text{-}feedback}$ takes significantly more iterations to achieve the final accuracy.

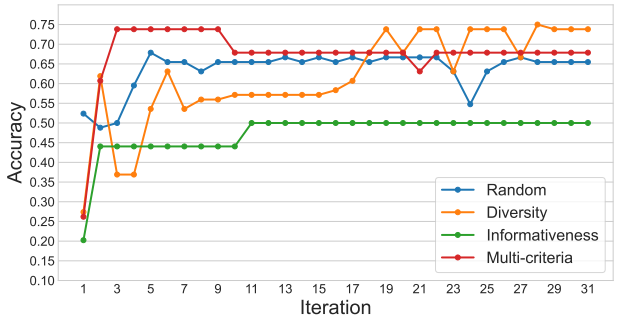
Table 4: Comparison of accuracy (%) between RAPID and four variants replacing the inductive logic learner with statistic and NN modules in the four tasks.

	Highly Specialized Domains		Common Domains	
	Glaucoma	Bird Species	Occupation	Traffic
RAPID_{SVM}	50.82	70.37	41.11	33.33
$\text{RAPID}_{XGBoost}$	79.23	58.33	78.33	74.60
$\text{RAPID}_{RandomForest}$	79.23	70.37	79.44	79.76
$\text{RAPID}_{NeuralNetwork}$	41.53	42.59	33.55	34.52
RAPID	**85.25**	**86.11**	**88.33**	**83.33**

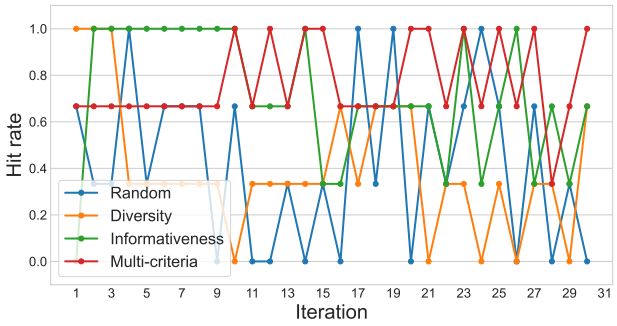
5.3 RQ3. Effectiveness of Inductive Logic Learning

Table 4 shows the comparison of accuracy between RAPID and variants replacing the inductive logic learner with statistical and neural network models. RAPID outperforms all variant models in all four tasks by 6.02%, 15.74%, 8.89% and 3.57%, respectively. The results show the great capability of the inductive logic learning model in learning image labeling tasks under a low resource setting.

Though $\text{RAPID}_{RandomForst}$ does not achieve an accuracy as high as RAPID, it is worth mentioning that $\text{RAPID}_{RandomForst}$ outperforms all the baseline models in Table 2, in the two highly specialized domain tasks. This implies that our pipeline, similar to Concept Bottleneck Models [20], which disentangles the perception process (Visual Attribute Extraction) and the learning process, has great capability in highly specialized image labeling tasks.



(a) Comparison of accuracy.



(b) Comparison of hit rate.

Figure 5: Comparison of accuracy and hit rate for different data selection strategies in the traffic scene labeling task.

5.4 RQ4. Sensitivity Analysis of Active Learning

Figure 5 shows the comparison of accuracy and hit rate between four data selection strategies in active learning in the traffic scene labeling task. Recall that the hit rate is defined as the percentage of selected data samples that RAPID mislabels in the current iteration and thus is worth fixing. A higher hit rate indicates better effectiveness of data selection. Due to the page limit, we only showcase the experiment results on this task. Figures for the other three tasks have been provided in Section 3 of the Supplementary Material [45].

Among the four data selection strategies, the multi-criteria strategy achieves the highest image labeling accuracy (73.81%) with the smallest number of iterations (3). By contrast, using the diversity criterion alone takes 19 iterations (6X more) to achieve similar accuracy. Both using the informativeness criterion alone and random selection achieve lower accuracy, 50.00% and 67.86%. Yet compared with random selection, the informative selection still helps, achieving significantly higher accuracy within fewer iterations.

In the traffic scene labeling task, random selection, single diversity criterion, single informativeness criterion, and multi-criteria

achieve 42.22%, 40.00%, 71.11%, and 76.67% average hit rate over the iterations. Combining informativeness and diversity, our multi-criteria strategy achieves the highest hit rate in this task. The single informativeness strategy maintains the highest hit rate at the beginning of the training process, which proves the effectiveness of our informativeness metric designed in Section 3.3.2. The results on accuracy and hit rate imply that the two metrics are complementary to each other. With the informativeness metric selecting training samples with high information and the diversity metric choosing representatives from the selected samples, Rapid can achieve high accuracy in a small number of iterations.

6 DISCUSSION

The experiment results demonstrate the effectiveness of Rapid in learning accurate labeling rules with a small amount of training data. Based on the ablation studies in RQ3 and RQ4, we found that both inductive logic learning and multi-criteria active learning play vital roles in the success of Rapid. Furthermore, given the inherent transparency and explainability of logic-based labeling rules, Rapid provides affordance for users to directly embed their expertise into the labeling rules via rule editing. This further improves the efficiency of rule inference, leading to significantly fewer feedback iterations compared with using active learning alone.

Our work is the first to apply inductive logic learning to neuro-symbolic learning. The experiment for RQ3 demonstrates the learning capability of our FOIL-based inductive logic learner. Specifically, our logic learner outperforms alternative statistical and neural network models. It is interesting to observe that when replacing the inductive logic learner with Random Forest, the resulting model can still achieve better performance than the fine-tuned models in Table 2 in highly specialized domains. This implies that simply disentangling the perception process from the reasoning process can still be beneficial when learning highly specialized tasks in a low-resource setting (i.e., training with a small amount of data).

Compared with Snorkel [31], our approach has two significant advantages. First, Rapid can learn an initial set of labeling rules from a small amount of labeled data (e.g., 3 training samples in our case) as a starting point for expert users to refine. In contrast, Snorkel requires users to manually write labeling functions from scratch, which can be effortful for expressing complex knowledge. Second, to use Snorkel, users need to be familiar with a programming language such as Python to write labeling functions. Thus, it comes with a steep learning curve for domain experts and end-users who do not have programming background. By contrast, the logic labeling rules in our approach are readable and intuitive. Thus, our approach comes with a more gentle learning curve compared with Snorkel.

This work is also a good demonstration of effective human-AI collaboration in challenging tasks. In our case, Rapid infers an initial set of labeling rules and continuously refines them based on human feedback in the form of direct edits or label corrections. Our experiment in RQ2 has shown that incorporating human feedback is critical to infer accurate labeling rules. Without humans in the loop, Rapid has significantly lower accuracy even when trained with the same amount of data as training with many iterations.

Despite the promising results, the final image labeling accuracy of Rapid may still not be on par with human experts. Our approach achieves an average of 85% in the four labeling tasks, while the accuracy of human labelers on ImageNet is about 95% [27]. Similar to Snorkel [31], we also want to argue that such relatively noisy data can still provide valuable supervision signals for model training, especially when used together with noise-robust learning [44, 49] and weak supervision [31, 48]. This can be extremely beneficial in highly specialized domains such as medical imaging, where human labelers are expensive and hard to acquire.

The current design of the inductive logic learner in Rapid has a rigid objective of learning a rule that matches all positive samples while rejecting all negative samples. Consequently, our approach is sensitive to user mistakes (e.g., incorrect labels and edits). A single incorrect training sample can lead to a labeling rule that makes no sense to users. In future work, we will improve the inductive logic learning method by providing the flexibility to relax the rigid rule satisfaction requirement, which is expected to increase the noisy label tolerance and robustness. Besides, when training Rapid in this work, we follow the FOIL algorithm and use an information gain-based search method. In future work, We will explore improving the search mechanism by adding heuristics using visual attributes.

The performance of Rapid highly depends on the visual attributes extracted by pre-trained computer vision models. For specialized domains, it is possible that no pre-trained computer vision model can extract meaningful visual attributes. However, we think this situation will not occur very often. There are two reasons. First, numerous large computer vision models have been developed and made available online (e.g., HuggingFace, GitHub) these days. These models can recognize a wide range of objects, shapes, colors, and other basic visual attributes that generally apply to different domains, including many highly specialized domains. Second, for rare visual attributes, the pre-trained models can be substituted with conventional handcrafted computer vision techniques, such as SIFT. Our inductive logic learning component can still act on those low-level attributes and synthesize meaningful labeling rules.

7 CONCLUSION

We present a rapid image labeling method based on neuro-symbolic learning. The proposed method uses pre-trained neural network models to extract visual attributes and use first-order inductive learning to infer labeling rules based on the visual attributes. This architecture disentangles perception from learning, enabling our method to be applied to new tasks by easily changing the pre-trained visual attribute extractor. Besides, the declarative nature of logic rules enables users to directly inspect and edit the inferred labeling rules, explicitly embedding human expertise into the rules. The experiments show that our method can achieve outstanding performance on highly specialized domains under an extremely low resource setting while generalizing to other general domains with reasonable performance.

ACKNOWLEDGEMENT

We would like to thank the anonymous reviewers for their valuable feedback. This work was in part supported by an Amazon Research Award.

REFERENCES

[1] Saeed Amizadeh, Hamid Palangi, Alex Polozov, Yichen Huang, and Kazuhito Koishida. 2020. Neuro-Symbolic Visual Reasoning: Disentangling "Visual" from "Reasoning". In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 279–290. http://proceedings.mlr.press/v119/amizadeh20a.html

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*.

[3] Francisco José Fumero Batista, Tinguaro Diaz-Aleman, Jose Sigut, Silvia Alayon, Rafael Arnay, and Denisse Angel-Pereira. 2020. Rim-one dl: A unified retinal image database for assessing glaucoma using deep learning. *Image Analysis & Stereology* 39, 3 (2020), 161–167.

[4] Francesco Bergadano, Daniele Gunetti, et al. 1993. An interactive system to learn functional logic programs. In *Proceedings of the International Joint Conference on Artificial Intelligence*. International Joint Conference on Artificial Intelligence, Inc., 1044–1049.

[5] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-End Object Detection with Transformers. https://doi.org/10.48550/ARXIV.2005.12872

[7] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. ACM, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[8] Nilaksh Das, Sanya Chaba, Renzhi Wu, Sakshi Gandhi, Duen Horng Chau, and Xu Chu. 2020. Goggles: Automatic image labeling with affinity coding. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 1717–1732.

[9] Luc De Raedt. 1992. *Interactive theory revision: An inductive logic programming approach*. Academic Press Ltd.

[10] Luc De Raedt and Maurice Bruynooghe. 1992. An overview of the interactive concept-learner and theory revisor CLINT. (1992).

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[12] Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö Arık, Larry S Davis, and Tomas Pfister. 2020. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *European Conference on Computer Vision*. Springer, 510–526.

[13] Leo Grady and Gareth Funka-Lea. 2004. Multi-label image segmentation for medical applications based on graph-theoretic electrical potentials. In *Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis*. Springer, 230–245.

[14] Jiannan Guo, Haochen Shi, Yangyang Kang, Kun Kuang, Siliang Tang, Zhuoren Jiang, Changlong Sun, Fei Wu, and Yueting Zhuang. 2021. Semi-supervised active learning for semi-supervised models: Exploit adversarial examples with graph-based virtual labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2896–2905.

[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] David Joon Ho, Narasimhan P Agaram, Peter J Schüffler, Chad M Vanderbilt, Marc-Henri Jean, Meera R Hameed, and Thomas J Fuchs. 2020. Deep interactive learning: an efficient labeling approach for deep learning-based osteosarcoma treatment response assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 540–549.

[17] Drew Hudson and Christopher D Manning. 2019. Learning by abstraction: The neural state machine. *Advances in Neural Information Processing Systems* 32 (2019).

[18] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. 2009. Multi-class active learning for image classification. In *2009 ieee conference on computer vision and pattern recognition*. IEEE, 2372–2379.

[19] Jinkyu Kim, Suhong Moon, Anna Rohrbach, Trevor Darrell, and John Canny. 2020. Advisable Learning for Self-Driving Vehicles by Internalizing Observation-to-Action Rules. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[20] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *International Conference on Machine Learning*. PMLR, 5338–5348.

[21] D.G. Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Vol. 2. 1150–1157 vol.2. https://doi.org/10.1109/ICCV.1999.790410

[22] Ahmed Al Mahrooqi, Dmitrii Medvedev, Rand Muhtaseb, and Mohammad Yaqub. 2022. GARDNet: Robust Multi-View Network for Glaucoma Classification in Color Fundus Images. arXiv:2205.12902 [eess.IV]

[23] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=rJgMlhRctm

[24] Julian McAuley and Jure Leskovec. 2012. Image labeling on a network: using social-network metadata for image classification. In *European conference on computer vision*. Springer, 828–841.

[25] Adithya Murali, Atharva Sehgal, Paul Krogmeier, and P. Madhusudan. 2022. Composing Neural Learning and Symbolic Reasoning with an Application to Visual Discrimination. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, Luc De Raedt (Ed.). ijcai.org, 3358–3365. https://doi.org/10.24963/ijcai.2022/466

[26] Kun-Peng Ning, Xun Zhao, Yu Li, and Sheng-Jun Huang. 2022. Active Learning for Open-set Annotation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 41–49. https://doi.org/10.1109/CVPR52688.2022.00014

[27] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749* (2021).

[28] José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel van Keer, Deepti R. Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, Joonho Lee, Joonseok Lee, Xiaoxiao Li, Peng Liu, Shuai Lu, Balamurali Murugesan, Valery Naranjo, Sai Samarth R. Phaye, Sharath M. Shankaranarayana, and Hrvoje Bogunovic. 2020. REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical Image Anal.* 59 (2020). https://doi.org/10.1016/j.media.2019.101570

[29] Ariadna Quattoni, Michael Collins, and Trevor Darrell. 2007. Learning visual representations using images with captions. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.

[30] J. Ross Quinlan. 1990. Learning logical definitions from relations. *Machine learning* 5, 3 (1990), 239–266.

[31] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, Vol. 11. NIH Public Access, 269.

[32] Alexander J. Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data Programming: Creating Large Training Sets, Quickly. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). 3567–3575. https://proceedings.neurips.cc/paper/2016/hash/6709e8d64a5f47269ed5cea9f625f7ab-Abstract.html

[33] Md. Kamruzzaman Sarker, Lu Zhou, Aaron Eberhart, and Pascal Hitzler. 2021. Neuro-Symbolic Artificial Intelligence: Current Trends. *CoRR* abs/2105.05330 (2021). arXiv:2105.05330 https://arxiv.org/abs/2105.05330

[34] Samrath Sinha, Sayna Ebrahimi, and Trevor Darrell. 2019. Variational Adversarial Active Learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 5971–5980. https://doi.org/10.1109/ICCV.2019.00607

[35] Aishwarya Sivaraman, Tianyi Zhang, Guy Van den Broeck, and Miryung Kim. 2019. Active inductive logic programming for code search. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*. IEEE, 292–303.

[36] Jayanthi Sivaswamy, S Krishnadas, Arunava Chakravarty, G Joshi, A Syed Tabish, et al. 2015. A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. *JSM Biomedical Imaging Data Papers* 2, 1 (2015), 1004.

[37] Markus Suchi, Timothy Patten, David Fischinger, and Markus Vincze. 2019. EasyLabel: A semi-automatic pixel-wise object annotation tool for creating robotic RGB-D datasets. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 6678–6684.

[38] 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.

[39] Joseph J Titano, Marcus Badgeley, Javin Schefflein, Margaret Pain, Andres Su, Michael Cai, Nathaniel Swinburne, John Zech, Jun Kim, Joshua Bederson, et al. 2018. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nature medicine* 24, 9 (2018), 1337–1341.

[40] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. *The Caltech-UCSD Birds-200-2011 Dataset*. Technical Report CNS-TR-2011-001. California Institute of Technology.

[41] Dong Wang, Yuan Zhang, Kexin Zhang, and Liwei Wang. 2020. FocalMix: Semi-Supervised Learning for 3D Medical Image Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[42] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. 2017. Cost-Effective Active Learning for Deep Image Classification. *CoRR* abs/1701.03551 (2017). arXiv:1701.03551 http://arxiv.org/abs/1701.03551

[43] Shujun Wang, Lequan Yu, Kang Li, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. 2019. Boundary and Entropy-driven Adversarial Learning for Fundus Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 102–110.

[44] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. 2019. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 322–330.

[45] Yifeng Wang, Zhi Tu, Yiwen Xiang, Shiyuan Zhou, Xiyuan Chen, Bingxuan Li, and Tianyi Zhang. 2023. Rapid Image Labeling via Neuro Symbolic Learning Supplementary Material. (6 2023). https://doi.org/10.6084/m9.figshare.23292182

[46] Jian Wu, Anqian Guo, Victor S Sheng, Pengpeng Zhao, Zhiming Cui, and Hua Li. 2017. Adaptive low-rank multi-label active learning for image classification. In *Proceedings of the 25th ACM international conference on Multimedia*. 1336–1344.

[47] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[48] Jia Xu, Alexander G Schwing, and Raquel Urtasun. 2015. Learning to segment under various forms of weak supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3781–3790.

[49] Cheng Xue, Qi Dou, Xueying Shi, Hao Chen, and Pheng-Ann Heng. 2019. Robust learning at noisy labeled medical images: Applied to skin lesion classification. In *2019 IEEE 16th International symposium on biomedical imaging (ISBI 2019)*. IEEE, 1280–1283.

[50] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems* 31 (2018).

[51] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/5e388103a391daabe3de1d76a6739ccd-Paper.pdf

[52] Qihang Yu, Dong Yang, Holger Roth, Yutong Bai, Yixiao Zhang, Alan L. Yuille, and Daguang Xu. 2020. C2FNAS: Coarse-to-Fine Neural Architecture Search for 3D Medical Image Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[53] Lei Zhang, Yan Tong, and Qiang Ji. 2008. Active image labeling and its application to facial action labeling. In *European Conference on Computer Vision*. Springer, 706–719.

[54] Wenqiao Zhang, Lei Zhu, James Hallinan, Shengyu Zhang, Andrew Makmur, Qingpeng Cai, and Beng Chin Ooi. 2022. Boostmis: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20666–20676.