A Survey on Deep Active Learning: Recent Advances and New Frontiers

Dongyuan Li, *Graduate Student Member, IEEE*, Zhen Wang, Yankai Chen, Renhe Jiang, *Member, IEEE*, Weiping Ding, *Senior Member, IEEE*, and Manabu Okumura

***Abstract*— Active learning seeks to achieve strong performance with fewer training samples. It does this by iteratively asking an oracle to label newly selected samples in a human-in-the-loop manner. This technique has gained increasing popularity due to its broad applicability, yet its survey papers, especially for deep active learning (DAL), remain scarce. Therefore, we conduct an advanced and comprehensive survey on DAL. We first introduce reviewed paper collection and filtering. Second, we formally define the DAL task and summarize the most influential baselines and widely used datasets. Third, we systematically provide a taxonomy of DAL methods from five perspectives, including annotation types, query strategies, deep model architectures, learning paradigms, and training processes, and objectively analyze their strengths and weaknesses. Then, we comprehensively summarize the main applications of DAL in natural language processing (NLP), computer vision (CV), data mining (DM), and so on. Finally, we discuss challenges and perspectives after a detailed analysis of current studies. This work aims to serve as a useful and quick guide for researchers in overcoming difficulties in DAL. We hope that this survey will spur further progress in this burgeoning field.**

***Index Terms*— Active learning, adaptive sampling, computer vision (CV), deep learning, natural language processing (NLP), sequential optimal design, uncertainty quantification.**

I. Introduction

THE remarkable success of deep learning relies heavily on large-scale datasets with human-annotated labels [1]. However, continually labeling large-scale datasets is an extremely time-consuming, expensive, and laborious task, which tends to become a bottleneck for deep learning with limited labeled data. To tackle this issue, deep active learning (DAL) recently exhibited great potential. As shown in Fig. 1, DAL models are first trained on an initial training dataset. Then, query strategies can be iteratively applied to select the most informative and representative samples from a large pool of unlabeled data. Finally, an oracle labels the selected samples and adds them to the training dataset for retraining or fine-tuning the DAL models. DAL aims to achieve competitive performance while reducing annotation costs within a reasonable time [2], [3], [4]. Benefiting from the strong representation capabilities of various neural networks, such as graph neural networks (GNNs) [5], convolutional neural networks (CNNs) [6], and transformers [7], as well as leveraging prior knowledge from pretrained models, such as contrastive language-image pretraining (CLIP) [8] and generative pretrained transformer (GPT) [9], DAL has made significant advances.

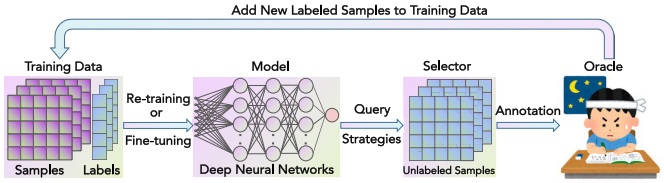


Fig. 1. General pipeline in deep active learning.

As a methodology for selecting or generating a subset of training data in data-centric AI, DAL is closely related to learning settings and practical techniques, including curriculum learning [10], transfer learning [11], data augmentation or pruning [12], [13], and dataset distillation [14]. The commonality of these methods is to train or fine-tune a model using a small number of samples, aiming to remove noise and redundancy while improving training efficiency without decreasing models' performance on downstream tasks. However, one primary difference from DAL is that these approaches have *full access* to all labels when selecting, distilling, or generating training subsets. DAL defaults to that all data should be unlabeled during the training subset selection process, making it better suited for real-world scenarios where labels are initially unavailable.

To summarize DAL methodologies, recent efforts have focused on specific tasks such as text classification [15] and image analysis [16], [17], specific domains such as natural language processing (NLP) [18] and computer vision (CV) [19], [20], or reproducing mainstream baselines [21], [22]. As for most early survey work, one common inadequacy is that they may not have enough discussion of recent advances [23], [24], [25], or lack summarization of emerging learning paradigms (contrastive learning and so on) and challenges [26], [27], especially in light of rapidly developing deep learning techniques (e.g., fine-tuning on pretrained models). To assist researchers in reviewing, summarizing, and planning for future exploration, we provide a comprehensive review encompassing the latest advancements and insights in the field. While some survey papers focus on stream-based DAL [28], this article concentrates on pool-based DAL.

Manuscript received 24 February 2024; accepted 23 April 2024. *(Dongyuan Li and Zhen Wang contributed equally to this work.) (Corresponding authors: Weiping Ding; Manabu Okumura.)*

Dongyuan Li, Zhen Wang, and Manabu Okumura are with the Institute of Innovative Research, School of Information and Communication Engineering, Tokyo Institute of Technology, Tokyo 152-8550, Japan (e-mail: lidy@lr.pi.titech.ac.jp; wzh@lr.pi.titech.ac.jp; oku@pi.titech.ac.jp).

Yankai Chen is with the School of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: ykchen@cse.cuhk.edu.hk).

Renhe Jiang is with the Center for Spatial Information Science, The University of Tokyo, Tokyo 113-8654, Japan (e-mail: jiangrh@csis.u-tokyo.ac.jp).

Weiping Ding is with the School of Information Science and Technology, Nantong University, Nantong 226019, China (e-mail: dwp9988@163.com).

Digital Object Identifier 10.1109/TNNLS.2024.3396463

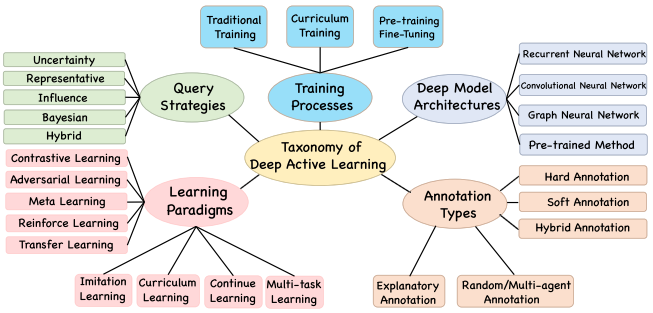


Fig. 2. Taxonomy for deep active learning methods.

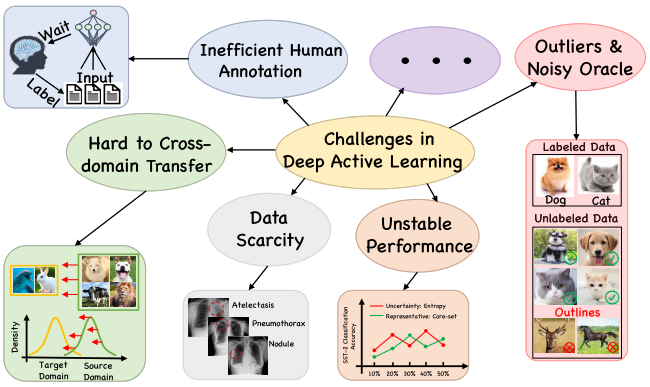


Fig. 3. Emerging challenges in deep active learning.

Specifically, we first introduce our strategy for collecting reviewed papers and explain our criteria for selecting them in Section II. Then, we give a specific formal definition for DAL in Section III-A and chronologically summarize the most influential DAL baselines and the widely used datasets in Section III-C. As shown in Fig. 2, in Section IV, we develop a high-level taxonomy to provide a broad overview of this field, categorizing previous studies from five perspectives. In Section IV-A, we classify the annotation types into hard, soft, hybrid, explanatory, and random/multiagent annotations and give a detailed introduction to each annotation type. In Section IV-B, we summarize query strategies into five distinct categories, including uncertainty-based, representative-based, influence-based, Bayesian-based, and their hybrid methods, and analyze the strengths and weaknesses of each query type. As for deep model architectures, in Section IV-C, they are mainly categorized into recurrent neural networks (RNNs), CNNs, GNNs, and pretrained methods. We discuss the benefits and drawbacks of each type of architecture. In Section IV-D, we are pleased to discover that various emerging learning paradigms, such as curriculum learning and continual learning, have shown promising results when combined with DAL. For each learning paradigm, we provide a detailed description of its definition and how to integrate it with DAL. Finally, in Section IV-E, three different training processes, including traditional training, curriculum learning-based training, and pretraining and fine-tuning (Pre + FT), will be introduced with typical examples.

In Section V, we comprehensively show some domains in which DAL methods have been successfully applied, including NLP, CV, data mining (DM), and so on. As depicted in Fig. 3, despite the remarkable progress in DAL, this rapidly developing field is still fraught with several crucial emerging challenges. In Section VI, we analyze the causes and opportunities of each challenge, which can be summarized as follows.

1) *Pipeline-Related:* Inefficient and costly human annotation, insufficient research on stopping strategies, and cold start.
2) *Task-Related:* Difficulty in cross-domain transfer, unstable performance, and lack of scalability and generalizability.
3) *Dataset-Related:* Outlier data and oracles, data scarcity and imbalance, and class distribution mismatch.

Finally, after organizing and summarizing the current DAL-related research, we have four intriguing findings that we would like to share with the readers.

1) As shown in Section IV-E, DAL has great potential as a sample selection strategy to apply few-shot or one-shot setting for large-scale pretrained models with billions of parameters [29], [30]. Furthermore, as discussed in Section III-C, many studies have shown that using only 10∼20% labeled samples for fine-tuning the pretrained language models (PLMs) with billions of parameters can yield even better performance and be 5∼10 times more efficient than training with a full labeled dataset [31], [32].
2) Intuitively, having more high-quality samples can promote model performance for some tasks. Thus, as shown in Section IV-D, many works integrate DAL with semisupervised strategies, allowing to obtain more high-quality labeled samples without increasing the need for human labor. However, as discussed in Section VI-C, semisupervised methods are highly sensitive to outliers and error labels, easily fueling a vicious cycle, i.e., models continue to label samples with wrong pseudolabels. How to effectively integrate DAL with semisupervised strategies, using human-labeled true signals to guide semisupervised annotation and avoid the mislabeled circular, remains an open and challenging issue waiting to be solved.
3) From the detailed analysis of scalability and generalizability in Section VI-B, although DAL has achieved great success in classification tasks, comparing various DAL methods to choose the optimal one for a given task remains time-intensive and unrealistic in practice. Thus, there is an urgent need for a universal framework that is friendly to various downstream tasks.
4) By summarizing DAL applications for NLP in Section V-A, we find only a few DAL studies focused on *generative tasks*. Generative tasks, such as summarization and question answering, urgently require more attention and research compared to classification tasks. This is because generating informative objects, such as annotations, is more difficult and time-consuming. Defining the most meaningful samples for generation tasks and explaining why those samples play an important role are two core problems that need to be solved. We hope that future research can promote the development of DAL for generation tasks.

Overall, the main contributions of this article are given as follows.

1) This is the latest comprehensive and systematic survey paper on DAL to help researchers review, summarize, and look forward to the future of DAL.
2) Based on the novel DAL taxonomy, we detail the explanations and discussions of the methodology, ranging from annotation types, query strategies, deep

Group Name	Keywords
Fundamental	Active Learning.
Scope	Artificial Intelligence, Machine Learning, Neural Network, Computer Vision, Bioinformatics, Deep Learning.
Context	Transfer Learning, Self-training, Data Augmentation, Semi-supervised, unsupervised, User Demands, Discrete Annotation, Human-in-the-loop, Human feed back, Crowd Annotation Framework, Label-Efficient, Continual Learning, Open-set, Online Learning, Robust, Data Acquisition, Interactive Learning, Decision Boundaries.

Publication trend analysis of 3,967 unique papers on DAL.

Year	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Number	142	175	184	198	267	370	456	542	671	685	268

Fig. 4. Keywords and publication trend on DAL.

model architectures, learning paradigms, and training processes.

3) The difficult challenges in DAL are presented from multiple perspectives. Through a detailed analysis of challenges and current studies, we discuss possible advanced solutions for them.
4) A GitHub repository[1] is available with the most up-to-date DAL techniques, including papers, code, and datasets.

The remaining part of this survey is organized as follows. Section II shows the collection of DAL papers. Section III introduces important DAL baselines and datasets. Section IV details the taxonomy of DAL methods. Section V reviews DAL-related applications. Section VI introduces DAL challenges and opportunities. Section VII concludes this article with the conclusions.

II. Paper Collection and Filtering

We first determine relevant keywords used to search articles and create an initial keyword list, as shown in Fig. 4. We perform searches across multiple databases using all possible three-keyword combinations from defined keyword groups, such as "active learning," "machine learning," and "open-set." The databases searched include Google Scholar, Scopus, Semantic Scholar, and Web of Science. We limit the number of papers collected per query to 200, and the publication date ranges from January 2013 to March 2023.

We collect a total of 10 000 research papers from various sources and obtain 3967 unique papers after removing any duplicates. Fig. 4 shows the trend of these articles over time, revealing a growing interest in the topic that we are investigating. To ensure the relevance of the collected articles to DAL, we conduct a detailed manual inspection of their abstracts. As a result, we identify 1273 articles that are considered interesting and pertinent to our study. Based on the collected materials, we employ these keywords to perform a final filtering process and also consider the reputation of conferences or journals in which the papers were published, as well as their impact. This approach further refines our dataset, resulting in 405 articles that are selected for systematic analysis, and 220 articles are finally summarized and discussed, focusing on their key findings and contributions. This rigorous analysis ensures that the articles are relevant and provide valuable insight into the field of DAL.

[1]https://github.com/Clearloveyuan/Awesome-Active-Learning

III. Deep Active Learning

In this section, we first introduce the basic notation and definition of DAL and then discuss the most important DAL baselines based on their relevance and chronological order.

Algorithm 1 DAL Procedure

Input: Unlabeled Data $\mathcal{D}_{\mathbf{pool}}$
Parameter: Batch Size b, Iteration Times T, Query Function α
Output: The final trained model $\mathcal{M}$

$\mathcal{Q}_0 \leftarrow$ Initialization sampling from $\mathcal{D}_{\mathbf{pool}}$ where $|\mathcal{Q}_0| = b$;
$\mathcal{D}^0_{\mathbf{train}} \leftarrow \mathcal{Q}_0$ [Initialization of training dataset];
$\mathcal{M}_0 \leftarrow$ Train $\mathcal{M}_0$ on $\mathcal{D}^0_{\mathbf{train}}$;
while not stop-criterion( ) & i $\leq T$ **do**
  $\mathcal{Q}_i \leftarrow \alpha(\mathcal{M}_{i-1}, \mathcal{D}^{i-1}_{\mathbf{pool}}, b)$ [Annotating b samples];
  $\mathcal{D}^i_{\mathbf{train}} = \mathcal{D}^{i-1}_{\mathbf{train}} \cup \mathcal{Q}_i$; $\mathcal{D}^i_{\mathbf{pool}} \leftarrow \mathcal{D}^{i-1}_{\mathbf{pool}} \backslash \mathcal{Q}_i$;
  $\mathcal{M}_i \leftarrow$ Train $\mathcal{M}_{i-1}$ on $\mathcal{D}^i_{\mathbf{train}}$;
end while

A. Notations and Definitions

We focus on pool-based DAL methods since most DAL methods belong to this category. Pool-based DAL methods iteratively select the most informative samples from a large pool of unlabeled datasets until either the base model reaches a certain level of performance or a predefined budget is exhausted. As shown in Algorithm 1, we use a classification task as an example for illustration, while other tasks follow the typical definition of their task domains. Given an initial labeled training dataset $\mathcal{D}_{\mathbf{train}} = \{\boldsymbol{x}_i, y_i\}_{i=1}^m$ and a large-scale pool of unlabeled data $\mathcal{D}_{\mathbf{pool}} = \{\boldsymbol{x}_i\}_{i=1}^n$, where m$\ll$n, $\boldsymbol{x}_i$ represents the feature vector of the ith sample, and $y_i \in \{0, 1\}$ is the class label for binary classification (or $y_i \in \{1, \ldots, k\}$ for multilabel classification), the DAL procedure is carried out in T iterations. In the ith iteration, a batch of samples $\mathcal{Q}^i$ with batch size b is selected from $\mathcal{D}^{i-1}_{\mathbf{pool}}$ on the basis of the base model $\mathcal{M}$ and an acquisition function $\alpha(\,)$. These samples $\mathcal{Q}^i$ are then labeled by an oracle and added to the ith training dataset $\mathcal{D}^i_{\mathbf{train}}$, with which the model $\mathcal{M}$ is then retrained. DAL terminates when the labeled budget Q is exhausted or the desired performance of the model is reached.

B. Comparisons Between Traditional and Deep AL

The differences between traditional and deep AL mainly lie in the following two aspects.

1) Most traditional AL methods use fixed preprocessed features to calculate uncertainty/representativeness. In deep learning tasks, feature representations are jointly learned with deep neural networks (DNNs). Therefore, feature representations dynamically change during DAL processes, and thus, pairwise distances/similarities used by representativeness-based measures need to be recomputed in every stage. In contrast, for traditional AL with classical ML tasks, these pairwise terms should be precomputed [22].
2) DAL can leverage advanced large-scale PLMs to achieve comparable performance in few-shot or one-shot settings. In contrast, traditional AL methods with few-shot

TABLE I
DETAILED TAXONOMY OF IMPORTANT DEEP ACTIVE LEARNING BASELINES. REFER TO SECTION IV FOR A DETAILED EXPLANATION OF EACH CATEGORY. **ANY TYPES** IN QUERY STRATEGY MEAN THAT THE PROPOSED FRAMEWORKS CAN BE COMBINED WITH ANY TYPES OF DAL QUERY STRATEGIES

Method	Query Strategy	Architecture	Learning Paradigm	Annotation	Training	Tasks
BCBA [2016] [34]	Bayesian	CNNs	Traditional	Hard	Traditional	Image Classification
DBAL [2017] [35]	Bayesian	CNNs	Semi-supervised Learning	Hard	Traditional	Image Classification
CEAL [2017] [36]	Uncertainty	CNNs	Curriculum Learning	Hybrid	Curriculum	Image Classification
ESNN [2017] [37]	Uncertainty	BNNs	Adversarial Learning	Hard	Traditional	Image Classification
PAL [2017] [38]	Uncertainty	BNNs	Reinforcement Learning	Hard	Traditional	Named Entity Recognition
LAL [2017] [39]	Influence	Random Forest	Traditional	Hard	Traditional	Regression Tasks
GAAL [2017] [40]	Uncertainty	GNNs	Adversarial Learning	Hard	Traditional	Image Classification
CoreSet [2018] [41]	Representative	CNNs	Semi-supervised Learning	Hard	Traditional	Image Classification
DFAL [2018] [42]	Uncertainty	CNNs	Adversarial Training	Hard	Traditional	Image Classification
ASM [2019] [43]	Uncertainty	CNNs	Curriculum Learning	Hybrid	Curriculum	Objective Detection
MIAL [2019] [44]	Representative	SVM	Traditional	Hard	Traditional	Image Classification
BatchBALD [2019] [45]	Uncertainty	BNNs	Traditional	Hard	Traditional	Image Classification
DRAL [2019] [46]	Uncertainty	CNNs	Reinforcement Learning	Hard	Pre+FT	Person Re-Identification
DLER [2019] [47]	Uncertainty	PLMs	Transfer Learning	Hard	Pre+FT	Entity Resolution
BGADL [2019] [48]	Hybrid	BNNs	Semi-supervised Learning	Hard	Traditional	Image Classification
VAAL [2019] [49]	Representative	VAE	Adversarial Learning	Hard	Traditional	Image Classification
AADA [2020] [50]	Hybrid	CNNs	Transfer Learning	Hard	Pre+FT	Object Detection
CSAL [2020] [51]	Hybrid	CNNs	Traditional	Hard	Pre+FT	Image Classification
SRAAL [2020] [52]	Uncertainty	CNNs	Adversarial Learning	Hard	Pre+FT	Image Classification
ALPS [2020] [31]	Uncertainty	PLMs	Traditional	Hard	Pre+FT	Cold-start Issue
Ein-Dor et al. [2020] [53]	Any Types	PLMs	Traditional	Hard	Pre+FT	Text Classification
TOD [2021] [54]	Uncertainty	CNNs	Traditional	Hard	Pre+FT	Image Classification
Cluster-Margin [2021] [55]	Representative	CNNs	Traditional	Hard	Pre+FT	Image Classification
LADA [2021] [56]	Uncertainty	CNNs	Semi-supervised Learning	Hard	Traditional	Image Classification
TA-VAAL [2021] [57]	Influence	VAE	Adversarial Learning	Hard	Pre+FT	Image Classification
Karamcheti et al. [2021] [58]	Hybrid	PLMs	Traditional	Hard	Pre+FT	Visual Question Answering
MAML [2022] [59]	Any Types	PLMs	Meta Learning	Hard	Pre+FT	Text Classification
BATL [2022] [32]	Hybrid	PLMs	Traditional	Hard	Pre+FT	Text Classification
TYROGUE [2022] [60]	Hybrid	PLMs	Traditional	Hard	Pre+FT	Text Classification
Schroder et al. [2022] [61]	Uncertainty	PLMs	Traditional	Hard	Pre+FT	Text Classification

or one-shot settings may not meet the minimum requirements for the number of training samples needed to achieve comparable performance [30], [33]. On the other hand, the most similar aspect between traditional and deep AL methods is their utilization of a small number of the most informative samples to train models, thereby improving efficiency and reducing reliance on labeled samples.

C. Important DAL Baselines and Datasets

The most important baselines for DAL are carefully categorized in Table I from six perspectives to provide readers with a complete understanding of the development of DAL and the identification of the most relevant works. These influential studies have achieved breakthroughs in designing new DAL methods, tackling novel tasks, or integrating with emerging learning paradigms. They have been published in influential international conferences or high-quality journals in machine learning, CV, NLP, and so on and have been highly cited with more than 100 total citations or more than ten citations per year.

BCBA [34] pioneers the combination of AL with Bayesian neural networks (BNNs), using Monte Carlo dropout for a variational Bayesian approximation to apply for image classification. Based on this, **DBAL** [35] proposes an uncertainty-based query strategy for high-dimensional image classification. To expand the number of labeled samples without increasing human labors, **CEAL** [36] combines DAL with semisupervised strategies by assigning pseudolabels to high-confidence samples while requesting annotations for the most uncertain samples. Relying on a single query strategy may lead to errors. Thus, **ESNN** [37] uses a deep ensemble of DNNs to measure sample uncertainty from multiple aspects and achieves good robustness for unbalanced datasets. However, the aforementioned methods are criticized for being less effective for batch DAL [45]. To address this issue, **CoreSet** [41] selects informative batches that cover the whole data distribution, and **BatchBALD** [45] uses mutual information to identify the most informative batches. **Cluster-Margin** [55] aims to select informative and diverse minibatches to improve accuracy and efficiency.

To better help DAL adjust to different tasks, reinforcement learning provides detailed rewards for dynamically controlling query strategies. For example, **PAL** [38] learns a deep reinforcement learning-based Q-network as an adaptive policy to select data samples for labeling. Similarly, **DRAL** [46] uses a reinforcement learning framework to dynamically adjust the acquisition function via rewards to obtain high-quality queries. **UCBVI** [62] provides a new modification to the Q-network formulation for reward-free exploration, significantly reducing query complexity. However, reinforcement learning requires a large amount of training data and human-designed rewards,

which is difficult for many real-world applications. To address this issue, meta learning and transfer learning have become main solutions. **LAL** [39] trains a regressor to learn optimal query strategies for downstream tasks. **MAML** [59] combines meta learning and DAL by initializing an active learner with meta-learned parameters obtained through meta training on tasks similar to the target task during DAL. **DLER** [47] designs an architecture to learn a transferable model from a high-resource setting to a low-resource one, allowing DAL to select a few informative samples based on the knowledge of the source domain. **AADA** [50] jointly considers domain alignment, uncertainty, and diversity for sample selection.

To enlarge the labeled training dataset for DNNs without incurring additional human labor costs, semisupervised, semisupervised, and self-supervised DAL methods have been proposed. **MIAL** [44] pioneers semisupervised DAL using cluster-based strategies to measure sample informativeness. **ASM** [43] collaborates with self-learning and DAL, designing a selector function to selectively and seamlessly determine the confidence of the samples, where high-confidence samples are labeled by a pseudolabeling module, and low-confidence samples are labeled by humans. **CSAL** [51] first uses semisupervised learning to distill information from unlabeled data during the training stage and then uses consistency-based sample selection for DAL. **TOD** [54] leverages a novel unlabeled data sampling strategy for data annotation in conjunction with a semisupervised training scheme to improve the performance of the task model with unlabeled data. Recently, data augmentation has expanded to become a deep neural model that generates virtual instances to help expand training datasets. **GAAL** [40] introduces a generative adversarial network to the DAL query method to generate informative samples to train the model. **BGADL** [48] expands GAAL and combines generative adversarial DAL with Bayesian data augmentation to generate diverse and informative samples. **DFAL** [42] uses adversarial DAL to select samples close to the decision boundary as the most informative samples for DAL. **VAAL** [49] learns a latent space using a variational autoencoder (VAE) to generate new informative samples and trains an adversarial network to discriminate labeled and unlabeled data. Inspired by these works, **TA-VAAL** [57] incorporates a learning loss prediction module and a task ranker to enable task-aware sample selection. **SRAAL** [52] proposes a relabel adversarial model that aims to obtain the most informative unlabeled samples. **LADA** [56] anticipates data augmentation impact by scoring both real and virtually augmented instances, allowing training in informative labeled and augmented data.

Large-scale PLMs achieve great success and become a milestone in artificial intelligence. Due to sophisticated pretraining objectives and huge model parameters, large-scale PLMs effectively capture knowledge from massive labeled and unlabeled data. DAL also ushers in a new paradigm by leveraging the prior knowledge in PLMs to enable few-shot or zero-shot learning for many downstream tasks. **ALPS** [31] extracts knowledge from PLMs to select the first batch of data using masked language modeling loss, which successfully solves the cold-start problem of DAL. **Ein-Dor et al.** [53] use multiple DAL methods to select samples for fine-tuning in BERT-based text classification. It achieves comparable or higher performance than fine-tuning on full datasets only with 10%∼20% labeled samples. **Karamcheti et al.** [58] use DAL to identify and remove noisy data, select balanced samples to fine-tune PLMs, and achieve better performance in visual question-answering. **BATL** [32] is a task-independent batch acquisition method on PLMs with triplet loss to determine hard samples, which have similar features but difficult to identify labels in an unlabeled data pool. **TYROGUE** [60] designs an interactive DAL framework to flexibly select samples to fine-tune PLMs for multiple low-resource tasks. **Schroder et al.** [61] extend the PLMs using available unlabeled data for greater adaptability and introduce effective fine-tuning for the robustness of DAL in low-resource and high-resource settings.

TABLE II
WIDELY USED DAL DATASET INFORMATION

Dataset	Size	Domain	Tasks
MNIST [6]	70,000	Images	Classification
CIFAR-10 [63]	60,000	Images	Classification
SVHN [64]	600,000	Images	Classification, Localization
ImageNet [65]	1.2M	Images	Classification, Detection
MSCOCO [66]	123,287	Images	Object detection
Cityscapes [67]	5,000	Images	Semantic segmentation
Caltech-101 [68]	9,000	Images	Classification
SST [69]	11,855	Text	Sentiment analysis
TREC [70]	5,952	Text	Question answering
SNLI [71]	570,000	Text	Natural language inference
IMDB [72]	50,000	Text	Sentiment analysis
AGNews [73]	31,900	Text	Classification
PubMed [74]	19,717	Text	Document classification
YouTube-8M [75]	237,000	Audio	Classification
MIMIC-III [76]	112,000	Medical	Healthcare analytic

As shown in Table II, we also conclude the most widely used datasets in DAL including images, text, and audio.

IV. TAXONOMY OF DAL

A. Annotation Type

1) Hard Annotations: These provide one or multiple discrete categorical labels independently for each sample. For example, Citovsky et al. [55] annotate each image with a specific label such as "balloon" or "strawberry" for an image classification task. Wiechman et al. [77] design an online annotation system to assign multiple labels to long documents based on their sentiments, topics, and spam/nonspam status.

2) Soft Annotations: These allow continuous and subjective labels for samples. For instance, ReDAL [78] annotates continuous 2-D region labels for 3-D point clouds in semantic segmentation. Kothawade et al. [79] use mutual information as an auxiliary metric to select annotation regions in images for autonomous vehicles. Xie et al. [80] propose a region-based approach to automatically query a small subset of image regions to label while maximizing segmentation performance.

3) Hybrid Annotations: These combine automatic pseudolabels of high-confidence predictions with human labeling of low-confidence samples in an iterative self-paced manner [43]. For example, Wang et al. [36] propose a complementary sample selection strategy to progressively choose the most informative samples, pseudolabeling high-confidence predictions for training. Yu et al. [81] jointly use the expertise of different annotation groups, interrelations between workers, and label correlations within groups. By weighting groups, they reduce the impact of low-quality workers and calculate reliable consensus labels.

4) Explanatory Annotations: These provide a hard or soft label along with an explanation for each annotation. For example, Schröder et al. [82] use topic-related annotations for environmental texts. Similarly, Yan et al. [83] annotate

the text and list keywords as evidence of the accuracy of the label. Unlike the above methods, Zhou et al. [84] annotate samples by minimizing correlations between tasks and provide explainable medical knowledge to distinguish selected samples.

5) Random/Multiagent Annotations: These use multiple independent pseudoannotators to randomly label new unlabeled samples without human input [85]. For example, Gong et al. [86] use an agent team to collaboratively select informative images for annotation based on the decisions from the other agents.

B. Query Strategy

1) Uncertainty-Based Methods: These aim to select the most ambiguous samples according to model predictions. Given an input $\boldsymbol{x}_i$

$$\text{Entropy}(\boldsymbol{x}_i) = \arg\max_{\boldsymbol{x}_i} \left(\sum_j P(\hat{y}_j|\boldsymbol{x}_i) \log P(\hat{y}_j|\boldsymbol{x}_i) \right) \quad (1)$$

where $P(\hat{y}_i|\boldsymbol{x}_i)$ represents the likelihood that $\boldsymbol{x}_i$ is classified into the ith class [87]. Uncertainty-based methods focus on designing various score functions to measure sample uncertainty and informativeness, including predictive entropy [87], least confidence [88], highest estimated dual variables [89], and mutual information between model posterior and predictions [79]. Some strategies check samples near the decision boundary as the most uncertain ones [90], such as instances close to the hyperplane [44] or close to the margin [91]. Others combine multiple query strategies, forming a query-by-committee [92] or disagreement-based [93] DAL strategy to decrease errors made by a single query strategy. With the development of adversarial learning, instead of selecting samples from unlabeled datasets, models tend to generate the most informative and uncertain synthetic samples to expand the training dataset [48].

However, they have some common drawbacks: 1) redundant samples, as uncertain points, are continually selected yet in short of coverage; 2) simply focusing on a single sample lacks robustness to outliers; and 3) these task-specific designs exhibit limited generalizability.

2) Representative-Based Methods: These aim to sample the most prototypical data points that effectively cover the distribution of the entire feature space. Existing methods can be categorized into density- and diversity-based approaches. **Density-based** methods prefer to select samples that can represent all unlabeled samples. They use clustering methods to select cluster centers [94] as the most representative samples or select samples that can maximize probability coverage of the whole feature space of unlabeled datasets [41]. For example, Kim and Shin [95] design the density awareness CoreSet approach to estimate sample densities and preferentially select diverse points from sparse regions. Given the input $\boldsymbol{x}_i$

$$\text{Density}(\boldsymbol{x}_i) = \frac{1}{k} \sum_{j \in \mathcal{N}(\boldsymbol{x}_i, k)} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 \quad (2)$$

where $\mathcal{N}(\boldsymbol{x}_i, k)$ represents the k-nearest neighbors of $\boldsymbol{x}_i$ [95]. Coleman et al. [96] and Gudovskiy et al. [97] achieve efficiency by only considering nearest neighbors rather than all data or matching feature densities with self-supervised methods. **Diversity-based** methods prefer to select samples that are different from the labeled samples. They use context-sensitive methods [98] that take into account the distance between a sample and its surrounding labeled samples to enrich the diversity of the labeled dataset. BMAL [99] performs DAL for the image labeling problem, where diversity is measured by the KL divergence of the class probabilities distribution of similar neighboring instances, formulated as

$$\text{Divergence}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_j P(\hat{y}_j|\boldsymbol{x}_i) - P(\hat{y}_j|\boldsymbol{x}_j) \log \frac{P(\hat{y}_j|\boldsymbol{x}_i)}{P(\hat{y}_j|\boldsymbol{x}_j)}. \quad (3)$$

Other diversity-based methods tend to train a model, such as adversarial networks [57], contrastive networks [100], hierarchical clustering [44], and pretrained models [53], to help discriminate labeled and unlabeled sets and select the most different unlabeled samples. For example, Li et al. [101] explicitly learn a nonlinear embedding to select representative samples. Parvaneh et al. [102] explore neighborhoods around unlabeled data by interpolating features with labeled points. Li et al. [103] propose an acquisition function that measures mutual information between a batch of queries to encourage diversity. To further increase label efficiency, Citovsky et al. [55] use hierarchical clustering to diversify batches, requiring only 40% of the labels to achieve the same target performance. However, since they use ResNet-101 as their backbone, which contains only 170-MB parameters, more than 20% labeled samples are required for fine-tuning the model.

However, the aforementioned representative-based methods, which solely focus on sampling diverse samples, are always insensitive to samples that are close to the decision boundary (excluding hybrid methods that jointly consider representative and uncertainty), despite the fact that such samples are probably more important to the prediction model, as suggested by Zhao et al. [104]. In addition, representative-based methods work well for a small sample of data and classifiers with a small number of classes since their computational complexity is almost quadratic with respect to data size [55].

3) Influence-Based Methods: These aim to select samples that will have the greatest impact on the performance of the target model. These techniques can be categorized into three main groups.

1) The first group is directly measuring the expected impact on the modal through metrics such as gradient norm [105], query complexity [106], kernel approximation [107], KL divergence [97], change of loss function [108], or model parameters [54], and expected error reduction (EER) [109]. Specifically, EER can be formulated as

$$\text{EER}(\boldsymbol{x}_i) = \mathbb{E}_{\boldsymbol{x}_s} \left\{ \mathbb{E}_{y_i|\boldsymbol{x}_i} [\max_{y_s} p(y_s|\boldsymbol{x}_s, \boldsymbol{x}_i, y_i)] - \max_{y_s} p(y_s|\boldsymbol{x}_s) \right\} \quad (4)$$

where $\boldsymbol{x}_s$ refers to the labeled sample.

2) The second group incorporates different learning policies, such as reinforcement learning and imitation learning, to select samples based on reward signals or demonstrated actions. Despite the promising advantages, this requires significant additional training [110]. For

example, Wertz et al. [111] propose reinforced DAL, a reinforcement learning policy that uses multiple elements of the data and the task to dynamically pick the most useful unlabeled subset during the DAL process.

3) The last group is training a separate model to estimate the impact on the target model [89]. For example, Peng et al. [14] propose a knowledge distillation framework to evaluate the impact of samples based on the knowledge learned by the student model. Elenter et al. [89] use the dual variables of the original model to measure the impact on the target model.

However, despite recent advances, influence-based DAL remains challenging. Directly measuring model changes or incorporating new learning policies always requires huge time and space costs, and training a new model will overrely on its accuracy and often lead to unstable results.

4) Bayesian Methods: These aim to minimize classification errors and improve model beliefs by leveraging Bayes' rule. Most studies have treated Bayesian models (e.g., Gaussian process [109], BNNs [35], and Bayesian probabilistic ensemble [112]) as uncertainty-based methods, using them to estimate the informativeness of the sample. However, Bayesian DAL is better viewed as its own distinct system, with methods that select batches by directly measuring impact on the target model, such as BatchBALD [45] and Causal-BALD [113]. For example, we define a Bayesian model with model parameters $\boldsymbol{w} \sim p(\boldsymbol{w}|\mathcal{D}_{\text{train}})$, and BALD can be defined to estimate the mutual information between the model predictions and the model parameters, formulated as

$$\begin{aligned}\mathbb{I}(y; \boldsymbol{w}|\boldsymbol{x}, \mathcal{D}_{\text{train}}) = \mathbb{H}(y|\boldsymbol{x}, \mathcal{D}_{\text{train}}) \\ - \mathbb{E}_{p(\boldsymbol{w}|\mathcal{D}_{\text{train}})}[\mathbb{H}(y|\boldsymbol{x}, \boldsymbol{w}, \mathcal{D}_{\text{train}})] \quad (5)\end{aligned}$$

where $\mathbb{H}$ represents the entropy and $\mathbb{E}$ is the expectation.

Compared to standard DNNs, the aforementioned Bayesian DAL methods, which leverage the advantages of probabilistic graphical theory [35], can often provide reasonable explanations for why these samples should be selected [45]. However, they often require extensive accurate prior knowledge and tend to underperform deep learning models in representation learning and fitting capacity.

5) Hybrid Methods: These aim to take advantage of the above multiple query strategies and achieve a tradeoff among them. Hybrid methods can be further categorized according to interaction patterns. **Serial-form hybrids** apply criteria sequentially within a DAL cycle, filtering out noninformative samples until the batch is filled [55]. **Criteria-selection hybrids** use only one query strategy in one DAL iteration, in which they select the best query strategy or network architecture with the highest criterion. For example, DUAL [114] switches between density- and uncertainty-based selectors to choose the best criterion for each DAL cycle. Unlike DUAL, iNAS [115] searches a restricted candidate set to find the optimal model architecture incrementally in each DAL iteration. **Parallel-form hybrids** use multiobjective optimization methods or a weighted sum to merge multiple query criteria into one for sample selection. For example, Gu et al. [2] efficiently acquire batches with discriminative and representative samples by proposing procedures to update labeled and unlabeled sets based on path-following optimization techniques. Citovsky et al. [55] jointly optimize the uncertainty and diversity criteria in batch mode using multiobjective acquisition functions. TOD [54] selects samples with high model uncertainty and outputs discrepancy through a weighted combination of both metrics.

Hybrid methods combine the advantages of different query strategies. However, determining the most effective combinations and tradeoffs between criteria is time-consuming and still remains open for further investigation.

C. Model Architecture

1) Traditional Machine Learning: Architectures, such as forest [39] and support vector machine (SVM) [44], are statistical-based models that do not use neural networks. They attract great attention in the early stage of the DAL development.

2) Bayesian Neural Networks: BNNs combine neural networks with Bayesian inference, quantifying the uncertainty introduced by the models in terms of outputs and weights to explain the trustworthiness of the prediction [116]. Many studies propose DAL strategies based on BNNs, aiming to improve efficiency and explainability in samples' selection [38], [45].

3) Recurrent Neural Networks: RNNs [117] use their reasoning from previous experiences to predict upcoming events and are able to learn features with long-term dependencies. They have been widely used for sequential data such as text and audio. DAL is seldom combined with RNNs since they require large-scale labeled datasets for training. Some special tasks that easily recognizable patterns, such as malicious word detection on social networks [118], can be solved with DAL.

4) Convolutional Neural Networks: CNNs [6] are feedforward neural networks that can extract features from data with convolution structures and have been widely used for image processing with three advantages: local connections, weight sharing, and downsampling dimensionality reduction. DAL can be effectively combined with CNNs since Sener and Savarese [41] proved that a subset of samples (CoreSet) can geometrically characterize all features of the entire image set and can be selected by minimizing a rigorous bound. Following their study, more studies have been conducted [49], [55].

5) Graph Neural Networks: GNNs [5] learn node representations by aggregating neighborhood information and achieve great success in various tasks, such as node classification. However, effectively handling graph data with dense interconnections between samples using limited labeled data remains an open challenge [119]. DAL can help address this by selectively querying labels for the most informative samples and executing only one training epoch to reduce the annotation cost for various types of graphs, such as homogeneous graphs [120], heterogeneous graphs [121], and attribute graphs [122].

6) Variational Autoencoders: VAE is a class of neural network architecture designed with an encoder–decoder framework [123]. It aims to capture the underlying data distribution and learn to generate samples that closely resemble the input data. VAEs-based DAL methods usually generate samples to fool discriminators in an adversarial training manner, thus improving discriminators' ability to select the most challenging-to-distinguish samples for training DAL models [49], [57].

7) Pretrained Language Models: These, based on transformers, utilize multihead self-attention to capture long-term dependencies. By pretraining on large unlabeled corpora, PLMs embed substantial general knowledge and transfer to

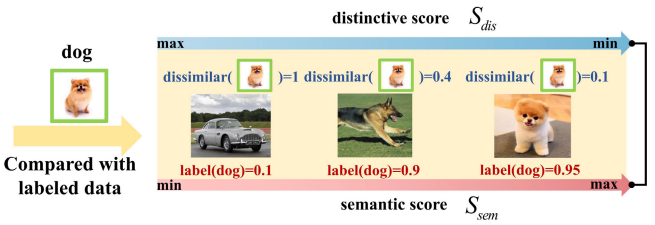


Fig. 5. Example for contrastive learning-based query strategies.

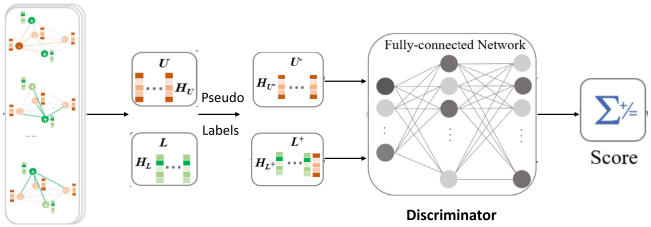


Fig. 6. The detailed processes of SEAL [122] method.

downstream tasks, enabling state-of-the-art (SOTA) performance [30]. For example, Seo et al. [32] identify the most informative samples for a given task, focusing on PLMs fine-tuning, to learn salient patterns with minimal annotation cost. The combination of a pretraining rich knowledge foundation and DAL's sample-efficient tuning unlocks PLMs' further potential for many applications.

D. Learning Paradigm

1) Traditional Learning Paradigm: This, as illustrated in Algorithm 1, iteratively queries and labels samples to train the models in a vanilla supervised learning manner, without incorporating any advanced learning paradigms [32], [34].

2) Semisupervised Learning: This, also known as weakly supervised learning, aims to jointly use real-labeled samples and pseudolabeled samples to train the models. Current DAL methods are designed with various efficient strategies to obtain pseudolabels for unlabeled samples. For instance, DBAL [35] and CoreSet [41] first predict pseudolabels using their models and then calculate samples' confidence scores to judge whether these pseudolabels should be trusted or not. On the other hand, LADA [56] and BGADL [48] propose new data augmentation methods to create more samples based on original labeled samples using their original real-labeled samples as pseudolabels. These studies effectively reduce human labor and achieve comparable performance compared with traditional supervised learning using larger labeled samples.

3) Contrastive Learning: This improves feature representation by pulling similar instances closer together while pushing dissimilar instances apart [124]. Contrastive methods extract discriminative features, such as semantics [100] and distinctiveness [57], to estimate the sample uncertainty during acquisition. For example, as shown in Fig. 5, Du et al. [125] extract both semantic and distinctive features with contrastive learning and then combine them in a query strategy to choose the most informative unlabeled samples with matched categories.

4) Adversarial Learning: This enables a model to train fully differentiable by solving minimax optimization problems [49]. This approach can be used as a generative query technique for DAL. For example, DAL can be combined with the generative adversarial network, which consists of a generator and a discriminator, where the DAL model acts as the discriminator and the generator explores the distribution of unlabeled data to generate the most informative and uncertain synthetic samples for training [57]. Li et al. [122] propose SEAL, as shown in Fig. 6, which consists of two adversarial components. The graph embedding network encodes all nodes into a shared space, with the intention of making the discriminator treat all nodes as labeled. In addition, a semisupervised discriminator is used to differentiate unlabeled nodes from labeled ones. The divergence score of the discriminator is used as an informativeness measure to actively select the most informative node for labeling. The two components form a loop to mutually improve DAL.

5) Meta Learning: This enables DNNs to leverage the knowledge acquired from multiple tasks, represented in the network with their weights, to adapt faster to new tasks. Meta learning can provide an acquisition function for DAL [39], [126] or favorable model initialization during DAL by controlling the transfer of knowledge from multiple source tasks. For example, Shao et al. [127] propose a learning-to-sample, where a boosting model and sampling model dynamically learn from each other and iteratively improve performance. Zhu et al. [59] combine both paradigms by initializing an active learner with meta-learned parameters via meta training on tasks similar to the target task.

6) Reinforcement Learning: This involves an agent that can interact with its environment and learn to alter its behavior in response to received rewards [119]. Given that almost all DAL methods use heuristic acquisition functions with limited effectiveness, reinforcement learning frames DAL as a reinforcement learning problem to explicitly optimize an acquisition policy. In the DAL with reinforcement learning setup, an autonomous agent (acquisition selector) controlled by a deep learning algorithm observes a state s_t from its environment (predictor) at time t. It takes an action a_t to maximize the reward r_t (prediction accuracy), where a_t decides whether to query unlabeled samples [62].

7) Curriculum Learning: This mimics human and animal learning processes, where the training progresses gradually from simple to complex samples. This provides a natural way to exploit labeled data for robust learning [10], [128]. Specifically, curriculum learning uses a predefined learning constraint to incrementally incorporate additional labeled samples during training. Curriculum learning introduces a weighted loss on all labeled samples, acting as a general regularizer over the sample weights. For example, Lin et al. [129] use a pseudolabel strategy that iteratively assigns pseudolabels to unlabeled samples with high prediction confidence.

8) Continual Learning: This is developed for constraints on task-based settings, where the model continuously learns a sequence of tasks one at a time, where all data for the current task are labeled and available in increments. However, real-world systems do not have the luxury of large labeled datasets for each new task. To address this issue, Mundt et al. [130] present a detailed analysis of continual learning-based DAL and out-of-distribution detection works. They suggest a unified perspective with open-set recognition as a natural interface between continual learning and DAL. Ayub and Fendley [30] develop a method that allows an agent to continually learn new object classes from a few labeled examples.

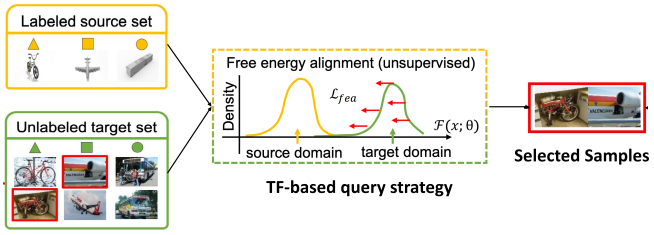


Fig. 7. Example for transfer learning-based query strategies.

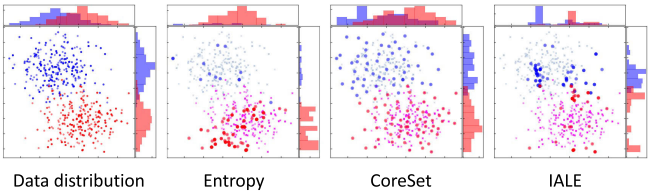


Fig. 8. Example for imitation learning [131].

9) Transfer Learning: This extracts knowledge from one or more source tasks and applies it to a target task. It has two broad categories: transductive and inductive. While transductive methods adapt models learned from a labeled source domain to a different unlabeled target domain with the same task, inductive methods ensure that the domains of source and target are the same but tasks are different. DAL with transfer learning can better enhance each other's performance by selecting the best target samples with a distribution similar to the source domain [50]. In addition, transfer learning can minimize the number of annotation labels needed and provide auxiliary information for DAL acquisition functions. For example, as shown in Fig. 7, Xie et al. [87] propose an energy-based active domain adaptation that balances domain representation and uncertainty when selecting target data.

10) Imitation Learning: This provides SOTA results in many structured prediction tasks by learning near-optimal search policies [92]. Such methods assume access to an expert during training that can provide the optimal action in any queried state, essentially asking "what would you do here?" and learning to mimic that choice. For example, Bullard et al. [132] use imitation learning to allow an agent in a constrained environment to concurrently reason about both its internal learning goals and externally impose environmental constraints within its objective function. Löffler and Mutschler [131] propose an imitation learning scheme (IALE) that mimics the selection of the best-performing expert heuristic at each stage of the learning cycle in a batch-mode setting. As shown in Fig. 8, IALE can well imitate the entropy- and CoreSet-based methods and, thus, obtain better performance.

11) Multitask Learning: MTL focuses on formulating methods to maintain performance across multiple tasks rather than a single task. Multitask DAL (MTAL) methods combine multiple individual task-related query strategies into a single unified approach and jointly optimize the unified one. In contrast to single-task query settings, where the uncertainty of a single selected task classifier is used to query unlabeled samples, in MTAL, the uncertainty of an instance is determined by the uncertainties from classifiers across all tasks. For example, Ikhwantri et al. [133] propose an MTAL framework for semantic role labeling with entity recognition as an auxiliary task. This alleviated data needs and leverages entity information to aid role labeling. Their experiments show that MTAL can outperform single-task DAL and standard MTL, using 12% less training data than passive learning. Zhou et al. [84] propose a multitask adversarial DAL framework, where adversarial learning maintains the effectiveness of the MTL and DAL modules. A task discriminator eliminates irregular task-specific features, while a diversity discriminator exploits heterogeneity between samples to satisfy diversity constraints.

E. Training Process

1) Traditional Training: This first trains a model on an initialized training dataset and then selects unlabeled samples to annotate based on the predictions of the current model. The newly annotated samples are added to the training set for retraining the model in the next iteration [134]. This iterative process continues, with the model parameters randomly reinitialized before each epoch of retraining [36], until either the sample budget or the number of DAL iterations is reached.

2) Curriculum Learning Training: This gradually progresses from easy to complex samples, mimicking human and animal learning processes. This provides a natural and iterative way to exploit labeled data for robust learning. For example, Tang and Huang [135] propose a self-paced DAL approach that jointly considers the value and difficulty of a sample. It queries samples from easy to hard to minimize annotation costs. Wang et al. [43] show that curriculum learning alone improves the accuracy of object detection by 3.6%, while the combination of curriculum learning and DAL improves the accuracy by 4.3%.

3) Pretraining and Fine-Tuning: These have become a primary training process with the development of large-scale PLMs [58]. It leverages the rich prior knowledge in PLMs to solve different downstream tasks. DAL attracts attention as a sample selection strategy for fine-tuning with only 10%∼20% of labeled data achieving competitive performance compared to full data fine-tuning [32]. DAL iteratively selects and annotates batches of informative samples to fine-tune the PLMs for the downstream task. This satisfies task-specific needs while also enabling a few-shot learning [30].

V. Applications of DAL

As shown in Table III, the integration of DL and AL is leading to an increasing application of AL methods in various domains of life, ranging from agricultural development [82] to industrial revitalization [82] and from artificial intelligence [137] to biomedical fields [160]. In this section, we aim to provide a systematic and detailed overview of existing DAL-related work from a broad application perspective.

A. Applications in NLP

With the emergence of large-scale language models, NLP has achieved great success using computers to help understand intricate languages. However, fine-tuning these language models requires a substantial amount of data, computation resources, and time. DAL provides a strategy for searching for high-quality small and high-quality samples to help fine-tune the model and save resources. In the following, we introduce some of the most influential DAL methods in NLP.

TABLE III
ILLUSTRATION OF DAL-RELATED APPLICATIONS IN MAIN FIELDS, INCLUDING CLASSIC METHODS WITH THEIR ADVANTAGES AND DISADVANTAGES

Areas	Applications	Classic Methods	Advantages	Disadvantages
NLP	Text Classification	generate samples for training [83], [136]. uncertainty sampling [61]. use pre-trained language models [137].	make the selection process efficient. high efficiency and performance. easily adapt to new datasets.	high time consumption, unstable performance. vulnerable to outliers, unstable performance. vulnerable to outliers and imbalanced datasets.
	Text Summarization	PLMs with Monte Carlo dropout [138]. diverse sampling [139].	efficient and effectiveness. remove outliers and diverse sampling.	vulnerable to outliers, unstable performance. vulnerable to document embeddings.
	Question Answering	DataMap [58]. interactive query strategy [140].	eliminate outliers and improve accuracy. efficiently minimize costly data annotations.	high time consumption, lack of generalizability. wait for human reaction, need expert knowledge.
	Information Extraction	label identical subsequences [141] label most novel words [142].	high efficiency and effectiveness. high efficiency and effectiveness.	lack of generalizability, cold-start. unstable performance, cold-start.
	Semantic Parsing	hyperparameter selection [143]. hybrid query strategies [144].	reduce data annotation. select the most semantically varied samples.	high time consumption, lack of generalizability. vulnerable to outliers, lack of scalability.
CV	Image Captioning	semantic adversarial DAL [145] domain transfer learning [146].	overcome scarcity of labeled data. transfer knowledge from high-resource.	difficulty in cross-domain transfer, cold-start. vulnerable to outliers, data scarcity.
	Semantic Segmentation	uncertainty-based DAL [147]. region-based selection [80], [148].	high efficiency and effectiveness. balance between label efforts and effect.	unstable performance, easily select outliers. vulnerable to outliers, imbalance datasets.
	Object Detection	hybrid selection [43], [149]. instance uncertainty learning [150].	avoid noisy samples and outliers. suppress noisy instances.	data scaricity, unstable performance. unstable performance, lack of scalability.
	Pose Estimation	traditional DAL strategy [151], [152]. meta learning [86].	effectiveness, easy to apply. can learn an optimal sampling policy.	vulnerable to outliers, cold-start. vulnerable to outliers and imbalance datasets.
	Target Tracking	multi-frame collaboration [153] multi-target object tracking [154].	eliminate background noise, ensure diversity. high efficient and effectiveness.	unstable performance, lack of scalability. high time consumption, cold-start
	Person Re-identification	human-in-the-loop [46]. incremental annotation [155].	improve model performance. select diverse samples without redundancy.	high time consumption, lack of generalizability. vulnerable to outliers, cold-start.
DM	Node Classification	semi-supervised adversatial DAL [122]. graph policy network [120].	better performance gains. stable performance.	unstable performance, cold-start. single sample selection costs much time.
	Link Prediction	multi-view DAL [156]. transfer learning DAL [157].	query informative samples from multi-view. easily apply to new datasets.	lack of scalability and generalizability. unstable performance, cold-start.
	Community Detection	topic-based [158]. geometric block model [159].	reducing the unreliable dataset. efficient and effectiveness.	high time consumption, unstable performance. unstable performance, cold-start.

1) Text Classification: This aims to classify large-scale text with particular labels such as topic or sentiment. Researchers propose several methods to efficiently select informative samples for training. For example, Yan et al. [83] generate the most informative examples for training, efficiently skipping the sample selection process. They approximate the generated example with a few summary words, which significantly reduces the labeling cost for annotators, as they only need to read a few words instead of a long document. Tan et al. [136] develop the Bayesian estimate of mean proper scores (BEMPS) framework for DAL, which allows the calculation of scores such as logarithmic probability to better help select informative and uncertainty samples. Experiments demonstrate that BEMPS is more effective than baselines in various text classification datasets. On the other hand, Schröder et al. [61] use transformers for uncertainty-based sample selection. Interestingly, they achieve comparable performance in widely used text classification datasets while training in less than 20% of the labeled data, which demonstrates their ability to utilize limited labeled data. In another study, Jelenic et al. [137] conduct an initial empirical study to investigate the transferability of the DAL by using PLMs. They find that DAL can effectively adapt to new datasets with pretrained models.

2) Abstractive Text Summarization: Abstractive text summarization (ATS) aims to compress a document into a brief, informative, and readable summary that retains the key information of the original document. However, constructing human-annotated datasets is a time-consuming and costly endeavor. DALs are explored to reduce the amount of annotation needed while achieving a certain level of ATS performance. For example, Gidiotis and Tsoumakas [138] address the issue from a Bayesian view and study uncertainty estimation for SOTA text summarization models. They augment the pretrained summarization models with Monte Carlo dropout, forming the corresponding variational Bayesian PLM models. By generating multiple summaries from these models, they approximate Bayesian inference and estimate the summarization uncertainty. Experiments on multiple benchmark datasets consistently demonstrate their improved summarization performance with higher Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores. Unlike the above method, as shown in Fig. 9(a), Tsvigun et al. [139] propose an alternative query strategy for ATS based on diversity principles. This strategy, known as in-domain diversity sampling, involves selecting instances that are dissimilar from annotated documents but similar to the core documents of the domain. Given a limited annotation budget, they can improve model performance and consistency scores.

3) Question Answering: This involves answering questions about images or passages of text [161]. However, current models require large-scale training data to achieve high performance. DAL methods, such as Datamap [58] and hierarchical dialogue policies [140], are designed to maximize performance with minimal labeling effort. Specifically, in Fig. 9(b), DataMap [58] is able to detect and eliminate outlier examples from the unlabeled set, resulting in a significant increase in model accuracy with fewer labeled examples. Padmakumar and Mooney [140] develop a joint policy for clarification and DAL in an interactive image retrieval task. Asking users for clarification while querying new examples improves the model performance.

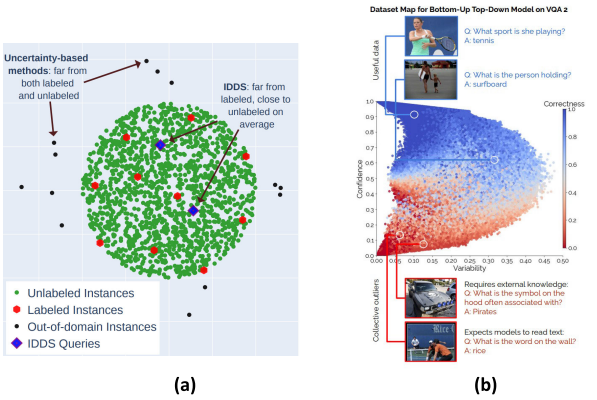


Fig. 9. Example for samples' selection of ATS and datamap. (a) ATS samples' selection. (b) Datamap for samples' selection.

4) Information Extraction: This refers to many NLP tasks, including named entity recognition, keyword extraction, word segmentation, and so on. Manual annotation of large-scale sequences is time-consuming, expensive, and, thus, difficult to realize. To address this, Brantley et al. [92] design a new DAL annotation manner. They use a noisy heuristic labeling function to provide initial low-quality labels, train a classifier to decide whether to trust these labels, and annotate the most uncertain samples with trustable labels. Their model achieves high efficiency and effectiveness on many information extraction tasks. Similarly, Radmard et al. [141] focus on improving the efficiency of DAL for naming entity recognition by querying subsequences within each sentence and propagating labels to unseen identical subsequences in the dataset. They demonstrate that the DAL strategy requires only 20% of the dataset to achieve the same results as training on the full dataset. Hua and Wang [142] propose two model-independent acquisition strategies for identifying and understanding the structure of argumentative discourse, achieving competitive results with fewer computations The former selects samples with the most novel words for labeling, while the latter seeks to identify more relation links by matching any of the 18 prominent discourse markers from a manual.

5) Semantic Parsing: This aims to convert a natural language utterance to a logical form: a machine-understandable representation of its meaning [162]. DAL can help reduce data requirements and improve efficiency for semantic parsing. For example, Duong et al. [143] design a simple hyperparameter selection technique for DAL to accelerate data annotation. Experiments show that their method significantly reduces the need for data annotation and improves the model's performance on semantic parsing. Li and Haffari [163] also design a hyperparameter tuning module to reduce the additional annotation cost. In addition, they design a novel query strategy that prioritizes examples with various logical form structures and more lexical choices, which further improves the performance of semantic parsing. Li et al. [144] propose a novel DAL method with two new annotation manners, called HAT. Experiments show that HAT can pick out the most semantically varied and illustrative utterances, leading to the highest possible gains in parser performance.

B. Applications in CV

With the remarkable success of CNNs and vision transformers, a valuable insight has been gained that more labeled image datasets can promote to obtain better performance of the task. However, as the amount of data increases, training DNNs becomes time-consuming and resource-consuming. In addition, even if the number of data increases, the presence of noise often leads to limited performance improvement. DAL can effectively reduce noise and time consumption in many CV tasks. Hereafter, we provide detailed information on specific tasks and their improvements achieved with DAL in CV.

1) Image Classification: This aims to accurately classify images based on the provided labels for many specific fields such as remote sensing [16], medical imaging [164], and face recognition [129]. We list the most successful DAL methods for image classification in Section III-C, such as BCBA, DBAL, and CEAL, which can be referred to for more detailed information.

2) Image Captioning: This aims to automatically generate descriptive text about the content of an image. Achieving high-quality captioning requires large-scale datasets with diverse images. Unfortunately, creating such a dataset is time-consuming and costly. To tackle this issue, Zhang et al. [145] devise a novel adversarial DAL model, which uses visual and textual information to select the most representative samples to optimize the performance of image captioning. Experiments show that they overcome the limitations of labeled data scarcity and improve the practicality and effectiveness of image captioning. In a similar vein, Cheikh and Zrigui [146] introduce a knowledge-transferable DAL framework for low-resource datasets. They take advantage of existing datasets, translate their captions into Arabic, and train the model with translated caption datasets as prior knowledge for low-resource ArabicFlickr1K datasets (which contain only 1095 images). Their model achieves the bilingual evaluation understudy (BLEU) score of 47%, serving as compelling evidence for the effectiveness of their approach.

3) Semantic Segmentation: This aims to understand images at the pixel level, serving as the basis for various applications, including autonomous driving [80] and robot manipulation [30]. However, training segmentation models requires an extensive amount of data with pixelwise annotations, a process that is burdensome and prohibitively expensive [78]. To solve this challenge, Konyushkova et al. [147] propose an uncertainty-based DAL method with geometric priors to expedite and simplify the annotation process for image segmentation. Experiments show that their method can be applied to both background-foreground and multiclass segmentation tasks. Qiao et al. [148] introduce a collaborative panoptic regional DAL framework for partial annotated semantic segmentation. By incorporating semantic-agnostic panoptic matching and region-based selection and extension, their model strikes a balance between labeling efforts and performance. Similarly, Xie et al. [80] propose an automated region-based DAL approach for semantic segmentation considering the spatial adjacency of image regions and the confidence in prediction. Experiments show that they can use a small number of labeled image regions while maximizing segmentation performance.

4) Object Detection: This is transformed into a region classification task by generating candidate regions of objects from the input image. Features are typically extracted from candidate object regions using CNNs, and classifiers are subsequently employed for the final detection. DAL can reduce labeled data to better fit numerous parameters of CNN. Wu et al. [149] propose a novel hybrid query strategy that jointly considers uncertainty and diversity. Extensive

experiments are conducted on two object detection datasets that effectively demonstrate the superiority and effectiveness of their model. Wang et al. [43] introduce active sample mining with switchable selection criteria to incrementally train robust object detectors using unlabeled or partially labeled samples, avoiding the influence of noisy samples and outliers. The effectiveness of the model is demonstrated through extensive experiments on publicly available object detection benchmarks. Yuan et al. [150] define an instance uncertainty learning module that takes advantage of the discrepancy of two adversarial instance classifiers trained in the labeled set to predict the instance uncertainty of the unlabeled set. With iterative instance uncertainty learning and reweighting, they suppress noisy instances, bridging the gap between instance and image-level uncertainty.

5) Pose Estimation: This aims to localize the positions of specific key points in images, which has a wide range of applications, such as augmented reality, translation of sign language, and human–robot interaction. Obtaining pose annotations can be extremely expensive and laborious. To address this issue, Caramalau et al. [151] propose distribution-based methods for the selection of diverse and representative samples. Experiments demonstrate their high efficiency and effectiveness for pose estimation. Similarly, Shukla et al. [152] use an uncertainty-based query strategy, annotate samples with the lowest confidence scores, and further improve the performance with fewer labeled samples. Gong et al. [86] design a novel meta-agent teaming DAL (MATAL) framework to actively select and label informative images for effective learning. MATAL formulates the sample selection procedure as a Markov decision process and learns an optimal sampling policy that effectively maximizes the performance of the pose estimator.

6) Target Tracking: This aims to accurately track targets in images, which can be applied for numerous applications, including video surveillance, autonomous vehicles, and so on. Using DAL can better help train neural networks with limited labeled samples for target tracking. Yuan et al. [153] present a new DAL sequence selection method in a multiframe collaboration way for target tracking. To ensure the diversity of selected sequences, they measure samples' similarity by their temporal relation between multiple frames in each video, and they use the nearest neighbor discriminator to select the representative samples. Experiments show that their method can eliminate background noise and improve efficiency.

7) Person Reidentification: Person reidentification (Re-ID) aims to match a specific pedestrian using different cameras, which is an essential task for public security. Previous efforts mainly concentrate on enhancing the performance of Re-ID models, relying on large labeled datasets. However, these efforts often overlook data redundancy issues that can arise in constructing Re-ID datasets. To address data redundancy in Re-ID datasets, Liu et al. [46] propose an alternative human-in-the-loop model based on reinforce learning. In their method, a human annotator provides binary feedback to fine-tune a pretrained CNNs Re-ID model. Extensive experiments prove the superiority of their method compared to existing unsupervised, transfer learning, and DAL models. On the other hand, Xu et al. [155] focus on learning from scratch with incremental labeling through human annotators and model feedback. They combine DAL with an incremental annotation process to select informative and diverse samples without redundancy from an unlabeled set in each iteration. These samples are then labeled by human annotators to further improve the performance of the model.

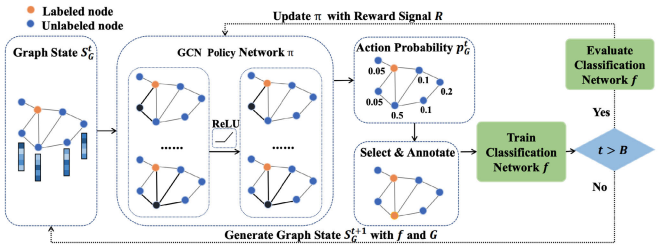


Fig. 10. Framework of graph policy network [120].

C. Applications in Graph DM and Learning

There is a substantial increase in content-rich networks from various domains, such as social networks, citation networks, and financial networks. Graphs have emerged as a powerful tool for representing and discovering knowledge, with nodes representing instances characterized by rich content features and edges denoting relationships or interactions between nodes.

1) Node Classification: This is to predict the labels of unlabeled nodes in a partially labeled network. GNNs rely heavily on a sufficient number of labeled nodes, which is costly and time-consuming. To address this problem, many graph-based DAL methods are proposed. For example, ICA-based methods [165] leverage label dependence among neighboring nodes to select diverse samples for node classification, while AGE [166] and ANRMAB [167] integrate GCNs with three traditional DAL query strategies and achieve good performance on many node classification datasets. As shown in Fig. 10, Hu et al. [120] present a graph policy network for transferable DAL on graphs, which formalizes DAL on graphs as a Markov decision process and learns the optimal query strategy with reinforce learning. The state is defined based on the current graph status, and the action is to select a node for annotation at each query step. The reward is defined as the performance gain of the GNNs trained with the selected nodes.

2) Link Prediction: This aims to predict missing or potential links between nodes in a given network. It involves using existing connections or relationships to infer the likelihood of forming new connections. In the context of link prediction, the challenge arises from the limited availability of existing link information between nodes in a network. DAL can help alleviate this issue; for example, DALAUP [168] uses neural networks to obtain vector representations of user pairs and utilizes multiple query strategies to select informative user pairs for labeling and model training, achieving superior performance compared to existing methods. Cai et al. [156] design a multiview DAL method that reduces the annotation cost by selectively querying metadata for the most informative examples, using a mapping function from the visual view to the text view. They demonstrate that multiview DAL can use richer information to help improve performance than using a single view. Zhao et al. [157] propose a DAL-based transfer learning framework for link prediction in recommender systems, which iteratively selects entities from source systems for target systems using uncertainty-based criteria. Experiments show that their method successfully improves efficiency and effectiveness.

3) Community Detection: This aims to accurately partition nodes into distinct classes based on the topological structure of the networks. However, in many practical scenarios, unsupervised methods struggle to achieve the exact community. To solve this issue, Gupta et al. [158] propose community trolling, a DAL-based method for topic-based community detection. Their method selects relevant samples from polluted big data, reducing the unreliable dataset to a reliable one for studying communities. Chien et al. [159] propose a novel DAL method for geometric community detection. They first remove many cross-cluster edges while preserving intracluster connectivity to avoid noise. Then, they interactively query the label of one node for each disjoint component to recover the underlying clusters. Experiments show that they can achieve SOTA performance in community detection.

D. Other Selected Interesting Applications

1) Engineering Systems: DAL methods exhibit remarkable performance in computationally demanding engineering systems by significantly reducing running time and computational costs. For example, Yue et al. [169] introduce two novel DAL algorithms: the variance-based weighted AL and the D-optimal weighted AL, designed specifically for Gaussian processes with uncertainties. Numerical studies demonstrate the effectiveness of their approach, notably improving predictive modeling for automatic shape control of composite fuselage structures. In another vein, Lee et al. [170] optimize their DAL acquisition function by jointly considering safe variance reduction and safe region expansion tasks, aiming to minimize failures without explicit knowledge of failure regions. This approach is tailored for real systems with uncertain failure conditions, as demonstrated in the predictive modeling of composite fuselage deformation, achieving zero failures by considering the composite failure criterion. Furthermore, Lee et al. [171] introduce a partitioned DAL method, comprising two systematic steps: global searching for uncertain design spaces and local searching using local Gaussian processes. They apply their method to aerospace manufacturing and materials science, achieving superior performance in prediction accuracy and computational efficiency compared to benchmarks.

2) Personalized Medical Treatment: This explores how patient health is affected by taking a drug and how user questions are answered by search recommendation [172]. Although modern methods can achieve impressive performance, they need a significant amount of labeled data. To solve this issue, Deng et al. [160] propose the use of DAL to recruit patients and assign treatments that reduce the uncertainty of an individual treatment effect model. Sundin et al. [173] propose to use a Gaussian process to model the individual treatment effect and use the expected information gain over the S-type error rate, defined as the error in predicting the sign of the conditional average treatment effect, as their acquisition function. Jesson et al. [113] develop epistemic uncertainty-aware methods for DAL of personalized treatment effects from high-dimensional observational data. In contrast to previous work that only uses information gain as the acquisition objective, they propose Causal-BALD because they consider both information gain and overlap between the treatment and control groups. Li et al. [174] used DAL to help people by recognizing their emotions.

VI. Challenges and Opportunities of DAL

As shown in Table IV, hereafter, we summarize the challenges and the corresponding potential solutions and opportunities.

A. Pipeline-Related Issues

1) Inefficient and Costly Human Annotation: DAL assumes that human annotators are readily available to label new samples once they are required. However, this assumption may not hold in some real-world applications. Human annotators can get tired or need breaks, forcing the DAL process to be suspended until they reappear. Moreover, human annotation is time-consuming and needs expert knowledge, resulting in long waits before models can be retrained with newly labeled data.

To improve efficiency, DAL methods incorporate additional techniques to reduce human annotation. Wang et al. [36] use self-supervised learning by adding pseudolabels with high confidence to help reduce human effort and improve the performance of the model. Going one step further, Yang and Loog [85] introduce multiple pseudoannotators that provide labels for unlabeled samples, achieving good performance without requiring human expert knowledge. On the other hand, as shown in Fig. 11, Huang et al. [134] propose a new annotation strategy to allow servers, workers, and annotators to cooperate efficiently for sharing candidate queries and annotations. Experiments show that their model can avoid annotation noise and save much time for rechecking annotations. To further reduce expert knowledge, others tend to reduce the search scope in each iteration to improve efficiency. For example, Yang and Loog [94] restrict candidate samples to their nearest neighbors of the labeled set rather than scanning all data.

2) Insufficient Research on Stopping Strategies: Few studies are designed for stopping strategies of DAL methods [196]. However, stopping strategies are essential for DAL because they reduce the amount of human labor by limiting the number of samples that need to be labeled and prevent the inclusion of noisy and redundant samples, which can negatively affect the performance of DAL models.

McDonald et al. [175] design two novel stopping strategies for DAL methods in the document classification task. The first strategy measures the overall confidence of the classifiers in correctly classifying the remaining unlabeled documents. It assumes that when the classifier's mean confidence level for the remaining documents stabilizes, the model stops the DAL process since its effectiveness will no longer improve. The second strategy measures the confidence of the classifiers among the selected documents to be reviewed. It assumes that when the classifier's confidence stops increasing for these documents, it has reached its maximal confidence and stops the DAL process. Benefiting from the idea of the margin exhaustion criterion, Yu et al. [176] identify two corresponding contour lines in the instance space and assume that the DAL process can only be stopped when all instances lying between these two contour lines have been labeled. They achieve good performance in many classification tasks. Based on the Bayesian theory, Ishibashi and Hino [177] derive a novel upper bound for the difference in expected generalization errors before and after obtaining new training data. They then combine this upper bound with a statistical test to derive a stopping criterion for DAL and significantly improve efficiency.

TABLE IV
SUMMARY OF VARIOUS CHALLENGES AND OPPORTUNITIES

Challenge Types	Challenges	Opportunities
Pipeline-related Issues	Inefficient & Costly human annotation	servers, workers and annotators share information [134]. self-supervised pseudo-labels to reduce human efforts [36], [85]. incorporate additional knowledge to reduce expert knowledge [94], [148].
	Insufficient research on stopping strategies	the confidence among the selected samples does not increase [175]. stop when all instances lie between two contour lines [176]. upper bound in expected generalization errors as stopping criterion [177].
	Cold-start	use pre-trained embeddings [31], [178]. design initial queries [179], [180]. use diverse sampling [181], [182].
Tasks-related Issues	Difficulty in cross-domain transfer	select samples in regions of joint disagreement between models [84], [154], [183]. source and target domain distribution matching [110], [184]. transferable DAL policies between the source and target graphs [120].
	Unstable performance	avoid DAL's sensitivity to the initial labeled set [31], [53], [94], [176], [185]. use distribution information to improve model's robustness [186], [187], [187]. use pre-trained language model [61], [188].
	Lack of scalability & generalizability	hybrid strategies for sample selection [60], [104]. nearest-neighbor classifiers [189]. combining annotation and counterfactual sample construction [190], [191].
Datasets-related Issues	Outlier Data & Noisy Oracles	find the best balance between purity and informativeness [89], [126]. knowledge distillation [14]. relabeling frameworks for correst oracle labels [81], [192], [193].
	Data Scarcity & Imbalance	data augmentation and large PLMs [12], [32], [97]. cost-sensitive learning [176], [194]. design new query strategies for imbalanced datasets [195]–[197].
	Class distribution mismatch	new DAL query strategy [125], [198]. new DAL framework [199]. incoporate additional detector [200].

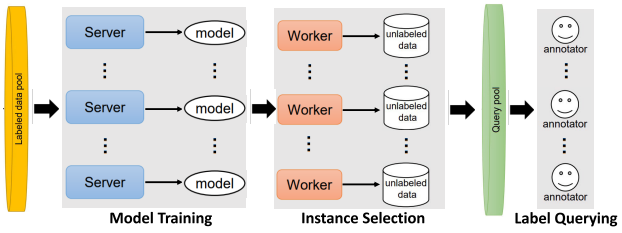


Fig. 11. Framework for efficient annotation.

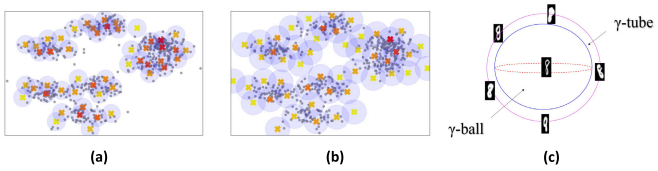


Fig. 12. Example for cold-start data selection. (a) ProbCover selection. (b) CoreSet selection. (c) γ-ball selection.

3) Cold Start: Most DAL methods fail to improve over random selection when the annotation budget is very small, a phenomenon sometimes termed "cold start" [179]. Uncertainty sampling has been shown to be inherently unsuitable for low budgets, possibly explaining the cold-start phenomenon [201]. Low budgets can be seen in many applications, especially those that require an expert tagger whose time is expensive. If we want to expand deep learning to new domains, overcoming the cold-start problem is an ever-important task.

To relieve the cold-start issue, Yuan et al. [31] use pretrained embeddings on unsupervised tasks, decreasing budget dependence while remaining faithful to uncertainty sampling. Similarly, Yu et al. [178] try to use pretrained knowledge from PLMs to avoid cold start. They select few shot samples to fine-tune large-scale PLM, achieve SOTA performance in six datasets, and improve the efficiency of labeling over existing baselines by 3.2%–6.9% on average. On the other hand, as shown in Fig. 12(a) and (b), Yehuda and Dekel [180] develop a new DAL initialization strategy to solve the cold-start issue for low-budget image classification, which significantly outperforms CoreSet initialization in the low-budget regime. They also theoretically analyze different DAL strategies in embedding spaces and improve performance on both low- and high-budget scenes. As shown in Fig. 12(c), Cao et al. [181] apply the informative sampling policy on the γ tube to solve the cold-start sampling problem. Mahmood et al. [182] query a diverse set of examples with minimal Wasserstein distance from unlabeled data. They report a significant performance boost in the low-budget regime.

B. Task-Related Issues

1) Difficulty in Cross-Domain Transfer: We discuss two difficulties of cross-domain transfer in DAL. First, machine learning systems are always deployed on various devices with the same labeled dataset. However, DAL is often model-dependent and not directly transferable, i.e., data queried for one model may be less effective for another [183]; Second, transfer learning biases DAL to select samples that match the distribution of the source domain to the target domain, leading to sampling bias and the high cost of transfer learning.

To benefit multiple target models, some methods aim to select samples in joint disagreement regions across models [183], adopt multiagent reinforcement learning for optimal

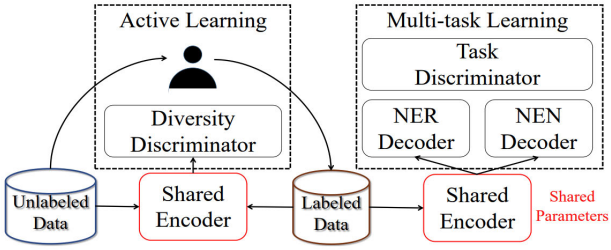


Fig. 13. MTL transfer knowledge from sources [84].

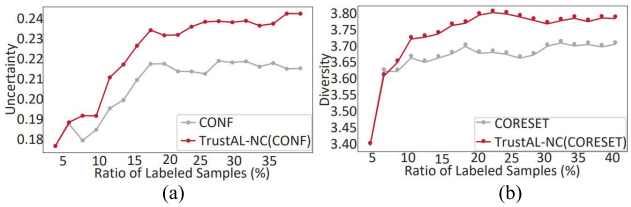


Fig. 14. Stable performance of TrustAL [186]. (a) CONF. (b) CoreSet.

selection [154], or leverage MTL to transfer common knowledge from the source domain, as shown in Fig. 13. To avoid sampling bias, Farquhar et al. [184] apply corrective weighting using an unbiased risk estimator to maintain the target distribution during pool-based sampling. Trang et al. [110] introduce a heuristic query strategy that matches the distribution of the source domain while retrieving valuable target samples. Hu et al. [120] learn transferable DAL policies on labeled source graphs that generalize selection to unlabeled target graphs. Experiments show that the above methods can achieve excellent performance and transferability.

2) Unstable Performance: DAL methods always have unstable performance, i.e., results for the same method vary significantly with different initialized seeds [108]. Two primary reasons can explain this instability. First, the DAL methods are sensitive to the initial labeled dataset. The initial selected samples have a great influence on the eventual outcome of the current approaches. With insufficient initial labeling, subsequent DAL cycles become highly biased, resulting in poor selection. Second, current DAL methods always separate active learning and deep learning methods into two separate processes, easily leading to suboptimal and unstable performance [202].

To solve DAL's sensitivity to the initialization, current methods always use diverse sampling and pretrained models. Yu et al. [176] adopt hierarchical clustering to select 10% samples near each clustering center as representative samples. Their new initialization greatly helps stabilize the performance. Zlabinger [185] takes into account both diversity and polarization to effectively select initial samples for DAL methods that further stabilize the performance of the DAL process. Yang and Loog [94] select initial samples by evaluating the total distance between the unlabeled samples and the initial samples, showing that the same distance between them can result in better and stable performance. On the other hand, Yuan et al. [31] incorporate language information as prior knowledge to help learn node representations and use clustering methods to select the initial data. Similarly, Dor et al. [53] use BERT to learn the representations of the input sentences and use a hybrid query strategy to select the most uncertain and diverse samples as the initialized training data.

To bridge the gap between AL and deep learning models, Kwak et al. [186] introduce trustworthy AL (TrustAL), a label-efficient DAL framework by transferring distilled knowledge from deep learning models to the data selection process. As shown in Fig. 14, they jointly optimize knowledge distillation and DAL to obtain a more consistent and reliable performance compared to the two best performing baselines on three benchmarks. Similarly, Ma et al. [187] learn nonlinear embeddings to map inputs into a latent space and introduce a selection block to choose representative samples in the learned latent space to achieve stable performance. Schröder et al. [61] extend the PLMs to continually pretrain on available unlabeled data to tailor it to the task-specific domain, where they can benefit from both labeled and unlabeled data at each DAL iteration. Their experiments show considerable enhancements in data efficiency and stability compared to the standard fine-tuning approach, emphasizing the importance of a suitable training strategy in DAL. Mamooler et al. [188] try to combine DAL with PLMs in the legal domain, where they use unlabeled data in three stages: training the model to adjust it to the downstream task, using knowledge distillation to direct the embeddings to a semantically meaningful space, and identifying the initial set.

3) Lack of Scalability and Generalizability: Current DAL methods lack scalability, as they always require significant modifications to neural network architectures to adapt to different query strategies. Another issue with current methods is their heavy reliance on DAL's weight parameters, while the parameters may not be generalizable to different datasets. Users are required to prepare additional labeled samples as a validation set to tune parameters by cross-validation, which contradicts the goal of minimizing the need for labeled data.

In response to the above issues, Maekawa et al. [60] introduce a novel DAL method, called TYROGUE, which uses a hybrid query strategy to improve model generalization and reduce labeling costs. As shown in Fig. 15, uncertainty-based methods tend to acquire similar data points from a specific area within an iteration, diversity-based methods tend to acquire data points similar to the samples acquired in previous iterations, and TYROGUE balances diversity and uncertainty by acquiring samples that are diverse and also closer to the model decision boundary. RMQCAL [104] is a novel scalable DAL method, which allows for any number and type of query criteria, eliminates the need for empirical parameters, and makes the tradeoffs between the query criteria self-adaptive. On the other hand, Wan et al. [189] propose an embedded network of nearest-neighbor classifiers to enhance the generalization ability of models trained in labeled and unlabeled subspaces in a simple but effective manner. Deng et al. [190] focus on combining sample annotation and counterfactual sample construction in the DAL procedure to enhance the model's out-of-distribution generalization. Wang et al. [191] introduce a new training manner to improve the model's generalizability and show a strong positive correlation between convergence speed and generalization performance under ultrawide conditions.

C. Dataset-Related Issues

1) Outlier Data and Noisy Oracles: DAL methods tend to acquire outliers since models always assign high uncertainty scores to outliers. Outliers can damage a model's learning ability and fuel a vicious cycle in which DAL methods continue to select them [43]. Identifying and removing outliers have

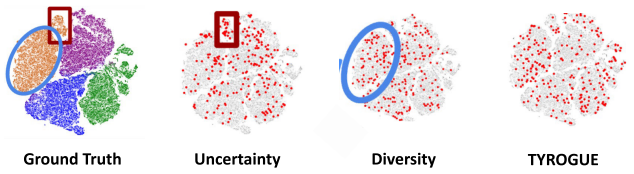


Fig. 15. TYROGUE can select better samples than baselines.

become important directions in improving DAL performance and robustness. On the other hand, classic DAL methods assume that annotators have high labeling accuracy. However, in real-world settings, sample difficulty and annotator expertise can significantly affect the quality and accuracy of annotation, which may further degrade model performance.

To remove outliers, Park et al. [126] propose MQ-Net to adaptively find the best balance between purity and informativeness of samples, filtering out noisy open-set data. Elenter et al. [89] introduce a new query strategy based on Lagrangian duality to select diverse samples, efficiently removing redundant data. Other studies [14] use knowledge distillation to compress useful knowledge into a small model, effectively identifying and removing outliers. To make high-quality annotations, AMCC [81] measures worker annotations considering both their commonality and individuality to reduce the impact of unreliable workers and improve effectiveness. Zhao et al. [192] actively select samples that are relabeled multiple times through crowd-sourcing majority voting. EMMA [193] relabels samples to remove noisy annotations by analyzing the stimulus based on model memory retention and greedy heuristics. BALT [203] improves human expertise during labeling to improve relabel quality and significantly improve model performance. Zlabinger [185] trains human annotators on a set of prelabeled samples to improve the quality of annotations. Huang et al. [134] propose a multiserver, multiworker framework for DAL, where servers and workers cooperate to select diverse samples and improve model performance.

2) Data Scarcity and Imbalance: Data scarcity poses two critical challenges. First, datasets are difficult to collect and annotate [204]; Second, DAL methods have the common underlying assumption that all classes are equal, while some classes have more samples than others (skewed class distribution [176]) or some classes may be more difficult to learn than others, leading to sampling bias in the acquisition process [205].

For scarce datasets, Chen et al. [12] used data augmentation to generate diverse samples to expand training data. Other studies used PLMs as prior knowledge and fine-tuned them to reduce the required labeled samples [32]. For difficult annotations, Gudovskiy et al. [97] introduce several novel self-supervised pseudolabels estimators to correct acquisition bias by minimizing the distribution shift between unlabeled data and weakly labeled validation data. To mitigate the class imbalance, Yu et al. [176] are the first to use cost-sensitive learning. They choose the extreme weighted learning machine as the base learner to select samples based on the class imbalance ratio, class overlap, and small disjunction. They investigate why DAL can be impacted by a skewed instance distribution and improve DAL performance on imbalanced datasets. Choi et al. [194] solve the issue of data imbalance by considering the probability of mislabeling a class, the probability of the data given a predicted class, and the prior probability

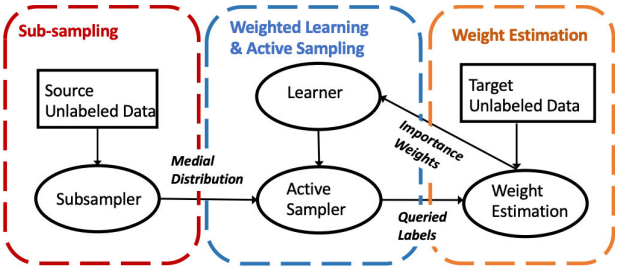


Fig. 16. Example of imbalanced sampling [195].

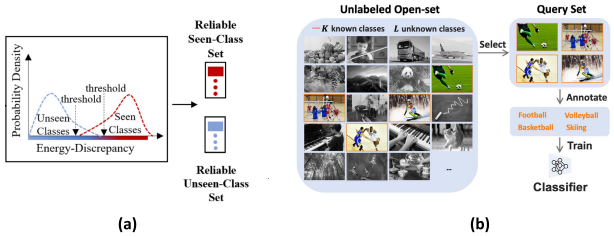


Fig. 17. Methods for solving class distribution mismatch. (a) Seen and unseen classes identification. (b) Find examples from known classes.

of the abundance of a predicted class, during querying samples of DAL. Experiments show that they can significantly enhance the ability of existing DAL methods to handle unbalanced datasets. As shown in Fig. 16, Zhao et al. [195] propose an alternate query strategy by using the medial distribution to find a compromise between importance weighting and class-balanced sampling. Experiments show that their model can be easily combined with various DAL methods and successfully select balanced samples in imbalanced datasets. Hartford et al. [196] present an exemplar-guided DAL method that shows strong empirical performance under extremely skewed label distributions by using exemplar embedding. Zhang et al. [197] propose a graph-based DAL method that applies a more sophisticated version of uncertainty sampling. Their strategy can select more evenly distributed examples for labeling than standard uncertainty sampling.

3) Class Distribution Mismatch: DAL methods assume that the labeled and unlabeled data are drawn from the same class distribution, which means that the categories of both datasets are identical [200]. However, in real-world scenarios, unlabeled data often come from uncontrolled sources, and a large portion of the examples may belong to unknown classes. For example, when crawling images for binary image classification using keywords such as "dog" and "cat," over 50% of the images in the unlabeled dataset are irrelevant to the task (e.g., "deer" and "horse"). Annotating these irrelevant images will lead to a waste of the annotation budget as they are unnecessary for training the desired classifier. Despite this challenge, existing DAL systems tend to select these irrelevant images for annotation, as they contain more uncertain knowledge.

To address this issue, as shown in Fig. 17(a), He et al. [198] propose the energy discrepancy to measure the density distribution between the seen and unseen classes. Then, they propose an iterative optimization strategy to facilitate the teacher–student distillation network to avoid selecting samples from unseen classes. Furthermore, Tang and Huang [199] propose a dual DAL framework that simultaneously performs model search and data selection. Their framework effectively addressed the issue of distribution mismatch and significantly improved model performance. As shown in Fig. 17(b),

Ning et al. [200] introduce a detector-classifier DAL framework, where the detector filters unknown classes using Gaussian mixture models and the classifier selects uncertain in-distribution samples for retraining. By actively acquiring purer in-distribution query sets, this framework improves the model generalization on class distribution mismatch.

VII. Conclusion

Due to the advantages of DAL, such as high efficiency, good effectiveness, and strong robustness, DAL has been deployed in both research and industry projects. This article provides a comprehensive survey of DAL, including its collection, definition, influential baselines and datasets, taxonomy, applications, challenges, and some inspiring prospects. First, we discuss the collection and filtering of DAL papers to ensure their high quality. Second, we give the definition of DAL tasks and present its basic pipeline, influential baselines, and widely used datasets. Third, we present our taxonomy for DAL methods from several perspectives and discuss their strengths and weaknesses. From them, we obtain some guidelines for selecting different query strategies, deep model architectures, and learning paradigms to apply to different tasks. In addition, different annotation strategies can significantly reduce manual labor while also bringing certain drawbacks. In terms of the training process, curriculum learning training and Pre + FT can better adapt to the current era of large language models. Fourth, we discuss some typical applications of DAL. Other than the commonly used and popular DAL methods used for CV tasks, we also introduce the carefully designed DAL method for NLP, DM, and so on. Finally, even though DAL has many benefits, we reckon that it can be refined further in terms of pipeline, tasks, and datasets. Specifically, there are many problems that DAL is hard to handle, such as inefficient human annotation, difficulty in cross-domain transfer, unstable performance, lack of scalability, data imbalance, and class distribution mismatch. We share DAL-related resources on GitHub. We hope that this work will be a quick guide for researchers and motivate them to solve important problems in the DAL domain.

References

[1] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.

[2] B. Gu, Z. Zhai, C. Deng, and H. Huang, "Efficient active learning by querying discriminative and representative samples and fully exploiting unlabeled data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 9, pp. 4111–4122, Sep. 2021.

[3] X. Cao and I. W. Tsang, "Shattering distribution for active learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 1, pp. 215–228, Jan. 2022.

[4] S. Liu et al., "Online active learning for drifting data streams," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 1, pp. 186–200, Jan. 2023.

[5] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. ICLR*, 2017. [Online]. Available: https://openreview.net/forum?id=SJU4ayYgl

[6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[7] A. Vaswani, "Attention is all you need," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, 2017, pp. 1–11.

[8] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[9] OpenAI et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[10] Y. Bengio, J. Louradour, and R. Collobert, "Curriculum learning," in *Proc. Int. Conf. Mach. Learn.*, Aug. 2009, pp. 41–48.

[11] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 5, pp. 1019–1034, May 2015.

[12] Z. Chen, J. Zhang, P. Wang, J. Chen, and J. Li, "When active learning meets implicit semantic data augmentation," in *Proc. IEEE Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 56–72.

[13] S. Yang et al., "Dataset pruning: Reducing training data by examining generalization influence," in *Proc. ICLR*, 2023. [Online]. Available: https://openreview.net/forum?id=4wZiAXD29TQ

[14] F. Peng, C. Wang, J. Liu, and Z. Yang, "Active learning for lane detection: A knowledge distillation approach," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15132–15141.

[15] Z. Zhang, E. Strubell, and E. Hovy, "A survey of active learning for natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 6166–6190.

[16] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari, "A survey of active learning algorithms for supervised remote sensing image classification," 2021, *arXiv:2104.07784*.

[17] H. Wang, Q. Jin, S. Li, S. Liu, M. Wang, and Z. Song, "A comprehensive survey on deep active learning in medical image analysis," 2023, *arXiv:2310.14230*.

[18] H. Hadian and H. Sameti, "Active learning in noisy conditions for spoken language understanding," in *Proc. ACL*, 2014, pp. 1081–1090.

[19] S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," *Med. Image Anal.*, vol. 71, Jul. 2021, Art. no. 102062.

[20] R. Takezoe et al., "Deep active learning for computer vision: Past and future," 2022, *arXiv:2211.14819*.

[21] X. Zhan, H. Liu, Q. Li, and A. B. Chan, "A comparative survey: Benchmarking for pool-based active learning," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 4679–4686.

[22] X. Zhan, Q. Wang, K.-H. Huang, H. Xiong, D. Dou, and A. B. Chan, "A comparative survey of deep active learning," 2022, *arXiv:2203.13450*.

[23] Y. Fu, X. Zhu, and B. Li, "A survey on instance selection for active learning," *Knowl. Inf. Syst.*, vol. 35, no. 2, pp. 249–283, May 2013.

[24] C. Aggarwal et al., "Active learning: A survey," in *Data Classification*. New York, NY, USA: Chapman & Hall, 2014. [Online]. Available: https://charuaggarwal.net/active-survey.pdf

[25] A. Tharwat and W. Schenck, "A survey on active learning: State-of-the-art, practical challenges and research directions," *Mathematics*, vol. 11, no. 4, p. 820, Feb. 2023. [Online]. Available: http://www.mdpi.com/2227-7390/11/4/820

[26] P. Ren et al., "A survey of deep active learning," *ACM Comput. Surv.*, vol. 54, no. 9, pp. 1–40, 2022.

[27] P. Liu, L. Wang, R. Ranjan, G. He, and L. Zhao, "A survey on active deep learning: From model driven to data driven," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–34, Jan. 2022.

[28] D. Cacciarelli and M. Kulahci, "Active learning for data streams: A survey," *Mach. Learn.*, vol. 113, no. 1, pp. 185–239, Jan. 2024.

[29] K. Margatina, L. Barrault, and N. Aletras, "On the importance of effectively adapting pretrained language models for active learning," in *Proc. ACL*, 2022, pp. 825–836.

[30] A. Ayub and C. Fendley, "Few-shot continual active learning by a robot," in *Proc. NeurIPS*, vol. 35, 2022, pp. 30612–30624.

[31] M. Yuan, H.-T. Lin, and J. Boyd-Graber, "Cold-start active learning through self-supervised language modeling," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 7935–7948.

[32] S. Seo, D. Kim, Y. Ahn, and K. Lee, "Active learning on pre-trained language model with task-independent triplet loss," in *Proc. AAAI*, 2022, pp. 11276–11284.

[33] K. Margatina, T. Schick, N. Aletras, and J. Dwivedi-Yu, "Active learning principles for in-context learning with large language models," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2023, pp. 5011–5034.

[34] Y. Gal and Z. Ghahramani, "Bayesian convolutional neural networks with Bernoulli approximate variational inference," in *Proc. ICLR*, 2016.

[35] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian active learning with image data," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 1183–1192.

[36] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2591–2600, Dec. 2017.

[37] L. Balaji and A. Pritzel, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. NeurIPS*, 2017.

[38] M. Fang, Y. Li, and T. Cohn, "Learning how to active learn: A deep reinforcement learning approach," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 595–605.

[39] K. Konyushkova, R. Sznitman, and P. Fua, "Learning active learning from data," in *Proc. NeurIPS*, 2017.

[40] J.-J. Zhu and J. Bento, "Generative adversarial active learning," 2017, *arXiv:1702.07956*.

[41] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *Proc. ICLR*, 2018. [Online]. Available: https://openreview.net/forum?id=H1aIuk-RW

[42] M. Ducoffe and F. Precioso, "Adversarial active learning for deep networks: A margin based approach," 2018, *arXiv:1802.09841*.

[43] K. Wang, L. Lin, X. Yan, Z. Chen, D. Zhang, and L. Zhang, "Cost-effective object detection: Active sample mining with switchable selection criteria," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 3, pp. 834–850, Mar. 2019.

[44] M.-A. Carbonneau, E. Granger, and G. Gagnon, "Bag-level aggregation for multiple-instance active learning in instance classification problems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1441–1451, May 2019.

[45] A. Kirsch, J. Amersfoort, and Y. Gal, "BatchBALD: Efficient and diverse batch acquisition for deep Bayesian active learning," in *Proc. NeurIPS*, 2019.

[46] Z. Liu, J. Wang, S. Gong, D. Tao, and H. Lu, "Deep reinforcement active learning for human-in-the-loop person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6121–6130.

[47] J. Kasai, K. Qian, S. Gurajada, Y. Li, and L. Popa, "Low-resource deep entity resolution with transfer and active learning," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 5851–5861.

[48] T. Tran, T. T. Do, and I. Reid, "Bayesian generative active deep learning," in *Proc. ICML*, 2019, pp. 6295–6304.

[49] S. Sinha, S. Ebrahimi, and T. Darrell, "Variational adversarial active learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 5972–5981.

[50] J.-C. Su, Y.-H. Tsai, K. Sohn, B. Liu, S. Maji, and M. Chandraker, "Active adversarial domain adaptation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 739–748.

[51] M. Gao, Z. Zhang, G. Yu, S. Ö. Arık, L. S. Davis, and T. Pfister, "Consistency-based semi-supervised active learning: Towards minimizing labeling cost," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 510–526.

[52] B. Zhang, L. Li, S. Yang, S. Wang, Z.-J. Zha, and Q. Huang, "State-relabeling adversarial active learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 8756–8765.

[53] L. Ein-Dor et al., "Active learning for BERT: An empirical study," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 7949–7962.

[54] S. Huang, T. Wang, H. Xiong, J. Huan, and D. Dou, "Semi-supervised active learning with temporal output discrepancy," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3447–3456.

[55] G. Citovsky et al., "Batch active learning at scale," in *Proc. NeurIPS*, 2021, pp. 11933–11944.

[56] Y. Kim, K. Song, J. Jang, and I. Moon, "LADA: Look-ahead data acquisition via augmentation for deep active learning," in *Proc. NeurIPS*, 2021, pp. 22919–22930.

[57] K. Kim, D. Park, K. I. Kim, and S. Y. Chun, "Task-aware variational adversarial active learning," in *Proc. CVPR*, 2021, pp. 8166–8175.

[58] S. Karamcheti, R. Krishna, L. Fei-Fei, and C. Manning, "Mind your outliers! Investigating the negative impact of outliers on active learning for visual question answering," in *Proc. ACL*, 2021, pp. 7265–7281.

[59] Z. L. Zhu, V. Yadav, Z. Afzal, and G. Tsatsaronis, "Few-shot initializing of active learner via meta-learning," in *Proc. EMNLP*, 2022, pp. 1117–1133.

[60] S. Maekawa, D. Zhang, H. Kim, S. Rahman, and E. Hruschka, "Low-resource interactive active labeling for fine-tuning language models," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2022, pp. 3230–3242.

[61] C. Schröder, A. Niekler, and M. Potthast, "Revisiting uncertainty-based query strategies for active learning with transformers," in *Proc. Findings Assoc. Comput. Linguistics: ACL*, 2022, pp. 2194–2203.

[62] P. Menard, O. D. Domingues, A. Jonsson, E. Kaufmann, E. Leurent, and M. Valko, "Fast active learning for pure exploration in reinforcement learning," in *Proc. ICML*, 2021, pp. 7599–7608.

[63] A. Krizhevsky. (2009). *Learning Multiple Layers of Features From Tiny Images*. [Online]. Available: https://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf

[64] Y. Netzer, T. Wang, and A. Coates. (2011). *Reading Digits in Natural Images With Unsupervised Feature Learning*. [Online]. Available: http://ufldl.stanford.edu/housenumbers/

[65] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[66] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[67] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. CVPR*, 2016, pp. 3213–3223.

[68] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop*, 2004, p. 178.

[69] R. Socher et al., "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. EMNLP*, 2013, pp. 1631–1642.

[70] X. Li and D. Roth, "Learning question classifiers," in *Proc. COLING*, 2002, pp. 1–7.

[71] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proc. EMNLP*, 2015, pp. 632–642.

[72] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proc. ACL*, 2011, pp. 142–150.

[73] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proc. NeurIPS*, 2015.

[74] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, "Collective classification in network data," *AI Mag.*, vol. 29, no. 3, p. 93, Sep. 2008.

[75] S. Abu-El-Haija et al., "YouTube-8M: A large-scale video classification benchmark," 2016, *arXiv:1609.08675*.

[76] A. E. Johnson et al., "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, pp. 1–9, 2016.

[77] M. Wiechmann, S. M. Yimam, and C. Biemann, "ActiveAnno: General-purpose document-level annotation tool with active learning integration," in *Proc. NAACL*, 2021, pp. 99–105.

[78] T.-H. Wu et al., "ReDAL: Region-based and diversity-aware active learning for point cloud semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15490–15499.

[79] S. Kothawade et al., "Targeted active learning for object detection with rare classes and slices using submodular mutual information," in *Proc. ECCV*, 2022, pp. 1–16.

[80] B. Xie, L. Yuan, S. Li, C. H. Liu, and X. Cheng, "Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8058–8068.

[81] G. Yu, J. Tu, J. Wang, C. Domeniconi, and X. Zhang, "Active multilabel crowd consensus," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 4, pp. 1448–1459, Apr. 2021.

[82] C. Schröder et al., "Supporting land reuse of former open pit mining sites using text classification and active learning," in *Proc. ACL*, 2021, pp. 4141–4152.

[83] Y. Yan, S.-J. Huang, S. Chen, M. Liao, and J. Xu, "Active learning with query generation for cost-effective text classification," in *Proc. AAAI*, 2020, pp. 6583–6590.

[84] B. Zhou et al., "MTAAL: Multi-task adversarial active learning for medical named entity recognition and normalization," in *Proc. AAAI*, 2021, pp. 14586–14593.

[85] Y. Yang and M. Loog, "Single shot active learning using pseudo annotators," *Pattern Recognit.*, vol. 89, pp. 22–31, May 2019.

[86] J. Gong, Z. Fan, Q. Ke, H. Rahmani, and J. Liu, "Meta agent teaming active learning for pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11069–11079.

[87] B. Xie, L. Yuan, S. Li, C. H. Liu, X. Cheng, and G. Wang, "Active learning for domain adaptation: An energy-based approach," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 8708–8716.

[88] D. Wang and Y. Shang, "A new active labeling method for deep learning," in *Proc. IJCNN*, 2014, pp. 112–119.

[89] J. Elenter, N. Naderializadeh, and A. Ribeiro, "A Lagrangian duality approach to active learning," in *Proc. NeurIPS*, 2022, pp. 37575–37589.

[90] W. Li, G. Dasarathy, K. N. Ramamurthy, and V. Berisha, "Finding the homology of decision boundaries with active learning," in *Proc. NeurIPS*, 2020, pp. 8355–8365.

[91] D. Roth and K. Small, "Margin-based active learning for structured output spaces," in *Proc. Eur. Conf. Mach. Learn.*, 2006, pp. 413–424.

[92] K. Brantley, H. Daumé III, and A. Sharaf, "Active imitation learning with noisy guidance," in *Proc. ACL*, 2020, pp. 2093–2105.

[93] S. Yan, K. Chaudhuri, and T. Javidi, "The label complexity of active learning from observational data," in *Proc. NeurIPS*, 2019, pp. 1808–1817.

[94] Y. Yang and M. Loog, "To actively initialize active learning," *Pattern Recognit.*, vol. 131, Nov. 2022, Art. no. 108836.

[95] Y. Kim and B. Shin, "In defense of core-set: A density-aware core-set selection for active learning," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2022, pp. 804–812.

[96] C. Coleman et al., "Similarity search for efficient active learning and search of rare concepts," in *Proc. AAAI*, 2022, pp. 6402–6410.

[97] D. Gudovskiy, A. Hodgkinson, T. Yamaguchi, and S. Tsukizawa, "Deep active learning for biased datasets via Fisher kernel self-supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9038–9046.

[98] M. Hasan, S. Paul, A. I. Mourikis, and A. K. Roy-Chowdhury, "Context-aware query selection for active learning in event recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 554–567, Mar. 2020.

[99] S. Chakraborty, V. Balasubramanian, Q. Sun, S. Panchanathan, and J. Ye, "Active batch selection via convex relaxations with guaranteed solution bounds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 1945–1958, Oct. 2015.

[100] Q. Jin, M. Yuan, Q. Qiao, and Z. Song, "One-shot active learning for image segmentation via contrastive learning and diversity-based sampling," *Knowl.-Based Syst.*, vol. 241, Apr. 2022, Art. no. 108278.

[101] C. Li, H. Ma, Z. Kang, Y. Yuan, X.-Y. Zhang, and G. Wang, "On deep unsupervised active learning," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 2626–2632.

[102] A. Parvaneh, E. Abbasnejad, D. Teney, R. Haffari, A. Van Den Hengel, and J. Q. Shi, "Active learning by feature mixing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12227–12236.

[103] S. Li, J. M. Phillips, X. Yu, R. M. Kirby, and S. Zhe, "Batch multi-fidelity active learning with budget constraints," in *Proc. NeurIPS*, 2022, pp. 995–1007.

[104] Y. Zhao, Z. Shi, J. Zhang, D. Chen, and L. Gu, "A novel active learning framework for classification: Using weighted rank aggregation to achieve multiple query criteria," *Pattern Recognit.*, vol. 93, pp. 581–602, Sep. 2019.

[105] T. Wang et al., "Boosting active learning via improving test performance," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 8, pp. 8566–8574.

[106] M. Thiessen and T. Gärtner, "Active learning of convex halfspaces on graphs," in *Proc. NeurIPS*, 2021, pp. 23413–23425.

[107] M. A. Mohamadi, W. Bae, and D. J. Sutherland, "Making look-ahead active learning strategies feasible with neural tangent kernels," in *Proc. NeurIPS*, vol. 35, 2022, pp. 12542–12553.

[108] D. Yoo and I. S. Kweon, "Learning loss for active learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 93–102.

[109] G. Zhao et al., "Efficient active learning for Gaussian process classification by error reduction," in *Proc. NeurIPS*, 2021, pp. 9734–9746.

[110] T.-T. Vu, M. Liu, D. Phung, and G. Haffari, "Learning how to active learn by dreaming," in *Proc. ACL*, 2019, pp. 4091–4101.

[111] L. Wertz, J. Bogojeska, K. Mirylenka, and J. Kuhn, "Reinforced active learning for low-resource, domain-specific, multi-label text classification," in *Proc. Findings ACL*, 2023, pp. 10959–10977.

[112] W. H. Beluch, T. Genewein, A. Nurnberger, and J. M. Kohler, "The power of ensembles for active learning in image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9368–9377.

[113] A. Jesson et al., "Causal-bald: Deep Bayesian active learning of outcomes to infer treatment-effects from observational data," in *Proc. NeurIPS*, 2021, pp. 30465–30478.

[114] P. Donmez, J. Carbonell, and P. N. Bennett, "Dual strategy active learning," in *Proc. ECML*, 2007, pp. 116–127.

[115] Y. Geifman and R. El-Yaniv, "Deep active learning with a neural architecture search," in *Proc. NeurIPS*, 2019, pp. 5974–5984.

[116] P. Izmailov, S. Vikram, M. D. Hoffman, and A. G. Wilson, "What are Bayesian neural network posteriors really like?" in *Proc. ICML*, 2021, pp. 4629–4640.

[117] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

[118] X. Zeng, S. Garg, R. Chatterjee, U. Nallasamy, and M. Paulik, "Empirical evaluation of active learning techniques for neural MT," in *Proc. 2nd Workshop Deep Learn. Approaches Low-Resource NLP*, 2019, pp. 84–93.

[119] A. Amirinezhad, S. Salehkaleybar, and M. Hashemi, "Active learning of causal structures with deep reinforcement learning," *Neural Netw.*, vol. 154, pp. 22–30, Oct. 2022.

[120] S. Hu et al., "Graph policy network for transferable active learning on graphs," in *Proc. NeurIPS*, 2020.

[121] Y. Ren, B. Wang, J. Zhang, and Y. Chang, "Adversarial active learning based heterogeneous graph neural network for fake news detection," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2020, pp. 452–461.

[122] Y. Li, J. Yin, and L. Chen, "SEAL: Semisupervised adversarial active learning on attributed graphs," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 3136–3147, Jul. 2021.

[123] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. ICLR*, 2014. [Online]. Available: https://openreview.net/forum?id=33X9fd2-9FyZd

[124] S. Albelwi, "Survey on self-supervised learning: Auxiliary pretext tasks and contrastive learning methods in imaging," *Entropy*, vol. 24, no. 4, p. 551, Apr. 2022.

[125] P. Du, H. Chen, S. Zhao, S. Chai, H. Chen, and C. Li, "Contrastive active learning under class distribution mismatch," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4260–4273, Apr. 2023.

[126] D. Park, Y. Shin, J. Bang, Y. Lee, H. Song, and J. Lee, "Meta-query-Net: Resolving purity-informativeness dilemma in open-set active learning," in *Proc. NeurIPS*, 2022, pp. 31416–31429.

[127] J. Shao, Q. Wang, and F. Liu, "Learning to sample: An active learning framework," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 538–547.

[128] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann, "Self-paced curriculum learning," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2694–2700.

[129] L. Lin, K. Wang, D. Meng, W. Zuo, and L. Zhang, "Active self-paced learning for cost-effective and progressive face identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 7–19, Jan. 2018.

[130] M. Mundt, Y. Hong, I. Pliushch, and V. Ramesh, "Forgotten lessons and the bridge to active and open world learning," *Neural Netw.*, vol. 160, pp. 306–336, Jan. 2023.

[131] C. Löffler and C. Mutschler, "IALE: Imitating active learner ensembles," *J. Mach. Learn. Res.*, vol. 23, pp. 1–29, Jul. 2022.

[132] K. Bullard, Y. Schroecker, and S. Chernova, "Active learning within constrained environments through imitation of an expert questioner," in *Proc. IJCAI*, 2019, pp. 2045–2052.

[133] F. Ikhwantri et al., "Multi-task active learning for neural semantic role labeling on low resource conversational corpus," in *Proc. Workshop Deep Learn. Approaches Low-Resource NLP*, 2018, pp. 43–50.

[134] S.-J. Huang, C.-C. Zong, K.-P. Ning, and H.-B. Ye, "Asynchronous active learning with distributed label querying," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 2570–2576.

[135] Y.-P. Tang and S.-J. Huang, "Self-paced active learning: Query the right thing at the right time," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 5117–5124.

[136] W. Tan, L. Du, and W. L. Buntine, "Diversity enhanced active learning with strictly proper scoring rules," in *Proc. NeurIPS*, 2021, pp. 10906–10918.

[137] F. Jelenić, J. Jukić, N. Drobac, and J. Snajder, "On dataset transferability in active learning for transformers," in *Proc. Findings Assoc. Comput. Linguistics, ACL*, 2023, pp. 2282–2295.

[138] A. Gidiotis and G. Tsoumakas, "Should we trust this summary? Bayesian abstractive summarization to the rescue," in *Proc. Findings Assoc. Comput. Linguistics, ACL*, 2022, pp. 4119–4131.

[139] A. Tsvigun et al., "Active learning for abstractive text summarization," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2022, pp. 5128–5152.

[140] A. Padmakumar and R. J. Mooney, "Dialog policy learning for joint clarification and active learning queries," in *Proc. AAAI*, 2021, pp. 13604–13612.

[141] P. Radmard, Y. Fathullah, and A. Lipani, "Subsequence based deep active learning for named entity recognition," in *Proc. ACL*, 2021, pp. 4310–4321.

[142] X. Hua and L. Wang, "Efficient argument structure extraction with transfer learning and active learning," in *Proc. Findings Assoc. Comput. Linguistics, ACL*, 2022, pp. 423–437.

[143] L. Duong, H. Afshar, D. Estival, G. Pink, P. R. Cohen, and M. Johnson, "Active learning for deep semantic parsing," in *Proc. ACL*, 2018, pp. 43–48.

[144] Z. Li, L. Qu, P. R. Cohen, R. Tumuluri, and G. Haffari, "The best of both worlds: Combining human and machine translations for multilingual semantic parsing with active learning," in *Proc. ACL*, 2023, pp. 9511–9528.

[145] B. Zhang et al., "Structural semantic adversarial active learning for image captioning," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1112–1121.

[146] M. Cheikh and M. Zrigui, "Active learning based framework for image captioning corpus creation," in *Proc. LION*, vol. 12096, 2020, pp. 128–142.

[147] K. Konyushkova, R. Sznitman, and P. Fua, "Geometry in active learning for binary and multi-class image segmentation," *Comput. Vis. Image Understand.*, vol. 182, pp. 1–16, May 2019.

[148] Y. Qiao et al., "CPRAL: Collaborative panoptic-regional active learning for semantic segmentation," in *Proc. AAAI*, 2022, pp. 2108–2116.

[149] J. Wu, J. Chen, and D. Huang, "Entropy-based active learning for object detection with progressive diversity constraint," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9387–9396.

[150] T. Yuan et al., "Multiple instance active learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 5330–5339.

[151] R. Caramalau, B. Bhattarai, and T.-K. Kim, "Active learning for Bayesian 3D hand pose estimation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3418–3427.

[152] M. Shukla and S. Ahmed, "A mathematical analysis of learning loss for active learning in regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3315–3323.

[153] D. Yuan, X. Chang, Y. Yang, Q. Liu, D. Wang, and Z. He, "Active learning for deep visual tracking," 2021, *arXiv:2110.13259*.

[154] Z. Chen, J. Zhao, M. Yang, W. Zhou, and H. Li, "Multi-target active object tracking with Monte Carlo tree search and target motion modeling," 2022, *arXiv:2205.03555*.

[155] X. Xu, L. Liu, X. Zhang, W. Guan, and R. Hu, "Rethinking data collection for person re-identification: Active redundancy reduction," *Pattern Recognit.*, vol. 113, May 2021, Art. no. 107827.

[156] J.-J. Cai, J. Tang, Q.-G. Chen, Y. Hu, X. Wang, and S.-J. Huang, "Multi-view active learning for video recommendation," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 2053–2059.

[157] L. Zhao, S. J. Pan, and Q. Yang, "A unified framework of active transfer learning for cross-system recommendation," *Artif. Intell.*, vol. 245, pp. 38–55, Apr. 2017.

[158] P. Gupta, R. Jindal, and A. Sharma, "Community trolling: An active learning approach for topic based community detection in big data," *J. Grid Comput.*, vol. 16, no. 4, pp. 553–567, Dec. 2018.

[159] E. Chien, A. M. Tulino, and J. Llorca, "Active learning in the geometric block model," in *Proc. AAAI*, 2020, pp. 3641–3648.

[160] K. Deng, J. Pineau, and S. Murphy, "Active learning for personalizing treatment," in *Proc. IEEE Symp. Adapt. Dyn. Program. Reinforcement Learn. (ADPRL)*, Apr. 2011, pp. 32–39.

[161] K. Jedoui, R. Krishna, M. Bernstein, and L. Fei-Fei, "Deep Bayesian active learning for multiple correct outputs," 2019, *arXiv:1912.01119*.

[162] M. Moradshahi, V. Tsai, G. Campagna, and M. Lam, "Contextual semantic parsing for multilingual task-oriented dialogues," in *Proc. EACL*, 2023, pp. 902–915.

[163] Z. Li and G. Haffari, "Active learning for multilingual semantic parser," in *Proc. Findings Assoc. Comput. Linguistics, EACL*, 2023, pp. 621–627.

[164] Z. Zhou, J. Shin, L. Zhang, S. Gurudu, M. Gotway, and J. Liang, "Fine-tuning convolutional neural networks for biomedical image analysis: Actively and incrementally," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4761–4772.

[165] M. Bilgic, L. Mihalkova, and L. Getoor, "Active learning for networked data," in *Proc. ICML*, 2010, pp. 79–86.

[166] H. Cai, V. W. Zheng, and K. C.-C. Chang, "Active learning for graph embedding," 2017, *arXiv:1705.05085*.

[167] L. Gao, H. Yang, C. Zhou, J. Wu, S. Pan, and Y. Hu, "Active discriminative network representation learning," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 2142–2148.

[168] A. Cheng et al., "Deep active learning for anchor user prediction," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 2151–2157.

[169] X. Yue, Y. Wen, J. H. Hunt, and J. Shi, "Active learning for Gaussian process considering uncertainties with application to shape control of composite fuselage," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 1, pp. 36–46, Jan. 2021.

[170] C. Lee, X. Wang, J. Wu, and X. Yue, "Failure-averse active learning for physics-constrained systems," *IEEE Trans. Autom. Sci. Eng.*, vol. 20, no. 4, pp. 2215–2226, Oct. 2023, doi: 10.1109/TASE.2022.3213827.

[171] C. Lee, K. Wang, J. Wu, W. Cai, and X. Yue, "Partitioned active learning for heterogeneous systems," *J. Comput. Inf. Sci. Eng.*, vol. 23, no. 4, pp. 041009–041019, Aug. 2023, doi: 10.1115/1.4056567.

[172] A. Rahman, "Algorithms of oppression: How search engines reinforce racism," *New Media Soc.*, vol. 22, no. 3, pp. 308–310, 2020.

[173] I. Sundin, P. Schulam, E. Siivola, A. Vehtari, S. Saria, and S. Kask, "Active learning for decision-making from imbalanced observational data," in *Proc. ICML*, 2019, pp. 6046–6055.

[174] D. Li, Y. Wang, K. Funakoshi, and M. Okumura, "After: Active learning based fine-tuning framework for speech emotion recognition," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2023, pp. 1–8.

[175] G. McDonald, C. Macdonald, and I. Ounis, "Active learning stopping strategies for technology-assisted sensitivity review," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 2053–2056.

[176] H. Yu, X. Yang, S. Zheng, and C. Sun, "Active learning from imbalanced data: A solution of online weighted extreme learning machine," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 4, pp. 1088–1103, Apr. 2019.

[177] H. Ishibashi and H. Hino, "Stopping criterion for active learning based on deterministic generalization bounds," in *Proc. AISTATS*, 2020, pp. 386–397.

[178] Y. Yu, R. Zhang, R. Xu, J. Zhang, J. Shen, and C. Zhang, "Cold-start data selection for better few-shot language model fine-tuning: A prompt-based uncertainty propagation approach," in *Proc. ACL*, 2023, pp. 2499–2521.

[179] L. Chen et al., "Making your first choice: To address cold start problem in vision active learning," 2022, *arXiv:2210.02442*.

[180] O. Yehuda and A. Dekel, "Active learning through a covering lens," in *Proc. NeurIPS*, vol. 35, 2022, pp. 22354–22367.

[181] X. Cao, I. W. Tsang, and J. Xu, "Cold-start active sampling via γ-tube," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 6034–6045, Jul. 2022.

[182] R. Mahmood, S. Fidler, and M. T. Law, "Low-budget active learning via Wasserstein distance: An integer programming approach," in *Proc. ICLR*, 2022.

[183] Y. Tang and S. Huang, "Active learning for multiple target models," in *Proc. NeurIPS*, 2022, pp. 38424–38435.

[184] S. Farquhar, Y. Gal, and T. Rainforth, "On statistical bias in active learning: How and when to fix it," in *Proc. ICLR*, 2021. [Online]. Available: https://openreview.net/forum?id=JiYq3eqTKY

[185] M. Zlabinger, "Efficient and effective text-annotation through active learning," in *Proc. SIGIR*, 2019, p. 1456.

[186] B. Kwak, Y. Kim, Y. J. Kim, S. Hwang, and J. Yeo, "Trustal: Trustworthy active learning using knowledge distillation," in *Proc. AAAI*, 2022, pp. 7263–7271.

[187] H. Ma, C. Li, X. Shi, Y. Yuan, and G. Wang, "Deep unsupervised active learning on learnable graphs," 2021, *arXiv:2111.04286*.

[188] S. Mamooler, R. Lebret, S. Massonnet, and K. Aberer, "An efficient active learning pipeline for legal text classification," in *Proc. Natural Legal Lang. Process. Workshop*, 2022, pp. 345–358.

[189] F. Wan, T. Yuan, M. Fu, X. Ji, Q. Huang, and Q. Ye, "Nearest neighbor classifier embedded network for active learning," in *Proc. AAAI*, 2021, pp. 10041–10048.

[190] X. Deng, W. Wang, F. Feng, H. Zhang, X. He, and Y. Liao, "Counterfactual active learning for out-of-distribution generalization," in *Proc. ACL*, 2023, pp. 11362–11377.

[191] H. Wang, W. Huang, Z. Wu, H. Tong, A. Margenot, and J. He, "Deep active learning by leveraging training dynamics," in *Proc. NeurIPS*, 2022, pp. 25171–25184.

[192] L. Zhao, G. Sukthankar, and R. Sukthankar, "Incremental relabeling for active learning with noisy crowdsourced annotations," in *Proc. IEEE 3rd Int. Conf. Privacy, Secur., Risk Trust IEEE 3rd Int. Conf. Social Comput.*, Oct. 2011, pp. 728–733.

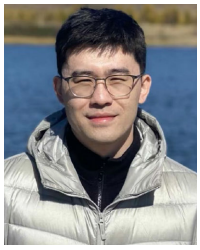
[193] Z. E. Ashari and H. Ghasemzadeh, "Mindful active learning," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 2265–2271.
[194] J. Choi et al., "VaB-AL: Incorporating class imbalance and difficulty with variational Bayes for active learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6745–6754.
[195] E. Zhao, A. Liu, A. Anandkumar, and Y. Yue, "Active learning under label shift," in *Proc. AISTATS*, vol. 130, 2021, pp. 3412–3420.
[196] J. S. Hartford, K. Leyton-Brown, H. Raviv, D. Padnos, S. Lev, and B. Lenz, "Exemplar guided active learning," in *Proc. NeurIPS*, 2020.
[197] J. Zhang, J. Katz-Samuels, and R. D. Nowak, "GALAXY: Graph-based active learning at the extreme," in *Proc. ICML*, 2022, pp. 26223–26238.
[198] R. He, Z. Han, X. Lu, and Y. Yin, "Safe-student for safe deep semi-supervised learning with unseen-class unlabeled data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14565–14574.
[199] Y. Tang and S. Huang, "Dual active learning for both model and data selection," in *Proc. IJCAI*, 2021, pp. 3052–3058.
[200] K. Ning, X. Zhao, Y. Li, and S. Huang, "Active learning for open-set annotation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 41–49.
[201] G. Hacohen, A. Dekel, and D. Weinshall, "Active learning on a budget: Opposite strategies suit high and low budgets," in *Proc. ICML*, vol. 162, 2022, pp. 8175–8195.
[202] M. Mosbach, M. Andriushchenko, and D. Klakow, "On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines," in *Proc. ICLR*, 2021.
[203] F. Tang, "Bidirectional active learning with gold-instance-based human training," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 5989–5996.
[204] S. Kothawade, N. Beck, K. Killamsetty, and R. Iyer, "Submodular information measures based active learning in realistic scenarios," in *Proc. NeurIPS*, 2021, pp. 18685–18697.
[205] S. Ertekin, J. Huang, and C. L. Giles, "Active learning for class imbalance problem," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2007, pp. 823–824.

Dongyuan Li (Graduate Student Member, IEEE) received the bachelor's degree in computer science from the Dalian University of Technology, Dalian, China, in 2018, and the master's degree from the School of Computer Science and Technology, Xidian University, Xi'an, China, in 2021. He is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering, Tokyo Institute of Technology, Tokyo, Japan.

His research interests include natural language processing, machine learning, data mining, social network analyses, and bioinformatics.

Zhen Wang received the bachelor's degree in computer science from Beihang University, Beijing, China, in 2018, and the master's degree from the Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, Delft, The Netherlands, in 2022. He is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering, Tokyo Institute of Technology, Tokyo, Japan.

His research interests include natural language processing, computer vision, and multimodal learning.



Yankai Chen received the B.S. degree from Nanjing University, Nanjing, China, in 2016, and the M.S. degree from The University of Hong Kong, Hong Kong, in 2018. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong.

His research interests include data mining and applied machine learning for database management and information retrieval.

Renhe Jiang (Member, IEEE) received the B.E. degree in software engineering from the Dalian University of Technology, Dalian, China, in 2012, the M.S. degree in information science from Nagoya University, Nagoya, Japan, in 2015, and the Ph.D. degree in civil engineering from The University of Tokyo, Tokyo, Japan, in 2019.

From 2019 to 2022, he was an Assistant Professor with the Information Technology Center, The University of Tokyo, where he is currently a Lecturer with the Center for Spatial Information Science. His research interests include spatiotemporal data mining, multivariate time series, and graph neural networks.

Weiping Ding (Senior Member, IEEE) received the Ph.D. degree in computer science from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2013.

In 2016, he was a Visiting Scholar with the National University of Singapore, Singapore. From 2017 to 2018, he was a Visiting Professor with the University of Technology Sydney, Ultimo, NSW, Australia. He is currently a Full Professor with the School of Information Science and Technology, Nantong University, Nantong, China, and a Supervisor of Ph.D. postgraduate students with the Faculty of Data Science, City University of Macau, Macao, China. He has published over 250 articles, including over 110 IEEE TRANSACTIONS. His 19 authored/coauthored papers have been selected as ESI Highly Cited Papers. He has coauthored four books. He holds 28 approved invention patents, including two U.S. patents and one Australian patent. His main research directions involve deep neural networks (DNNs), multimodal machine learning, and medical image analysis.

Dr. Ding serves as an Associate Editor/Editorial Board Member of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON FUZZY SYSTEMS, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON INTELLIGENT VEHICLES, IEEE TRANSACTIONS ON ARTIFICIAL INTELLIGENCE, *Information Fusion*, *Information Sciences*, *Neurocomputing*, and *Applied Soft Computing*. He is the Leading Guest Editor of special issues in several prestigious journals, including IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, IEEE TRANSACTIONS ON FUZZY SYSTEMS, and *Information Fusion*. He is the Co-Editor-in-Chief of *Journal of Artificial Intelligence and Systems* and *Journal of Artificial Intelligence Advances*.

Manabu Okumura was born in 1962. He received the B.E., M.E., and Dr.Eng. degrees from the Tokyo Institute of Technology, Tokyo, Japan, in 1984, 1986, and 1989, respectively.

He was an Assistant Professor with the Department of Computer Science, Tokyo Institute of Technology, from 1989 to 1992, and an Associate Professor with the School of Information Science, Japan Advanced Institute of Science and Technology, Nomi, Japan, from 1992 to 2000. He was a Visiting Associate Professor with the Department of Computer Science, University of Toronto, Toronto, ON, Canada, from 1997 to 1998. Since 2000, he has been an Associate Professor with the Precision and Intelligence Laboratory, Tokyo Institute of Technology, where he is currently a Professor with the Institute of Innovative Research. His current research interests include natural language processing, especially text summarization, computer-assisted language learning, and text data mining.