# Label-Agnostic Category Discovery

Yuwei Bian[1] Shidong Wang[2] Chunming Li[1] Haofeng Zhang[1,✉]
[1]Nanjing University of Science and Technology [2]Newcastle University
{yuwei.bian, chunmingli, zhanghf}@njust.edu.cn, shidong.wang@newcastle.ac.uk

## Abstract

*We introduce a label–agnostic paradigm for novel category discovery, designed to operate in open-world settings without relying on assumptions about train–test label space structure. Unlike prior approaches that infer categories from raw data or align to known labels, our method retrieves semantically meaningful concepts from a large-scale, ontology-grounded visual lexicon. This lexicon-guided framework enables discovery that is both scalable and semantically coherent. To support this paradigm, we propose Efficient Probabilistic Sampling (EPS) for prototype-level semantic querying, contrastive representation learning for instance and category discrimination, Adaptive Classifier Assembly (ACA) for dynamic classifier construction, and a hierarchical prototype–centroid alignment strategy for estimating category count. Taken together, these components instantiate Label-Agnostic Category Discovery (LACD) as a practical and principled solution for open-world discovery with explicit granularity control. Extensive experiments on standard benchmarks demonstrate that LACD exhibits strong clustering performance on specific unlabeled datasets when supported by a lexicon.*

## 1. Introduction

Category discovery [18, 27, 28, 42] is a fundamental challenge in computer vision, especially in open-world settings where novel unlabeled data emerge continuously. Traditional learning paradigms, including supervised learning (SL) , semi-supervised learning (SSL) [2, 41], few-shot learning (FSL) [44], zero-shot learning (ZSL) [34, 48], and Domain Adaptation (DA) [11] rely on assumptions about the relationship between training and test label spaces. These assumptions constrain generalization to predefined or semantically aligned categories, limiting scalability and adaptability. More recent efforts in novel category discovery (NCD) [18] and generalized category discovery (GCD) [42]attempt to relax these constraints by discovering unseen categories from unlabeled data. However, they still operate under disjoint or partially shared label space assumptions.

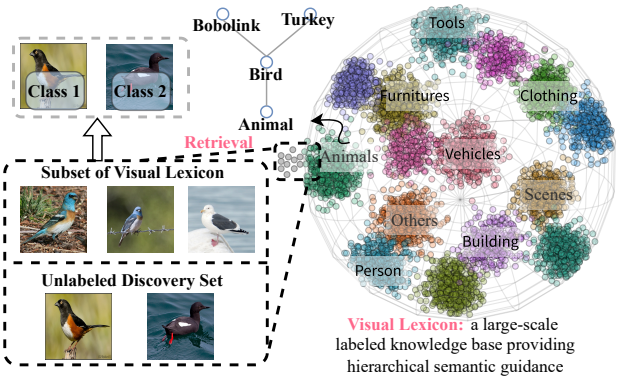


Figure 1. LACD setting. Cateogry discovery is guided by a visual lexicon via semantic similarity, transferring inter-class relations instead of label knowledge to support hierarchical, granularity-aware discovery.

In this work, we take a more pragmatic view of this problem. Rather than relying on a task-specific labeled set together with strong assumptions about how its label space relates to the target categories [3, 34, 41, 44, 48], we treat a large-scale ontology as an auxiliary semantic structure, which we refer to as a visual lexicon, not as supervision for the task but as a reusable source of semantic priors. These semantic priors offer weak, soft, and potentially mismatched priors, which can guide representation learning without assuming any explicit alignment between lexicon categories and the unknown classes in the unlabeled set.

The visual lexicon, illustrated in Figure 1, is constructed from curated datasets such as ImageNet [8] and organized into a hierarchical ontology that encodes rich semantic granularity across visual concepts. In the context of category discovery [18, 42], our objective is to identify relevant categories without imposing any assumptions on the structure or overlap of the label space. To this end, our method performs semantic retrieval over the structured lexicon using a small set of labeled samples, inducing implicit inter-class relationships that inform representation learning on unlabeled data. This retrieval-based mechanism enables category discovery to be grounded in semantic alignment

✉ Corresponding author.

Table 1. Comparison of LACD with existing learning paradigms across key capabilities. Symbols: ✓= Fully Supported, ✗= No/ Not Supported, ✎= Partial/Limited Supported.

| Paradigm | Label Space Assumption | Open-World Capability | Visual Lexicon Grounding | Ontology Use | Adaptive $K$ |
|---|---|---|---|---|---|
| Supervised Learning | Identical | ✗ | ✗ | ✗ | ✗ |
| Semi-Supervised Learning | Identical | ✎ | ✗ | ✗ | ✗ |
| Few-Shot Learning | Partial | ✎ | ✗ | ✗ | ✗ |
| Zero-Shot Learning | Disjoint (Semantic Space) | ✓ | ✗ | ✗ | ✗ |
| Domain Adaptation | Shared or Related | ✎ | ✗ | ✗ | ✗ |
| Novel Category Discovery | Disjoint | ✎ | ✗ | ✗ | ✓ |
| Generalized Category Discovery | Shared + Disjoint | ✎ | ✗ | ✗ | ✓ |
| **LACD (Ours)** | **Agnostic** | ✓ | ✓ | ✓ | ✓ |

between the lexicon and the unlabeled data, rather than relying on the label structure of the task-specific dataset.

To contextualize our contribution, we present a comparative analysis across existing learning paradigms (Table 1), demonstrating that our method uniquely satisfies all five key criteria. This capability stems from four core components. First, we propose Efficient Probabilistic Sampling (EPS), which performs prototype-level semantic querying over the visual lexicon and stochastically selects a compact, task-relevant support set. EPS enables efficient supervision in open-world settings, mitigates supervision dilution in large class spaces, and gracefully handles lexicon–task mismatch by distributing probability mass across related prototypes. Second, we adopt contrastive learning to jointly learn instance- and category-discriminative representations during semantic retrieval and relation discovery. Third, we introduce Adaptive Classifier Assembly (ACA), which episodically constructs task-specific classifiers from active prototypes and unlabeled centroids, reducing negative transfer and maintaining alignment with the evolving structure of the unlabeled data at the target granularity. Lastly, we estimate the number of categories via hierarchical clustering, aligning clusters to lexicon prototypes using Hungarian matching, scoring candidate counts by alignment accuracy and centroid similarity, and refining the estimate based on unlabeled-dominant clusters.

Taken together, these components instantiate Label-Agnostic Category Discovery (LACD) as a practical setting for open-world discovery with explicit granularity control. Our contributions are threefold: (1) We propose a label space–agnostic paradigm for category discovery that operates via semantic retrieval from hierarchical visual lexicons, enabling discovery without assumptions about label space structure or overlap; (2) We introduce a suite of mechanisms, including EPS, contrastive representation learning, and ACA, that collectively support compact, task-relevant supervision, mitigate supervision dilution, and maintain alignment with the evolving structure of unlabeled data; and (3) We develop a hierarchical prototype–centroid alignment strategy to estimate the number of categories, leveraging Hungarian matching and centroid similarity scoring to refine discovery under lexicon–task mismatch. We conducted extensive experiments on standard image classification benchmarks to evaluate the effectiveness of our proposed Label-Agnostic Category Discovery (LACD).

## 2. Related work

### 2.1. Open-Vocabulary Recognition

Open-vocabulary recognition (OVR) [47] builds upon the alignment between visual and textual representations, aiming to recognize unseen categories through textual descriptions. This task requires strong cross-modal semantic understanding and alignment, which are typically achieved by large-scale vision-language pretraining [31]. CLIP [37] learns a shared embedding space for images and texts via contrastive learning on a massive collection of image–text pairs, enabling cross-modal semantic alignment and prompt-based classification. However, CLIP relies heavily on manually curated data. ALIGN [22] demonstrates that, with sufficiently large-scale data, the presence of noise does not necessarily hinder the model's ability to learn robust visual-textual representations. In the line of prompt learning [24, 52], CoOP [53] replaces manually designed prompts with a set of learnable context vectors, easing the computational pressure of full fine-tuning. Unlike methods that use category names directly as textual prompts, CuPL [35] leverages responses generated by large language models as class-specific prompts, enriching the textual representation of each category. VPT [23] further extends prompt learning to the visual encoder, keeping the backbone frozen while introducing a small number of learnable parameters to adjust the Vision Transformer[10] input, achieving efficient adaptation.

### 2.2. Category Discovery

Category discovery abandons the assumption that unlabeled samples share the same label space as the labeled data [18]. Instead, it aims to represent and cluster unlabeled data belonging to related but disjoint semantic spaces, guided by the structure learned from labeled data. In the absence of explicit category supervision, representation learn-

ing for unlabeled data primarily relies on pairwise relationships among samples . Early studies [19–21, 51] focused on designing efficient pairwise pseudo-labeling strategies from statistical [17, 18, 51] perspectives. Some approaches [12, 40] further assumed the number of classes to be given and employed the Sinkhorn-knopp(SK) [6] algorithm to generate pseudo-labels, often incorporating residual learning [30], or prototype learning [49] to enhance clustering performance.

GCD [42] explicitly defines the labeled class space as a subset of the unlabeled one, significantly increasing the task's complexity. Its joint supervised [25] and unsupervised [16] contrastive learning on latent representations has become a foundation for subsequent research. Building upon this framework, PrCAL further incorporates prompt-driven feature enhancement; CMS [5] alleviates the strong repulsion towards negative samples in unsupervised contrastive learning through the mean-shift strategy [13], and SelEX [38] introduces a hierarchical architecture that assigns relevance weights to samples; SimGCD [46] adopts a prototype-based representation framework; SPTNet [43] leverages visual prompt learning to enhance image representations and DebGCD [29] mitigates the pseudo-label bias in self-distillation by introducing an additional debiasing classifier.

# 3. Label-Agnostic Category Discovery

## 3.1. Task Definition

The overarching goal of the proposed LACD is to discover semantic categories in unlabeled data by leveraging retrieval from a fundamental visual lexicon, without relying on label supervision or label space relationships. Let $\mathcal{V} = (\mathcal{C}, \mathcal{T}, \mathcal{X}_G)$ denote a hierarchical visual lexicon derived from a large-scale dataset, where $\mathcal{C}$ is a set of labeled categories, $\mathcal{T}$ is a coarse-to-fine taxonomy over $\mathcal{C}$, and $\mathcal{X}_G = \{(x_i^l, y_i^G)\}_{i=1}^N$ is a labeled dataset with $y_i^G \in \mathcal{C}$. The taxonomy $\mathcal{T}$ defines $G$ granularities $\{\mathcal{C}^{(1)}, \dots, \mathcal{C}^{(G)}\}$, ranging from coarse ($\mathfrak{g} = 1$) to fine ($\mathfrak{g} = G$), forming a disjoint cover $\biguplus_{\mathfrak{g}=1}^{G} \mathcal{C}^{(\mathfrak{g})} = \mathcal{C}$, with parent–child edges $\mathcal{C}^{(\mathfrak{g})} \to \mathcal{C}^{(\mathfrak{g}+1)}$.

Given an unlabeled task set $\mathcal{X}_U = \{x_j^u\}_{j=1}^n$, a target granularity $\mathfrak{g}^\star \in \{1, \dots, G\}$ (e.g., "Animal vs. Vehicle" at a coarse level or "bird species" at a fine level), the goal is to cluster $\mathcal{X}_U$ at granularity $\mathfrak{g}^\star$. Concretely, the task proceeds by retrieving a compact subset of lexicon categories $\mathcal{C}_S \subseteq \mathcal{C}^{(\mathfrak{g}^\star)}$ and their associated labeled samples: $\mathcal{X}_S = \{x_i^G \mid y_i^G \in \mathcal{C}_S\}_{i=1}^n$ that are most relevant to $\mathcal{X}_U$. Using $\mathcal{X}_S$ and $\mathcal{X}_U$, the objective is to learn $\mathfrak{g}^\star$-consistent representations to retrieve a relevant subset $\mathcal{C}_S \subseteq \mathcal{C}^{(\mathfrak{g}^\star)}$ and produce cluster assignments $\mathcal{Q} = \{q_j\}_{j=1}^n$ for $\mathcal{X}_U$, where $q_j \in \{1, \dots, K\}$ and $K$ is inferred from data.

## 3.2. Efficient Probabilistic Sampling

The key to realizing LACD lies in effectively identifying and efficiently retrieving category-relevant knowledge from the visual lexicon. Strong category-relevant relations can be established by comparing similarities between unlabeled instances and lexicon categories. Building on the success of GCD [42], we adopt a Vision Transformer (ViT-B/16) [9] pretrained with DINO self-supervision on unlabeled ImageNet-1K [4] as the backbone. This backbone induces holistic global representations, $f_\theta : \mathcal{X} \to \mathbb{R}^d$, which provide strong instance-level discrimination and guide the construction of semantic prototypes $\{\mathbf{p}_c^{(\mathfrak{g}^\star)}\}_{c \in \mathcal{C}^{(\mathfrak{g}^\star)}}$ at the target granularity $\mathfrak{g}^\star$.

Given the category set $\mathcal{C}^{(\mathfrak{g}^\star)}$ at granularity $\mathfrak{g}^\star$, we specifically construct one semantic prototype for each category $c \in \mathcal{C}^{(\mathfrak{g}^\star)}$. Formally,

$$\mathbf{p}_c^{(\mathfrak{g}^\star)} = \frac{1}{\left|\mathcal{X}_G^{(c,\mathfrak{g}^\star)}\right|} \sum_{(x_i^l, y_i^{\mathfrak{g}^\star}) \in \mathcal{X}_G^{(c,\mathfrak{g}^\star)}} f_\theta(x_i^l), \quad c \in \mathcal{C}^{(\mathfrak{g}^\star)}, \tag{1}$$

where $\mathcal{X}_G^{(c,\mathfrak{g}^\star)} = \{(x_i, y_i^{\mathfrak{g}^\star}) \in \mathcal{X}_G \mid y_i^{\mathfrak{g}^\star} = c\}$.

For each unlabeled instance $x_j^u$, we extract its representation $\mathbf{v}_j^u = f_\theta(x_j^u)$ and compute cosine similarities to all prototypes at granularity $\mathfrak{g}^\star$ through

$$s_{j,c} = \mathrm{sim}(\mathbf{v}_j^u, \mathbf{p}_c^{(\mathfrak{g}^\star)}) = \hat{\mathbf{v}}_j^{u\top} \hat{\mathbf{p}}_c^{(\mathfrak{g}^\star)}, \quad c \in \mathcal{C}^{(\mathfrak{g}^\star)}. \tag{2}$$

We then convert these similarities into a temperature-scaled softmax distribution

$$w_{j,c} = \frac{\exp\big(s_{j,c}/\tau\big)}{\sum_{c' \in \mathcal{C}^{(\mathfrak{g}^\star)}} \exp\big(s_{j,c'}/\tau\big)}, \tag{3}$$

where $w_{j,c}$ represents the softmax weight over category $c$ for instance $x_j^u$.

While contemporary retrieval methods [7, 14] offer sub-linear or otherwise improved per-query cost compared to brute-force search, the cumulative cost in LACD becomes prohibitive. This is because retrieval must be performed for every unlabeled instance at each training epoch, and sufficient epochs are required to achieve precise category alignment.

To mitigate this cost, we propose an efficient probabilistic sampling (EPS) strategy. For each unlabeled instance $x_j^u$, we sample one category from the softmax distribution

$$c_j \sim \mathtt{Cat}\big(\{w_{j,c}\}_{c \in \mathcal{C}^{(\mathfrak{g}^\star)}}\big). \tag{4}$$

Then, we uniformly sample one labeled instance from the corresponding category pool to form the support set

$$\mathcal{I}_j \sim \mathtt{Uniform}\big(\mathcal{X}_L^{(c_j,\mathfrak{g}^\star)}, 1\big), \tag{5}$$

where $\mathcal{I}_j$ denotes the $j^{\text{th}}$ support instance. This sampling enables the transfer of visual-lexicon priors to unlabeled

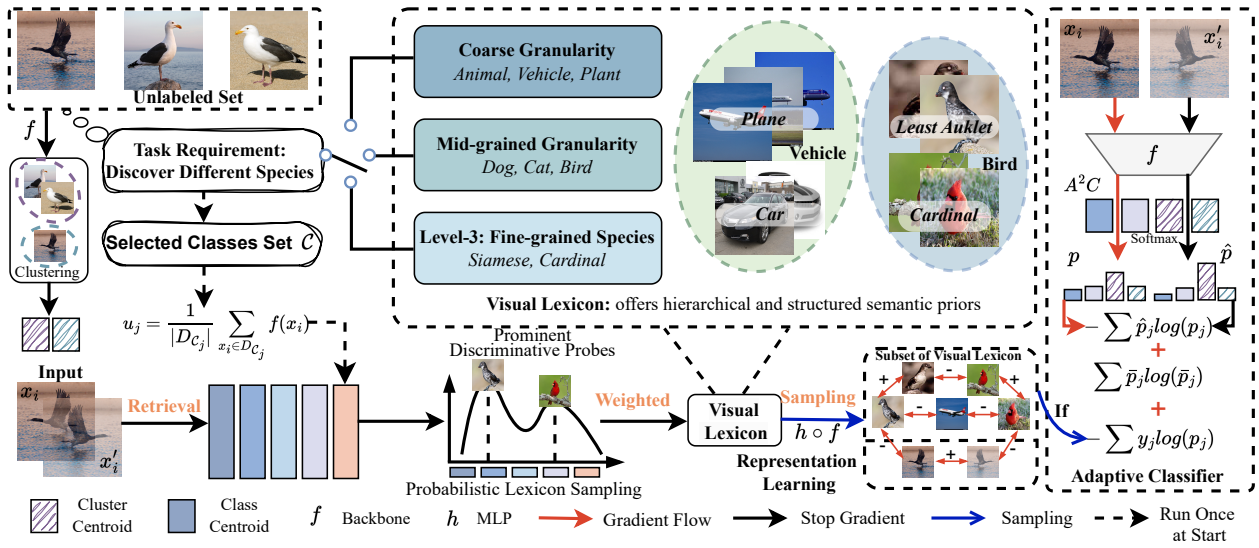


Figure 2. Pipeline Overview of LACD: We inject semantic priors via a Visual Lexicon. At the target granularity, we first perform a one-time initialization by extracting features from lexicon entries with a pretrained backbone and computing prototypes, which serve as structured semantic anchors. During training, unlabeled samples query these prototypes to retrieve visually consistent neighbors, yielding a candidate categories. Guided by the retrieved candidates, we sample data from the lexicon and conduct joint contrastive learning. On top of this, we assemble the classifier using both the prototypes of the candidate categories and the cluster centroids estimated from the unlabeled data. Finally, we apply self-distillation to regularize the classifier and features, improving stability and alignment.

data at the target granularity $\mathfrak{g}^\star$, facilitating contrastive learning to discover between-class relations at this granularity.

**Complexity and Scalability.** Let $n = |\mathcal{X}_U|$ be the number of unlabeled instances and $d$ the feature dimension. Computing similarities in Eq. (2) incurs a cost of $\mathcal{O}\big(n\,|\mathcal{C}^{(\mathfrak{g}^\star)}|\,d\big)$ per iteration. Sampling $s$ categories and one instance per category costs $\mathcal{O}\big(n\,(s+1)\big)$, which is negligible compared to the similarity computation. By contrast, exhaustive instance-level retrieval scales as $\mathcal{O}\big(n\,|\mathcal{X}_L|\,d\big)$ per iteration. Since typically $|\mathcal{C}^{(\mathfrak{g}^\star)}| \ll |\mathcal{X}_L|$, our EPS design yields an approximate speedup of $\mathcal{O}\big(|\mathcal{X}_L|/|\mathcal{C}^{(\mathfrak{g}^\star)}|\big)$ while preserving lexicon guidance.

Moreover, the stochastic sampling strategy is *robust-by-design* to lexicon–task mismatch: even when no exact category exists in the lexicon, probability mass is softly distributed over semantically related prototypes. This mitigates retrieval noise and injects semantic diversity, both of which are beneficial for category discovery.

### 3.3. Contrastive Representation Learning

Effective retrieval and relation alignment require feature representations that are both instance-discriminative and category-discriminative, achieved through joint supervised and unsupervised contrastive learning. Given an unlabeled mini-batch $\mathcal{B}_U = \{x_j^u\}_{j=1}^B$, EPS selects a labeled mini-batch of equal size, $\mathcal{B}_L = \{(x_i^l, y_i^{\mathfrak{g}^\star})\}_{i=1}^B$, with labels restricted to $\mathcal{C}^{(\mathfrak{g}^\star)}$, reducing the in-batch label space relative to the full lexicon and mitigating supervision dilution.

Let $h_\phi$ denote an MLP projection head. For the labeled batch $\mathcal{B}_L$, we apply a supervised contrastive loss to promote category-level consistency. Specifically, for each anchor $x_i^l$ we pull together all other instances in the batch with the same label and push apart those with different labels

$$\mathcal{L}_{\text{sup}} = \frac{1}{B} \sum_{x_i^l \in \mathcal{B}_L} \frac{-1}{|\mathcal{P}(x_i^l)|} \sum_{j \in \mathcal{P}(x_i^l)} \log \frac{\exp\big(\mathbf{z}_i^\top \mathbf{z}_j / \tau_s\big)}{\sum_{q \in \mathcal{B}_L \setminus \{x_i^l\}} \exp\big(\mathbf{z}_i^\top \mathbf{z}_q / \tau_s\big)}, \tag{6}$$

where $\mathcal{P}(x_i^l) = \{x_j^l \in \mathcal{B}_L \mid y_j^{(\mathfrak{g}^\star)} = y_i^{(\mathfrak{g}^\star)}, j \neq i\}$, $\mathbf{z}_i = (h_\phi \circ f_\theta)(x_i^l)$, and $\tau_s$ is a temperature hyperparameter. This promotes within-class consistency and encourages separation among semantically related classes at the target granularity $\mathfrak{g}^\star$.

To complement this, we apply an unsupervised InfoNCE loss to encourage instance-level discrimination. This objective pulls together different views of the same image while pushing apart views of other images in the batch. Specifically, let $\mathcal{B} = \mathcal{B}_U \cup \{x_i^l\}_{i=1}^B$ be the combined batch of unlabeled and sampled labeled instances, and draw two stochastic views per image. For each $x \in \mathcal{B}$, let $x'$ denote its corresponding augmented view. The unsupervised InfoNCE

objective is then defined as

$$\mathcal{L}_{\text{un}} = \frac{1}{|\mathcal{B}|} \sum_{x_i \in \mathcal{B}} -\log \frac{\exp\left(\mathbf{z}_i^\top \mathbf{z}_i'\right)}{\sum_{j \in \mathcal{B} \setminus \{x_i\}} \exp\left(\mathbf{z}_i^\top \mathbf{z}_j\right)}, \tag{7}$$

where $\mathbf{z}_i = (h_\phi \circ f_\theta)(x_i)$ and $\mathbf{z}_i' = (h_\phi \circ f_\theta)(x_i')$ are the projected representations of the original and augmented views, respectively.

Together, these objectives ensure that the learned features are both semantically aligned and robust to instance-level variation. The combined contrastive loss is

$$\mathcal{L}_{rep} = \mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{un}}. \tag{8}$$

### 3.4. Adaptive Classifier Assembly

Based on the learned representations, we construct an Adaptive Classifier Assembly (ACA) aligned with granularity $\mathfrak{g}^\star$, parameterized by lexicon prototypes and unlabeled cluster centroids. Concretely, at each iteration, we compute $\ell_2$-normalized features for the full unlabeled set $\mathcal{X}_U$ using $f_\theta$, yielding $\mathcal{V}_U = \{\mathbf{v}_j^u\}_{j=1}^n$, where $n = |\mathcal{X}_U|$. We specifically apply hierarchical clustering with Ward linkage [45] on $\mathcal{V}_U$ to obtain cluster assignments $\{q_j\}$ and centroids $\{\widehat{\mathbf{p}}_k\}$, indexed by $\mathcal{K}_t$ for the current iteration.

From the labeled batch $\mathcal{B}_L = \{(x_i^l, y_i^{\mathfrak{g}^\star})\}_{i=1}^B$, we identify the set of active granularity-$\mathfrak{g}^\star$ categories

$$\mathcal{Y}_B = \texttt{Unique}\left(\{y_i^{\mathfrak{g}^\star} \mid x_i^l \in \mathcal{B}_L\}\right). \tag{9}$$

We then assemble the classifier by concatenating the corresponding lexicon prototypes and the cluster centroids

$$W = \left[\{\mathbf{p}_c^{(\mathfrak{g}^\star)}\}_{c \in \mathcal{Y}_B} \cup \{\widehat{\mathbf{p}}_k\}_{k \in \mathcal{K}_t}\right], \tag{10}$$

$$T = |\mathcal{Y}_B| + |\mathcal{K}_t|, \tag{11}$$

where $W \in \mathbb{R}^{d \times T}$ is the assembled classifier head containing $T$ semantic anchors.

We use the ACA to classify normalized embeddings $\mathbf{v}_i$ by computing logits and predictions as follows:

$$\mathbf{o}_i = \frac{W^\top \mathbf{v}_i}{\tau_o}, \qquad p_i^{(t)} = \frac{\exp\left(o_i^{(t)}\right)}{\sum_{t=1}^T \exp\left(o_i^{(t)}\right)}, \tag{12}$$

where $\tau_o$ is the output temperature, and $p_i^{(t)}$ denotes the predicted probability for class $t$.

For labeled instances in $\mathcal{B}_L$, we apply a standard cross-entropy loss:

$$\mathcal{L}_{\text{cls}}^s = -\frac{1}{B} \sum_{x_j \in \mathcal{B}_L} \sum_{t=1}^T y_j^{(t)} \log p_j^{(t)}, \tag{13}$$

where $y_j^{(t)}$ is the one-hot ground truth label for class $t$.

For unlabeled instances and their augmented views, we generate pseudo-labels $\tilde{p}_j$ using a lower temperature to sharpen the output distribution. The unsupervised classification loss is:

$$\mathcal{L}_{\text{cls}}^u = -\frac{1}{|\mathcal{B}|} \sum_{x_j \in \mathcal{B}} \sum_{t=1}^T \tilde{p}_j^{(t)} \log p_j^{(t)}, \tag{14}$$

where $q_j^{(t)}$ is the pseudo-label distribution for class $t$.

To encourage confident and diverse predictions, we include a mini-batch mean-entropy maximization regularizer [1]. Let $M = |\mathcal{B}|$ and define the mean prediction:

$$\bar{\mathbf{p}} = \frac{1}{2M} \sum_{i=1}^M (\mathbf{p}_i + \mathbf{p}_i'), \tag{15}$$

where $\mathbf{p}_i$ and $\mathbf{p}_i'$ are the predictions for the original and augmented views of instance $i$. We then maximize the entropy:

$$H(\bar{\mathbf{p}}) = -\sum_{t=1}^T \bar{p}_t \log \bar{p}_t. \tag{16}$$

The full classification objective is:

$$\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{cls}}^s + \mathcal{L}_{\text{cls}}^u - H(\bar{\mathbf{p}}), \tag{17}$$

and the overall training objective combines representation learning and classification:

$$\mathcal{L} = \mathcal{L}_{\text{rep}} + \mathcal{L}_{\text{cls}}. \tag{18}$$

After each update, prototypes return to the pool, and the ACA is re-assembled using only the currently active categories and clusters. This episodic assembly exposes the model to a compact, task-relevant label space, reducing negative transfer from irrelevant classes and strengthening relation discovery at granularity $\mathfrak{g}^\star$.

### 3.5. Estimating Category Cardinality

Our LACD operates with a proxy labeled set retrieved from a visual lexicon, which is typically misaligned with the true open-world category space. As a consequence, selecting the number of clusters $K$ solely based on clustering accuracy over proxy labels can be misleading. For instance, when a novel class is mistakenly merged into a proxy category, the clustering accuracy may remain unchanged, introducing substantial bias in estimating $K$.

To address this, we propose a hierarchical clustering–based estimator for the number of categories. It assigns a distance-weighted score to each candidate centroid during clustering and generates the final estimate based on the proportion of unlabeled data assigned to each cluster.

We perform hierarchical clustering on the combined set $\mathcal{X}_U \cup \mathcal{X}_S$. For a given candidate number of clusters $k$, let

Table 2. Results on six benchmark datasets. **Bolding** indicates the best result and and the second-best score is underlined.

| | CUB-200 | | | Pet | | | Flower | | | Food | | | Caltech256 | | | Indoor | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | ACC | NMI | ARI | ACC | NMI | ARI | ACC | NMI | ARI | ACC | NMI | ARI | ACC | NMI | ARI | ACC | NMI | ARI |
| Kmeans | 35.8 | 0.66 | 0.22 | 73.0 | 0.84 | 0.67 | 64.3 | 0.82 | 0.48 | 39.1 | 0.55 | 0.22 | 64.7 | 0.81 | 0.52 | 46.9 | 0.63 | 0.35 |
| Take open-image V4 as the visual lexicon | | | | | | | | | | | | | | | | | | |
| GCD | 31.3 | 0.62 | 0.18 | 71.7 | 0.85 | 0.67 | 66.0 | 0.83 | 0.50 | 40.0 | 0.57 | 0.24 | 62.4 | 0.79 | 0.51 | 49.0 | 0.64 | 0.36 |
| NCE | 29.7 | 0.61 | 0.17 | 70.7 | 0.83 | 0.63 | 62.8 | 0.83 | 0.48 | 41.4 | **0.59** | **0.26** | 59.6 | 0.78 | 0.49 | 47.5 | 0.64 | 0.36 |
| LACD | 33.3 | 0.63 | 0.20 | 76.2 | 0.84 | 0.69 | 81.4 | **0.92** | 0.73 | **41.6** | 0.57 | 0.24 | 67.1 | 0.83 | 0.54 | 51.8 | 0.67 | 0.4 |
| Take ImageNet-1K as the visual lexicon | | | | | | | | | | | | | | | | | | |
| GCD | 35.1 | 0.65 | 0.22 | 75.7 | 0.86 | 0.7 | 65.5 | 0.84 | 0.53 | 39.0 | 0.57 | 0.23 | 67.0 | 0.82 | 0.56 | 50.0 | 0.66 | 0.38 |
| NCE | 29.4 | 0.61 | 0.17 | 74.8 | 0.85 | 0.69 | 66.4 | 0.84 | 0.52 | 39.1 | 0.56 | 0.23 | 60.8 | 0.78 | 0.48 | 47.4 | 0.64 | 0.36 |
| LACD | **52.0** | **0.76** | **0.4** | **86.8** | **0.90** | **0.81** | **83.6** | **0.92** | **0.74** | 40.0 | 0.58 | 0.25 | **70.5** | **0.84** | **0.58** | **52.4** | **0.68** | **0.41** |

$\hat{p}_i^k$ denote the predicted cluster assignment for sample $x_i$. We compute two optimal permutations via the Hungarian algorithm, an instance-level permutation $\pi_x$ that aligns per-sample predictions with the labeled set $\mathcal{C}_S$, and a centroid-level permutation $\pi_u$ that aligns predicted cluster centroids with class centroids from $\mathcal{C}_S$

Using these, we compute the alignment accuracy

$$\mathrm{Acc}(k) = \frac{1}{|\mathcal{X}_S|} \sum_{x_i \in \mathcal{X}_S} \mathbb{1}\{\, y_i = \pi_x(\hat{p}_i^k)\,\}\,. \tag{19}$$

The estimated number of classes is then given by:

$$\hat{K} = \arg\max_k \mathrm{Acc}(k) \cdot \left(c_i^\top \pi_u(\hat{c}_i^k)\right), \tag{20}$$

where the second term represents the optimal matching similarity between predicted cluster centers and class centroids.

Given an overclustering with $\hat{K}$ predicted clusters, we identify novel categories as those clusters dominated by unlabeled samples. For each predicted cluster $k \in \{1, \ldots, \hat{K}\}$, let $n_U(k)$ and $n_L(k)$ denote the number of unlabeled and labeled samples in cluster $k$, respectively. We define the number of novel clusters as:

$$K_n = \sum_{k=1}^{\hat{K}} \mathbb{1}\{\, n_U(k) > n_L(k)\}. \tag{21}$$

Finally, the adjusted estimate of the total number of classes is:

$$\tilde{K} = \hat{K} - |\mathcal{C}_S| + K_n, \tag{22}$$

where $\mathcal{C}_S$ is the set of known proxy categories. This formulation accounts for both known and newly discovered categories, yielding a more robust estimate of the open-world label space.

# 4. Experiment

## 4.1. Experimental Setup

**Datasets.** To assess the proposed approach for LACD, we conduct extensive evaluations on six public benchmarks (CUB-200 [39], Pet [33], Flower [32], Food [50], Caltech256 [15], Indoor [36]). We use ImageNet-1K [8] and Open Images V4 [26] as the Visual Lexicon (VL) to provide prior category knowledge for unlabeled data. To mitigate confounds introduced by hierarchical taxonomies, all experiments adopt the finest-grained category splits available in these large-scale datasets. In the LACD setting, ground-truth labels for the target tasks are unavailable; therefore, we retrieve from the VL a labeled category set matched in granularity $\mathrm{g}^\star$ and cardinality to the unlabeled set, from which we derive class priors and discriminative prototypes.

**Evaluation Protocol.** To comprehensively evaluate the performance, we adopt three widely used metrics: Clustering Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI). These metrics jointly assess the consistency between predicted cluster assignments and ground-truth labels from different perspectives, providing a holistic evaluation of clustering quality.

**Implementation Details.** Motivated by the empirical observation from GCD that Vision Transformers are effective for clustering, we adopt a DINO-pretrained ViT-B/16 as the backbone and use its 768-$d$ [CLS] token as the global representation. During training, we fine-tune only the last block of the backbone. A DINO head further maps the representation to a 256-$d$ embedding for contrastive learning.

We follow standard temperature settings with $\tau_s = 0.1$ and $\tau = 0.04$. Additionally, $\tau_o$ is linearly decayed from $0.07$ to $0.04$ over the first 30 epochs. We train with a batch size of 128. The initial learning rates are $0.1$ for the DINO head and $0.01$ for the backbone, both decayed to $1 \times 10^{-5}$ using cosine scheduling. For fair comparison, each dataset is trained for $100$ epochs, and we use the *last-epoch* checkpoint for evaluation. All experiments are conducted on NVIDIA GeForce RTX 3090 GPUs.

## 4.2. Comparison with Baseline

Open-world class discovery typically follows two lines: contrastive learning and prototype learning. We first as-

Table 3. Ablation study on the different components of our approach for LACD.

| | Unsup. | Sup. | EPS | ACA | CUB-200 | | | Pet | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | ACC | NMI | ARI | ACC | NMI | ARI |
| (1) | ✗ | ✗ | ✗ | ✗ | 35.8 | 0.66 | 0.22 | 73.0 | 0.84 | 0.67 |
| (2) | ✓ | ✗ | ✗ | ✗ | 33.0 | 0.64 | 0.20 | 71.9 | 0.84 | 0.66 |
| (3) | ✓ | ✓ | ✗ | ✗ | 37.9 | 0.68 | 0.26 | 74.8 | 0.86 | 0.70 |
| (4) | ✓ | ✓ | ✓ | ✗ | 47.3 | 0.74 | 0.35 | 77.6 | 0.88 | 0.73 |
| (5) | ✓ | ✓ | ✓ | ✓ | **52.0** | **0.76** | **0.40** | **86.8** | **0.90** | **0.81** |

Table 4. The impact of joint training on CUB-200 and Pet

| | CUB-200 | | | Pet | | |
|---|---|---|---|---|---|---|
| Method | ACC | NMI | ARI | ACC | NMI | ARI |
| w/o rep joint | 50.2 | 0.75 | 0.39 | 86.1 | 0.90 | 0.80 |
| only unlabeled | 47.0 | 0.73 | 0.36 | 85.7 | 0.89 | 0.79 |
| w/o cls joint | 48.7 | 0.75 | 0.37 | 71.5 | 0.81 | 0.62 |
| ours | **52.0** | **0.76** | **0.40** | **86.8** | **0.90** | **0.81** |

sess prototype-based methods under our LACD framework and find poor compatibility: even with targeted adaptations inspired by SimGCD, these methods still conflict with our setting, leading to unstable transfer to LACD. Therefore, we compare three representative baselines: KMeans (classical clustering), NCE (unsupervised contrastive fine-tuning), and an adapted version of GCD. To study the effect of the VL, we use Open-Images V4 and ImageNet-1K as VLs and evaluate all methods under a unified protocol on six benchmarks (as shown in Tab. 2).

Open-Images is originally a multi-label classification benchmark. To meet LACD's requirement, we reconstruct it into a 492-class single-label lexicon. However, intra-image label interference and a weaker class hierarchy than ImageNet substantially degrade overall results when using Open-Images as the VL (upper half of Tab. 2). WithOpen-Images as VL, LACD achieves 41.4/0.59/0.26 (ACC/NMI/ARI) on Food. Fine-grained inspection shows that Open-Images adopts food-related criteria that better align with the taxonomy of Food, yielding a more favorable target granularity match. This indicates that LACD is highly sensitive to the target granularity and the VL should be chosen accordingly.

Overall, NCE brings limited gains and can even damage class separability and Kmeans achieved an ACC of 35.8% on CUB-200, while NCE scored only 29.7% and 29.4% under the two VLs respectively. GCD is also constrained under LACD, primarily because the per-mini-batch supervision density from the VL is much lower than in its default setting, rendering the usual "joint contrastive + supervised guidance" nearly ineffective. These observations suggest that directly transplanting labels from a VL to assist unlabeled clustering is inefficient here, underscoring the necessity of the LACD task and method.

Table 5. Estimation of the of categories on public benchmarks.

| | CUB-200 | Pet | Flower | Food | Caltech256 | Indoor |
|---|---|---|---|---|---|---|
| GT | 200 | 37 | 102 | 100 | 257 | 67 |
| $\hat{K}$ | 328 | 50 | 192 | 123 | 320 | 106 |
| $\lvert\mathcal{C}_S\rvert$ | 150 | 20 | 100 | 50 | 150 | 50 |
| $\tilde{K}$ | 197 | 38 | 100 | 84 | 208 | 71 |
| error | 1.5% | 2.7% | 0.2% | 16.0% | 18.7% | 6.0% |

### 4.3. Ablation Study

In this subsection, we conduct a comprehensive analysis of each component of proposed method in Tab. 3.

**Representation Learning with EPS.** The first three rows of Tab. 3 show that plain unsupervised contrastive learning weakens the pretrained model's ability to preserve general features learned from large-scale unlabeled data, and it entirely ignores the rich labels available in the VL. Although supervised contrastive learning introduces labels, the supervision signal becomes ineffective to transfer when the VL spans an extremely large class space. The proposed PLS module focuses the learning supervised signal on highly relevant classes, thereby strengthening supervised contrastive modeling of inter-class relations and its transferability to unlabeled representations. As reported in the third/fourth rows of Tab. 3, adding PLS improves the ACC by 9.4% on CUB-200 and 2.8% on Pet.

**The Impact of ACA.** The proposed Adaptive Classifier Assembly breaks the constraint of a fixed classifier head by dynamically identifying and assembling active prototypes with cluster centroids into the classifier. This substantially enhances the transfer of labeled information to unlabeled data. With $A^2C$, the ACC reaches 52.8% on CUB-200 and 86.8% on Pet.

**The Impact of Joint Training.** Tab. 4 compares three settings: (i) unsupervised contrastive learning on unlabeled data only; (ii) joint representation learning on mixed mini-batches of labeled and unlabeled data without applying pseudo-supervision to the labeled data; and (iii) on top of (ii), adding pseudo-supervised learning on the labeled data. Results indicate that joint training at the representation level is critical. Such as, the ACC increases from 47.0% to 52.0% on CUB-200. Moreover, pseudo-supervision on the labeled data yields further gains. The ACC improves from 71.5% to 86.8% on Pet. Together, these two factors constitute the main sources of our overall performance improvements.

### 4.4. Estimating the Number of Classes

We report in Tab. 5 the results of estimating the number of classes using selected labeled categories as anchors. We find that the number of relevant categories is the key factor for class-number estimation, and we provide a practical selection procedure in the Appendix. By referring to both Tab. 2 and Tab. 5, it is clearly that datasets with smaller gains on performance tend to exhibit larger estimation er-

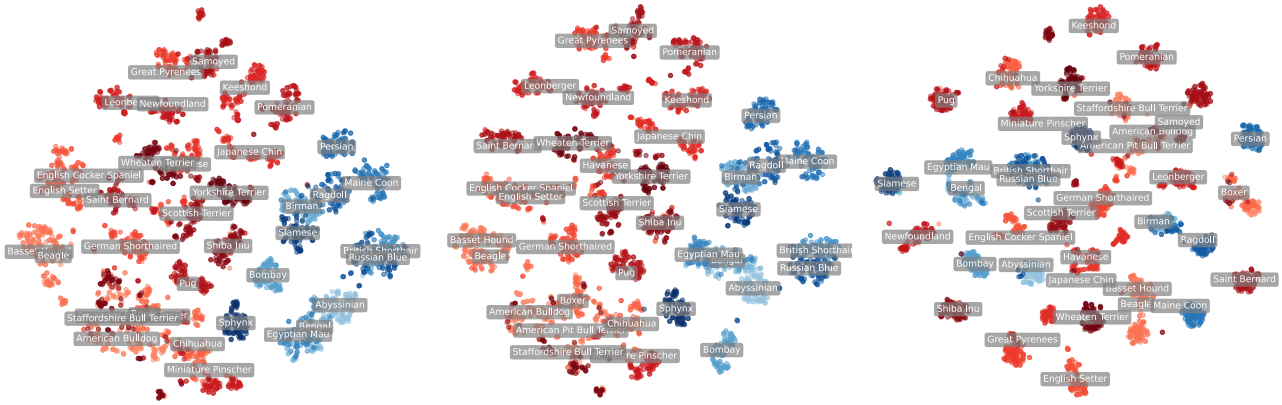


Figure 3. The t-SNE visualizations of the 25 dog classes and 12 cat classes from the PET dataset are shown in red and blue, respectively, across various stages. The initial model and simple warm-up can only distinguish the coarse-grained concepts of cats and dogs. However, our method effectively differentiates fine-grained semantics, forming well-defined, closely-knit clusters.

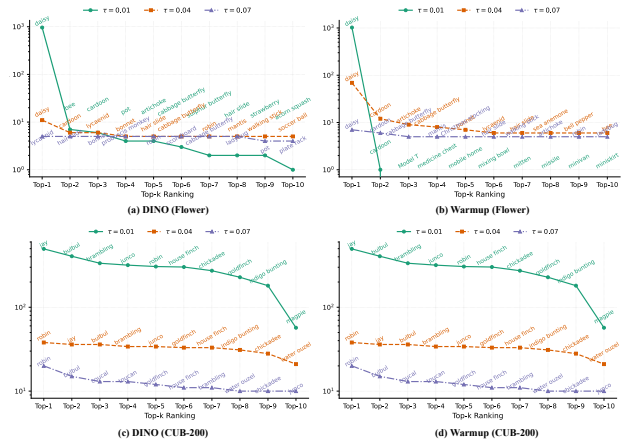


Figure 4. Distribution of selected labeled samples. Class-wise sample Distribution of top-10 categories with different $\tau$ for Flower and CUB-200, using ImageNet as the visual lexicon.

rors, and the relative error on Flower is only 0.2%, whereas on Caltech it reaches 18.7%.

### 4.5. Visualization

**Sample-Distribution Visualization.** Under the LACD setting, labels are unavailable, thus a validation split for hyperparameter selection cannot be constructed. We therefore resort to visualization to choose key hyperparameters. For the temperature $\tau_o$, which governs the selection of the labeled relevant subset, we plot on CUB-200 and Flower the class distribution and sample-count distribution of the selected labeled subset at both the initial stage and after the warm-up phase, under different $\tau_o$ values (see Fig. 4). Within the commonly used range, setting $\tau_o = 0.04$ stably retrieves classes aligned with the target granularity, avoiding over-concentration on a single class while preventing an overly dispersed selection that would dilute the supervision signal.

**t-SNE Visualization of Representations.** As shown in Fig. 3, we visualize the embeddings on Pet across training stages using t-SNE. After the warm-up fine-tuning, the embedding space exhibits limited separability for fine-grained classes and primarily reflects the coarse super-class distinction between cat and dog. In contrast, our method yields clear inter-class boundaries and compact clusters at the fine-grained level, indicating that the model's discriminative ability moves beyond coarse concepts to truly capture fine-grained category structure, thereby validating the effectiveness of LACD in granularity alignment.

## 5. Conclusion

We introduced Label-Agnostic Category Discovery (LACD), a paradigm designed to address the limitations of traditional and recent learning frameworks in open-world category discovery. By discarding assumptions about label space structure, LACD enables discovery grounded in semantic alignment rather than direct label transfer. Leveraging a hierarchical visual lexicon, our method retrieves task-relevant concepts through Efficient Probabilistic Sampling (EPS), learns robust representations via contrastive learning, dynamically assembles classifiers with Adaptive Classifier Assembly (ACA), and estimates category count through prototype–centroid alignment. Together, these components form a principled pipeline for scalable, granularity-aware discovery. Extensive experiments across standard benchmarks validate LACD as a practical and effective alternative to existing paradigms, consistently outperforming them in discovering novel categories under open-world conditions.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under the Grants No. 62371235 and No. U25A20444.

## References

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *ECCV*, pages 456–473, 2022. 5

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NIPS*, 2019. 1

[3] Yuwei Bian, Shidong Wang, Yazhou Yao, and Haofeng Zhang. Foundation-adaptive integrated refinement for generalized category discovery. *AAAI*, 40(4):2453–2461, 2026. 1

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 3

[5] Sua Choi, Dahyun Kang, and Minsu Cho. Contrastive mean-shift learning for generalized category discovery. In *CVPR*, pages 23094–23104, 2024. 3

[6] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2013. 3

[7] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2), 2008. 3

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1, 6

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 3

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2

[11] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R. Arabnia. A brief review of domain adaptation. In *Advances in Data Science and Information Engineering*, pages 877–894, 2021. 1

[12] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *ICCV*, 2021. 3

[13] K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975. 3

[14] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *ECCV*, 2016. 3

[15] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 6

[16] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010. 3

[17] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *ICLR*, 2020. 3

[18] Kai Han, Sylvestre-Alvise Rebuffi, Sébastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel: Automatically discovering and learning novel visual categories. *PAMI*, 44(10):6767–6781, 2022. 1, 2, 3

[19] Yen-Chang Hsu and Zsolt Kira. Neural network-based clustering using pairwise constraints. In *ICLR workshop*, 2016. 3

[20] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In *ICLR*, 2018.

[21] Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. In *ICLR*, 2019. 3

[22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021. 2

[23] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pages 709–727, 2022. 2

[24] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, pages 19113–19122, 2023. 2

[25] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NIPS*, pages 18661–18673, 2020. 3

[26] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 128(7):1956–1981, 2020. 6

[27] Chunming Li, Shidong Wang, and Haofeng Zhang. Adaptive gaussian expansion for on-the-fly category discovery. In *ICLR*, 2026. 1

[28] Chunming Li, Shidong Wang, and Haofeng Zhang. Learning a fix and explore framework for continuous generalized category discovery. In *AAAI*, pages 6064–6072, 2026. 1

[29] Yuanpei Liu and Kai Han. Debgcd: Debiased learning with distribution guidance for generalized category discovery. In *ICLR*, 2025. 3

[30] Yu Liu and Tinne Tuytelaars. Residual tuning: Toward novel category discovery without labels. *IEEE Transactions on Neural Networks and Learning Systems*, 34(10):7271–7285, 2023. 3

[31] Siqu Long, Feiqi Cao, Soyeon Caren Han, and Haiqin Yang. Vision-and-language pretrained models: A survey. In *IJCAI*, pages 5530–5537, 2022. 2

[32] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 6

[33] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012. 6

[34] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu. A review of generalized zero-shot learning methods. *PAMI*, 45(4):4051–4070, 2022. 1

[35] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*, pages 15691–15701, 2023. 2

[36] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *CVPR*, pages 413–420, 2009. 6

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2

[38] Sarah Rastegar, Mohammadreza Salehi, Yuki M. Asano, Hazel Doughty, and Cees G. M. Snoek. Selex: Self-expertise in fine-grained generalized category discovery. In *ECCV*, pages 440–458, 2025. 3

[39] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *CVPR*, pages 49–58, 2016. 6

[40] Mamshad Nayeem Rizve, Navid Kardan, and Mubarak Shah. Towards realistic semi-supervised learning. In *European Conference on Computer Vision*, pages 437–455. Springer, 2022. 3

[41] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440, 2020. 1

[42] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *CVPR*, 2022. 1, 3

[43] Hongjun Wang, Sagar Vaze, and Kai Han. Sptnet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. In *ICLR*, 2024. 3

[44] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.*, 53(3), 2020. 1

[45] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963. 5

[46] Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *ICCV*, pages 16544–16554, 2023. 3

[47] Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, Bernard Ghanem, and Dacheng Tao. Towards open vocabulary learning: A survey. *PAMI*, 46(7):5092–5113, 2024. 2

[48] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning - the good, the bad and the ugly. In *CVPR*, 2017. 1

[49] Lu Zhang, Lu Qi, Xu Yang, Hong Qiao, Ming-Hsuan Yang, and Zhiyong Liu. Automatically discovering novel visual categories with adaptive prototype learning. *PMAI*, 46(4):2533–2544, 2024. 3

[50] Weishan Zhang, Dehai Zhao, Wenjuan Gong, Zhongwei Li, Qinghua Lu, and Su Yang. Food image recognition with convolutional neural networks. In *UIC-ATC-ScalCom*, pages 690–693, 2015. 6

[51] Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. In *NeurIPS*, 2021. 3

[52] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 2

[53] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 2