# Parametric Classification for Generalized Category Discovery

**Abstract**

This paper focuses on the challenging problem of Generalized Category Discovery (GCD). Our aim is to identify unlabeled data in datasets containing both known and unknown categories, a critical task with practical applications, such as detecting new objects or phenomena in dynamic environments.

Building upon the existing SimGCD method, we enhance its performance by introducing mutual information maximization and similarity loss. Through detailed experimental comparisons, we observe that the improved method slightly outperforms SimGCD on certain datasets. The advantage lies in effectively identifying known categories. However, we also note a performance decline when processing unlabeled data, revealing the challenge of insufficiently mining unknown categories.

In addition, we explore strategies for handling imbalanced datasets. Class imbalance, a common issue in real-world datasets, can bias models towards more prevalent classes, neglecting the less common ones. To address this, we experiment with a resampling strategy to enhance the recognition of unknown categories while maintaining high accuracy for known categories. Nevertheless, effectively handling imbalanced datasets remains an open research question that requires further exploration and refinement in our future work.

**Keywords:** GCD, Mutual Information, Class Imbalance.

## 1 Introduction

Imagine a scene where a baby is sitting in a car, curiously observing the world outside. As the car drives, various objects appear in the baby's field of vision one after another. Babies may have learned to recognize certain objects, such as when an adult points outside and says, "Look, that's a dog," or "There's a car over there." However, there are also many objects that the baby has not seen before, such as cats or bicycles. Over time, after experiencing more and more of these scenes, we might expect an infant's visual recognition system to begin to classify these new objects into new and different categories.

This scenario raises a question worth pondering: how to deal with an image dataset in which only some images are labeled with categories and others are not labeled. Our task is to assign a category label to each unlabeled image, which may include new categories not seen in the labeled collection. We call this new problem Generalized Category Discovery (GCD) [12]. This concept is important in many machine vision applications, whether identifying products in supermarkets, analyzing lesions in medical images, or identifying

driving situations in autonomous vehicles. In these and other real-world visual environments, we often cannot predict whether a new image will belong to an already labeled category or an entirely new category.

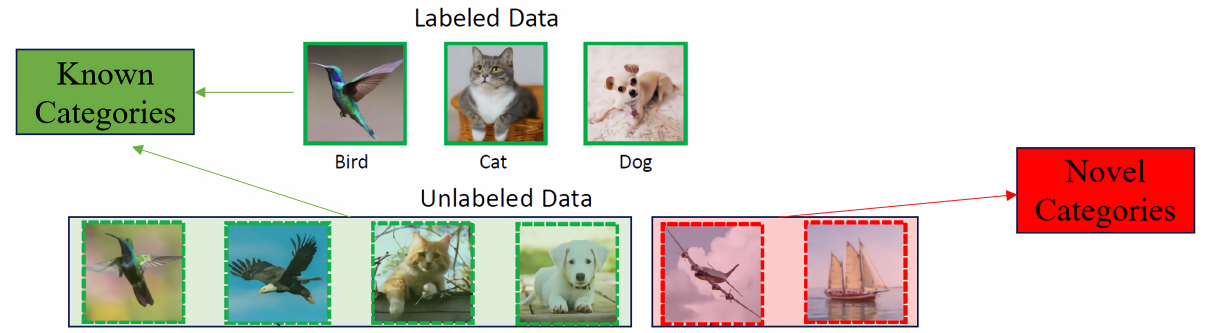


Figure 1. Schematic diagram of Generalized Category Discovery

As shown in Figure 1, the core of Generalized Category Discovery lies in how to effectively learn from limited labeled data and large amounts of unlabeled data, and apply this summarized and generalized knowledge to Realize the identification and classification of unlabeled data sets on images of known categories and unknown categories. This requires the model not only to accurately understand and remember known categories, but also to have certain innovative capabilities so that it can perform effective category inference when encountering unlabeled new categories of data.

## 2 Related works

In the field of image classification, the most common research scenario is that all images used for training have been given clear category labels, and the images in the test set belong to the existing categories in the training set. This approach may encounter limitations in practical applications because it cannot handle new categories that have not been seen during the training phase. In order to solve this problem, the semi-supervised learning (SSL) [9] method was introduced, which learns in labeled data and unlabeled data at the same time. However, SSL assumes that all unlabeled images still belong to known categories in the training set, which still limits its application in open-world environments.

### 2.1 Precursor to Generalized Category Discovery

The concepts of open set recognition (OSR) [1] and novel category discovery (NCD) [5] have become popular in recent years. The OSR task focuses on identifying images that do not belong to known classes in the training set at test time, but it does not require further classification of these images. The goal of NCD is to learn in labeled and unlabeled images and discover new categories in unlabeled collections. However the limitation of NCD is that it assumes that all unlabeled images belong to the new category, which is not always true in practical applications.

Generalized category discovery is seen as a natural extension of the NCD problem. GCD [12] proposes semi-supervised contrastive learning on large-scale pre-trained Vision Transformer (ViT), followed by semi-supervised KMeans to solve this challenging problem. However, this approach of GCD largely ignores the

potential relationships between instances of the same concept (e.g., classes, superclasses, and subclasses), which results in poor representation learning.

### 2.2 Research status

In order to solve the above shortcomings, Pu et al. [10] believed that samples belonging to the same concept should be similar to each other in the feature space, and proposed a dynamic concept contrastive learning framework, which effectively utilizes unlabeled data by considering the relationship between samples and estimating visual concepts. The underlying relationship between them improves representation learning performance.

On the other hand, Zhao et al. [16] proposed an idea to solve the problem of Generalized Category Discovery when the number of categories is unclear. They believe that representation learning and estimation of the number of categories should be considered together and can promote each other, using random A semi-supervised variant of the Gaussian mixture model of the splitting and merging mechanism is used to obtain prototypes, and these continuously updated prototypes are used for representation learning through prototype contrastive learning. They proposed an expectation-maximization EM-like framework that iteratively optimizes by alternating between representation learning and category number estimation.

There have been other important developments in this area of research. Zhang et al. [14] found that the potential of pre-trained ViT is actually suppressed by the practice of treating different unlabeled images from the same or similar semantic categories as false negatives, and freezing most of the structure of the pre-trained VIT backbone can alleviate the problem of Overfitting of known classes, but it limits the flexibility and adaptability of the model, so they proposed a two-stage process using contrastive affinity learning and visual cue learning to achieve known classes and new classes fine-grained semantic clustering.

There are also new findings and suggestions for parametric classifiers in GCD. Past studies [4, 12] have shown that parametric classifiers are prone to overfitting known categories, so it is recommended to use non-parametric classifiers formed by semi-supervised KMeans. However, Wen et al. [13] studied the failure of parametric classifiers and believed that unreliable pseudo-labels were the key factor. In order to solve this problem, they proposed to add an entropy regularization term to force the model to predict more evenly distributed labels to overcome Biased predictions, reaching SOTA on multiple datasets. In addition, Roy et al. [11] and Zhao et al. [15] also discussed how to solve the incremental learning problem of GCD.

## 3 Method

### 3.1 SimGCD

SimGCD [13] proposed a simple and effective parameter classification method, which benefited from entropy regularization, significantly improved model performance, and reached SOTA level on multiple data sets. Previous research work [4, 12] has shown that although parametric classifiers perform well when dealing with known categories, they tend to overfit these categories, resulting in performance degradation when identifying new categories. Therefore, it is recommended to use non-parametric classifiers, such as the semi-supervised K-Means clustering method. Although good results are achieved, non-parametric classifiers face huge chal-

lenges on large-scale data sets due to the time complexity required by the clustering algorithm. calculation cost. Furthermore, unlike learnable parametric classifiers, the results obtained by non-parametric methods may be suboptimal since they lack a learnable way to optimize the separating hyperplane for all classes.

SimGCD provides an in-depth analysis of the root causes of difficulties encountered by parametric classifiers in identifying new categories, pointing out that the key to the performance degradation of previous models lies in unreliable pseudo-labels. By observing the data predicted by the statistical parametric classifier, the researchers found significant prediction bias in the model, specifically the tendency to overpredict "old" classes and ignore the "new" classes, and to generate unbalanced pseudo predictions across all classes. labels, these problems eventually lead to the degradation of classification performance.

Based on these findings, SimGCD proposes a novel parameter classification method for Generalized Category Discovery. This method not only considers how to effectively utilize the information of known categories, but also optimizes the model through entropy regularization to better adapt and recognize new categories. The simplicity and effectiveness of this method are verified in the performance on multiple data sets. Figure 2 shows the similarities and differences between the original GCD method and the SimGCD method:

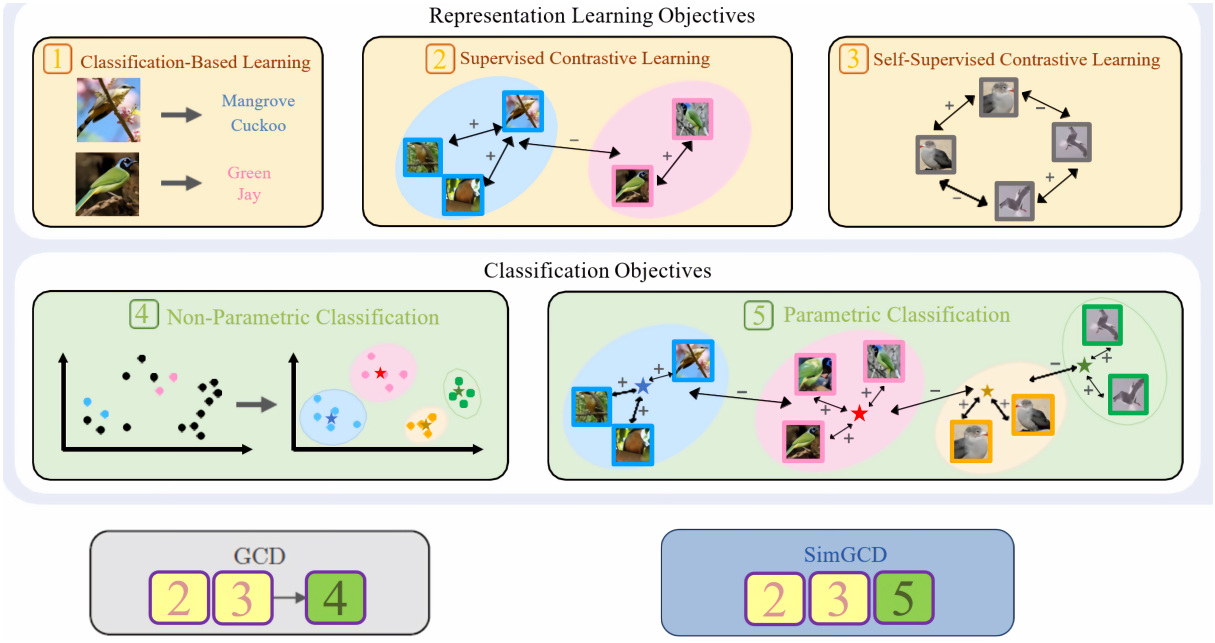


Figure 2. Similarities and differences between GCD and SimGCD

### 3.2 Representation learning

The representation learning goal of SimGCD completely follows the GCD method. Formally, let $\mathbf{x}_i$ and $\mathbf{x}_i'$ be two views of the same image (randomly enhanced) in mini-batch $B$. The loss of unsupervised contrastive learning can be written as:

$$\mathcal{L}_i^u = -\log \frac{\exp\left(\mathbf{z}_i \cdot \mathbf{z}_i'/\tau\right)}{\sum_n \mathbf{1}_{[n \neq i]} \exp\left(\mathbf{z}_i \cdot \mathbf{z}_n/\tau\right)} \tag{1}$$

where $\mathbf{z}_i = \phi\left(f\left(\mathbf{x}_i\right)\right)$; $\mathbf{1}_{[n \neq i]}$ is an indicator function, if $n \neq i$, its value is 1, otherwise it is 0; $\tau$ is the parameter used to control similarity in contrastive learning, the temperature value; $f$ is the backbone network; $\phi$ is a multi-layer perception machine (MLP) projection head.

The supervised contrastive loss can be written as:

$$\mathcal{L}_i^s = -\frac{1}{|\mathcal{N}(i)|} \sum_{q \in \mathcal{N}(i)} \log \frac{\exp\left(\mathbf{z}_i \cdot \mathbf{z}_q / \tau\right)}{\sum_n \mathbf{1}_{[n \neq i]} \exp\left(\mathbf{z}_i \cdot \mathbf{z}_n / \tau\right)} \tag{2}$$

$\mathcal{N}(i)$ represents the index of other images in mini-batch $B$ that have the same label as $\mathbf{x}_i$. Finally, the total loss function $\mathcal{L}_{\text{rep}}$ of the GCD method is:

$$\mathcal{L}_{\text{rep}} = (1-\lambda) \sum_{i \in B} \mathcal{L}_i^u + \lambda \sum_{i \in B_{\mathcal{L}}} \mathcal{L}_i^s \tag{3}$$

$B_{\mathcal{L}}$ corresponds to the labeled subset of $B$, and $\lambda$ is the weight coefficient. The contrastive learning component is only used to allow the network to learn semantically meaningful representations, thereby minimizing overfitting of the model to the labeled categories.

### 3.3 Parametric classification

The classification learning goals of SimGCD are cross-entropy of labeled samples and self-distillation of unlabeled samples. Additionally, SimGCD also employs an entropy regularization term to overcome biased predictions by forcing the model to predict more evenly distributed labels.

Formally, $K = |\mathcal{Y}_l \cup \mathcal{Y}_u|$ represents the total number of categories (assumed that the total number of categories $K$ is known), and a set of prototypes $\mathcal{C} = \{c_1, \ldots, c_K\}$ are randomly initialized, each prototype representing a category. During model training, the soft label of each augmented view $\mathbf{x}_i$ is calculated by the softmax of cosine similarity between the extracted latent feature $\mathbf{h}_i = f(\mathbf{x}_i)$ and prototype $\mathcal{C}$ scaled by $1/\tau_s$:

$$\boldsymbol{p}_i^{(k)} = \frac{\exp\left(\frac{1}{\tau_s}\left(\boldsymbol{h}_i / \|\boldsymbol{h}_i\|_2\right)^\top \left(\boldsymbol{c}_k / \|\boldsymbol{c}_k\|_2\right)\right)}{\sum_{k'} \exp\left(\frac{1}{\tau_s}\left(\boldsymbol{h}_i / \|\boldsymbol{h}_i\|_2\right)^\top \left(\boldsymbol{c}_{k'} / \|\boldsymbol{c}_{k'}\|_2\right)\right)} \tag{4}$$

The soft pseudo-label $\boldsymbol{q}_i'$ is generated in a similar way by another view $\mathbf{x}_i$ with a sharper temperature $\tau_t$ (making the similarities more obvious or discriminating, possibly causing the model to pay more attention to small differences between samples). By cross-entropy loss between predicted categories and pseudo-labels or ground truth labels,

$$\ell\left(\boldsymbol{q}_i', \boldsymbol{p}_i\right) = -\sum_k \boldsymbol{q}_i'^{(k)} \log \boldsymbol{p}^{(k)}, \ell\left(\boldsymbol{y}_i, \boldsymbol{p}_i\right) = -\sum_k \boldsymbol{y}_i^{(k)} \log \boldsymbol{p}^{(k)} \tag{5}$$

At the same time, an average entropy maximizing regularizer is used to achieve the unsupervised goal:

$$\bar{p} = \frac{1}{2|B|} \sum_{i \in B} \left(p_i + p_i'\right), H(\bar{p}) = -\sum_k \bar{p}^{(k)} \log \bar{p}^{(k)} \tag{6}$$

where $\boldsymbol{y}_i$ represents the one-hot encoding of $\mathbf{x}_i$, $\bar{p}$ represents the average prediction of the batch, and $H(\bar{p})$ is the entropy of $\bar{p}$.

$$\mathcal{L}_{\text{cls}}^u = \frac{1}{|B|} \sum_{i \in B} \ell\left(\boldsymbol{q}_i', \boldsymbol{p}_i\right) - \varepsilon H(\overline{\boldsymbol{p}}), \mathcal{L}_{\text{cls}}^s = \frac{1}{|B_{\mathcal{L}}|} \sum_{i \in B_{\mathcal{L}}} \ell\left(\boldsymbol{y}_i, \boldsymbol{p}_i\right) \tag{7}$$

Finally, the classification loss function of labeled samples and unlabeled samples is $\mathcal{L}_{\text{cls}} = (1-\lambda)\mathcal{L}_{\text{cls}}^{u} + \lambda\mathcal{L}_{\text{cls}}^{s}$. Combined with the loss function of the above representation learning, the overall loss function of SimGCD is $\mathcal{L}_{\text{rep}} + \mathcal{L}_{\text{cls}}$.

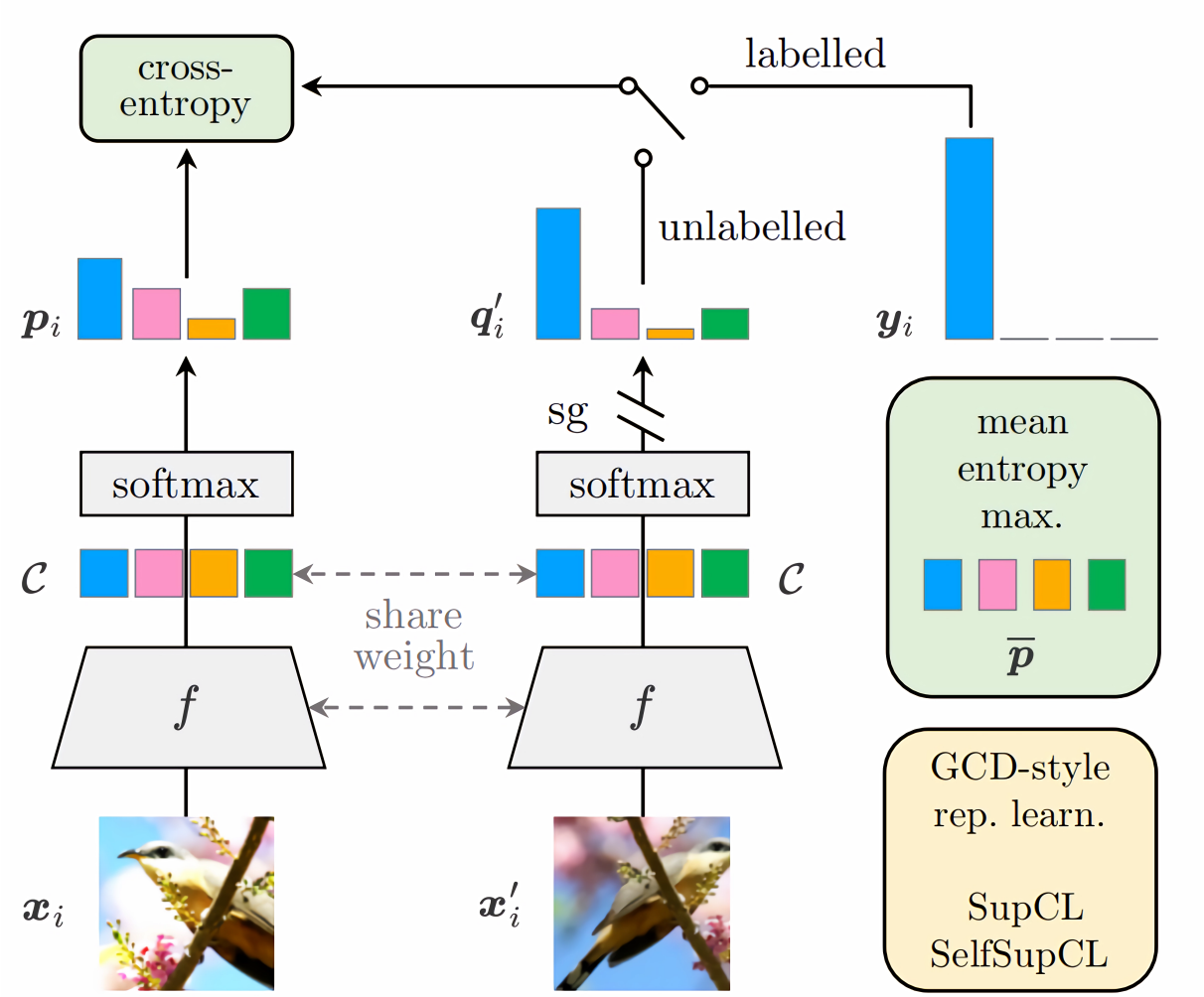


Figure 3. SimGCD overall framework diagram

## 4 Implementation details

### 4.1 Comparing with the released source codes

From the above introduction, we can know that SimGCD mainly uses contrastive learning and cross-entropy as its main training strategies, effectively using data information to improve model performance. How to make better use of data information? From the research of Chiaroni et al. [2], we can learn that the mutual information maximization method is a common method in unsupervised clustering. By maximizing the mutual information of sum, we can more accurately predict unlabeled samples. Generally, mutual information consists of the following two parts. The optimization goal is to make the mutual information as large as possible:

$$I(Y,Z) = \mathcal{H}(Y) - \mathcal{H}(Y \mid Z) \tag{8}$$

The former $\mathcal{H}(Y)$ is edge entropy, the latter $\mathcal{H}(Y \mid Z)$ is conditional entropy, and $\pi_k$ is defined as the edge

probability of class $k$, then the formula of edge entropy is defined as

$$\begin{aligned}\mathcal{H}(Y) &= -\sum_{k=1}^{K} \mathbb{P}(Y=k;\boldsymbol{W}) \log \mathbb{P}(Y=k;\boldsymbol{W}) \\ &= -\sum_{k=1}^{K} \pi_k \log \pi_k\end{aligned} \tag{9}$$

Here we can see that edge entropy actually limits the edge probability to approach a uniform distribution, reflecting the uniformity of category distribution. However, in most application scenarios, the distribution of different categories is not uniform, and long tails often appear. phenomenon, which results in the limitation of the application of edge entropy on imbalanced data sets.

In the scenario of Generalized Category Discovery, the formula for maximizing mutual information can add partial constraints, that is, constraints on label information, thereby improving the optimization goal:

$$\max_{\boldsymbol{w}} \mathcal{H}(Y) - \mathcal{H}(Y \mid Z) \quad \text{s.t. } \boldsymbol{y}_i = \boldsymbol{p}_i \quad \forall \boldsymbol{z}_i \in \mathcal{Z}_L \tag{10}$$

Transfer the constraint part to the optimization objective, then the optimization objective becomes

$$\min_{W} \sum_{k=1}^{K} \pi_k \log \pi_k - \frac{1}{|\mathcal{Z}|} \sum_{i \in \mathcal{Z}} \sum_{k=1}^{K} h_{i,k} \log p_{i,k} \tag{11}$$

$y_i$ is the real label, $p_i$ is the predicted label, where $h$ is defined as:

$$\begin{cases} h_{i,k} = y_{i,k} \text{ if } \boldsymbol{z}_i \in \mathcal{Z}_L \\ h_{i,k} = p_{i,k} \text{ otherwise} \end{cases} \tag{12}$$

It can be seen that in fact, for labeled samples, this optimization goal is equivalent to cross-entropy loss, which is reasonable. Cross-entropy strengthens the model's prediction confidence for labeled data because it pushes these predictions to a simple A vertex of the simplex shape, cross entropy and conditional entropy will reach the minimum value at the vertex of the simplex shape. For unlabeled samples, this goal is equivalent to minimizing the prediction entropy of each sample.

Since edge entropy tends to be evenly distributed, in order to further consider the imbalance problem of data, parameters can be introduced for control to obtain the outer optimization function.

$$\begin{aligned}F(\boldsymbol{W}, \lambda) = &\underbrace{\sum_{k=1}^{K} \pi_k \log \pi_k}_{\mathcal{H}(Y)} - \underbrace{\frac{1}{|\mathcal{Z}_L|} \sum_{i \in \mathcal{Z}_L} \sum_{k=1}^{K} y_{i,k} \log p_{i,k}}_{\text{CE}} \\ &- \underbrace{\frac{\lambda}{|\mathcal{Z}_U|} \sum_{i \in \mathcal{Z}_U} \sum_{k=1}^{K} p_{i,k} \log p_{i,k}}_{\propto \mathcal{H}(Y|Z)}\end{aligned} \tag{13}$$

Parameter $\lambda$ controls the role of unsupervised terms, which is equivalent to striking a balance between conditional entropy and edge entropy. The value of $\lambda$ will be obtained from the labeled data set, which can maximize the accuracy of the labeled data set, that is,

$$\min_{\boldsymbol{W}} F(\boldsymbol{W}, \lambda) \quad \text{s.t} \quad \lambda \in \arg\max_{\lambda \in (0,1]} A_L(\lambda) \tag{14}$$

where $A_L$ represents the accuracy in the labeled data set

$$A_L(\lambda) = \frac{1}{|\mathcal{Z}_L|} \sum_{i=1}^{|\mathcal{Z}_L|} \mathbf{1}_{\{\hat{\boldsymbol{y}}_i(\lambda) = \boldsymbol{y}_i\}} \tag{15}$$

At the same time, we know that data imbalance is an intrinsic feature of the real visual world. A small number of classes dominate the distribution, and many tail classes have only a few samples. The imbalance of data distribution in the real world needs to be considered.

Kang et al. [7] pointed out that in the case of class imbalance, the classifier trained only using cross entropy may not be robust enough. They can maximize the similarity between the class probability distribution a given by the classifier and the one-hot encoding b of the real label. Degree to improve the model's robustness when dealing with class-imbalanced data.

$$\mathcal{L}_{\text{sim}} = -\sum_{p \in P(i)} \log \frac{\exp\left(g_i \cdot l_p\right)/\tau}{\sum_{i \in C} \exp\left(g_i \cdot l_j\right)/\tau} \tag{16}$$

Let $\mathcal{L}_{\text{ent}} = F(\boldsymbol{W}, \lambda)$, the total loss $\mathcal{L}_{\text{total}}$ of the optimized model can be written as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rep}} + \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{sim}} + \mathcal{L}_{\text{ent}} \tag{17}$$

### 4.2 Experimental data set

To comprehensively evaluate the effectiveness of our proposed method, experiments cover several types of datasets, including common image recognition benchmarks (CIFAR10/100, ImageNet-100), as well as datasets focused on specific domains, such as CUB, Stanford Cars, and FGVC-Aircraft, these datasets represent semantic transfer challenges. Additionally, we consider Herbarium 19, a naturally unbalanced dataset, to demonstrate the ability of the method in handling real, complex data distributions.

For each data set, a subset of categories is selected as labeled category $\mathcal{Y}_l$ (Old) according to the settings of GCD, 50% of the images in these labeled categories are used to construct $\mathcal{D}_L$, and all the remaining images are regarded as unlabeled data $\mathcal{D}_U$, and the number of unlabeled categories is $|\mathcal{Y}_u| = |\mathcal{Y}_l| + |\mathcal{Y}_n|$, $\mathcal{Y}_n$ is the number of new categories. The detailed statistics of the data sets used in the experiment are as follows:

Table 1. Statistics used to evaluate the data sets

| | CIFAR10 | CIFAR100 | ImageNet100 | CUB | SCars | FGVCAircraft | Herbarium19 |
|---|---|---|---|---|---|---|---|
| $\lvert\mathcal{Y}_l\rvert$ | 5 | 80 | 50 | 100 | 98 | 50 | 941 |
| $\lvert\mathcal{Y}_u\rvert$ | 10 | 100 | 100 | 200 | 196 | 100 | 683 |
| $\lvert D_L\rvert$ | 12.5k | 20k | 31.9k | 1.5k | 2.0k | 1.7k | 8.9k |
| $\lvert D_{\mathcal{U}}\rvert$ | 37.5k | 30k | 95.3k | 4.5k | 6.1k | 5.0k | 25.4k |

### 4.3 Evaluation plan

Model performance is mainly evaluated by clustering accuracy (ACC). In the specific evaluation process, we first determine the real label $y_i^*$ and the label $\hat{y}_i$ predicted by the model, and then find the best label correspondence by solving a bipartite graph matching problem to calculate ACC:

$$ACC = \frac{1}{M} \sum_{i=1}^{M} \mathbf{1} \left( y_i^* = p \left( \hat{y}_i \right) \right) \tag{18}$$

where $M = |\mathcal{D}_U|$, $p$ is the best permutation matching the predicted cluster assignments to the true class labels.

During the experiment, we used the DINO pre-trained Vision Transformer (ViT-B/16) as the backbone network for feature extraction. The feature extraction mainly focused on the [CLS] token output with dimension 768, which was used as the image Feature representation, fine-tuning only the last block of the backbone network. For each dataset, the model was trained for 200 epochs with a batch size of 128, starting with an initial learning rate of 0.1 and gradually decaying the learning rate according to cosine scheduling as training progressed.

## 5 Results and analysis

### 5.1 Experimental results

By comparing the experimental results of SimGCD and the improved method (Ours), it can be found that after introducing mutual information maximization and data imbalance processing methods on the basis of SimGCD, the overall performance of Ours slightly exceeds SimGCD on some data sets. Because more supervision information constraints are added, Ours has a greater advantage in identifying known categories, but at the same time, due to the lack of further mining of unlabeled data, the model's performance on unknown categories declines.

On the general image recognition dataset, Ours is slightly higher than SimGCD in overall and known category recognition accuracy, but slightly lower in unknown category recognition accuracy, which shows that our method is slightly better at handling known categories. Advantages, but slightly less adaptable to unknown categories.

Table 2. Results for general image recognition datasets

| | CIFAR10 | | | CIFAR100 | | | ImageNet-100 | | |
|---|---|---|---|---|---|---|---|---|---|
| Methods | All | Old | New | All | Old | New | All | Old | New |
| SimGCD | 97.1 | 95.1 | **98.1** | **80.1** | 81.2 | **77.8** | **83.0** | 93.1 | **77.9** |
| Ours | **97.2** | **95.4** | 98.0 | 76.8 | **82.9** | 64.5 | 81.2 | **94.1** | 74.7 |

On the CUB and Stanford Cars datasets, Ours outperforms SimGCD in all categories, known categories and unknown categories, indicating that it is more effective in processing datasets with specific semantic features, but at the same time it performs better on the FGVC-Aircraft dataset There is a large drop in performance,

which may indicate that the large amount of supervised information of known categories seriously interferes with the learning of unknown categories when processing more complex and diverse data sets.

Table 3. Results for the semantic transfer benchmark datasets

| | CUB | | | Stanford Cars | | | FGVC-Aircraft | | |
|---|---|---|---|---|---|---|---|---|---|
| Methods | All | Old | New | All | Old | New | All | Old | New |
| SimGCD | 60.3 | 65.6 | 57.7 | 53.8 | 71.9 | 45.0 | **54.2** | **59.1** | **51.8** |
| Ours | **64.9** | **73.9** | **60.4** | **54.5** | **72.1** | **46.0** | 47.1 | 58.2 | 41.6 |

In the naturally imbalanced Herbarium 19 dataset, Ours slightly outperforms SimGCD, however, both methods face challenges, especially in the identification of new categories, which may require specific strategies to address the problem of data imbalance.

Table 4. Results for naturally imbalanced dataset

| | Herbarium 19 | | |
|---|---|---|---|
| Methods | All | Old | New |
| SimGCD | 44.0 | 58.0 | **36.4** |
| Ours | **44.5** | **60.7** | 35.8 |

### 5.2 Discussion on imbalanced data sets

In many practical applications, the data set is often unbalanced, that is, the number of samples of some categories is much more than that of other categories. This imbalance will seriously affect the training effect of the machine learning model.

Existing research on Generalized Category Discovery implicitly or explicitly assumes that each category (whether known or unknown) appears approximately equally frequently in unlabeled data. However, in nature, we are more likely to encounter known/common categories than unknown/uncommon categories, which is consistent with the long-tail nature of the real world.

For the long-tail data set CIFAR-10-LT [3], use SimGCD and Ours methods respectively to observe their performance:

Table 5. Results for CIFAR-10-LT and CIFAR-10 datasets

| | CIFAR-10-LT | | | CIFAR10 | | |
|---|---|---|---|---|---|---|
| Methods | All | Old | New | All | Old | New |
| SimGCD | 58.3 | 75.1 | **41.5** | 97.1 | 95.1 | **98.1** |
| Ours | **61.2** | **83.8** | 38.6 | **97.2** | **95.4** | 98.0 |

It can be seen that comparing CIFAR-10-LT and CIFAR-10, both SimGCD and Ours methods have significantly reduced performance due to data imbalance. We need to improve the applicability and robustness of the model in real-world applications.

Imbalanced solutions generally start from the perspectives of sampling, loss function adjustment, generating synthetic samples and changing decision thresholds. Referring to the open world sampling proposed by Jiang et al. [6], they used three key training strategies to improve the sample sampling effect of tail categories:

(1) Tailness: Sampling samples from tail categories is encouraged by ranking the samples based on empirical contrastive loss expectations (ECLE) based on random data augmentation. This strategy aims to ensure that more attention is paid to tail categories in training and to improve learning of these categories;

(2) Proximity: Reject outliers that may interfere with training and belong to samples outside the distribution. By excluding these outliers, it helps the training model to be more concentrated inside the data distribution, improving the stability and generalization performance of the model;

(3) Diversity: Ensure diversity in the sample set. This means that the samples sampled should not only come from tail categories, but also cover various categories to maintain the diversity of training data and help the model learn features of different categories more comprehensively.

We choose to use tailness and diversity to sample the CIFAR-10-LT data set. The distribution before sampling is as follows:

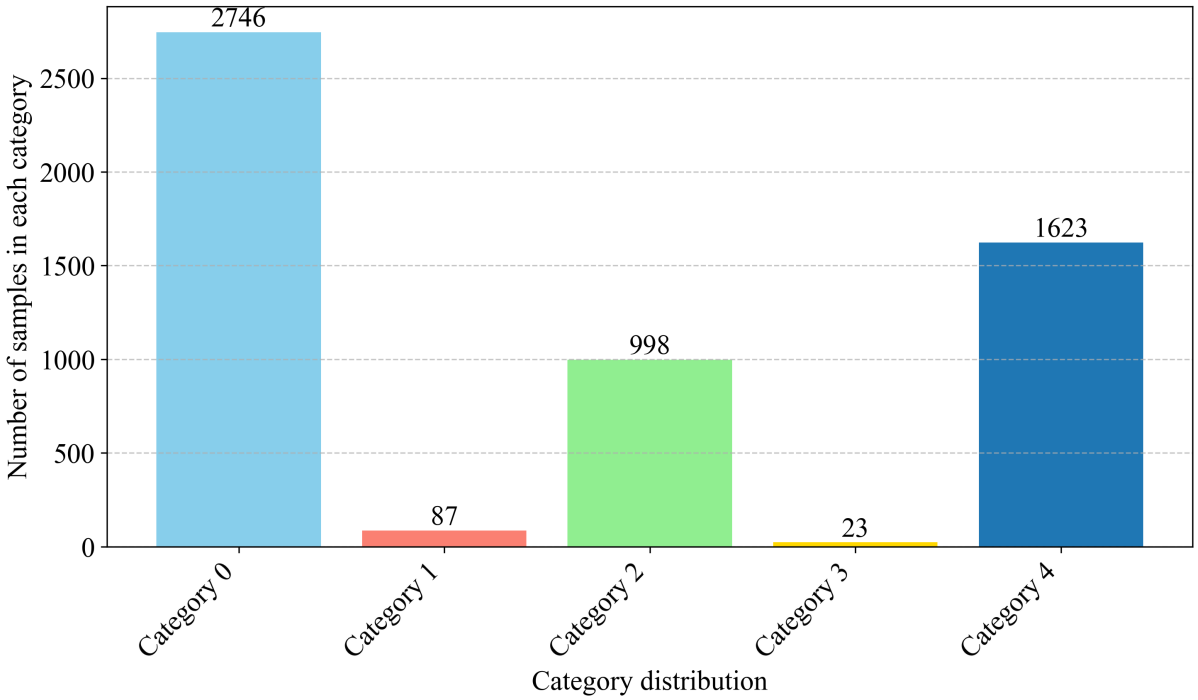


Figure 4. Labeled dataset distribution before sampling

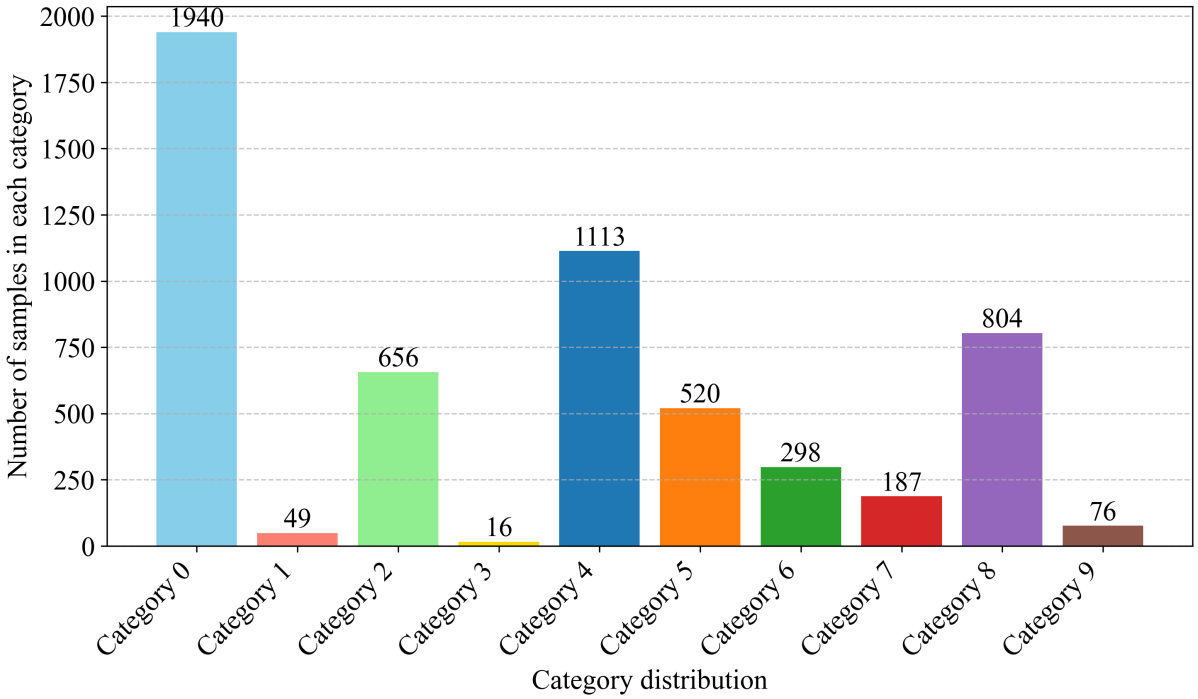


Figure 5. Unlabeled dataset distribution before sampling

Using 10% of the original data set size as the size of the sampling data set, the distribution after sampling is as follows:

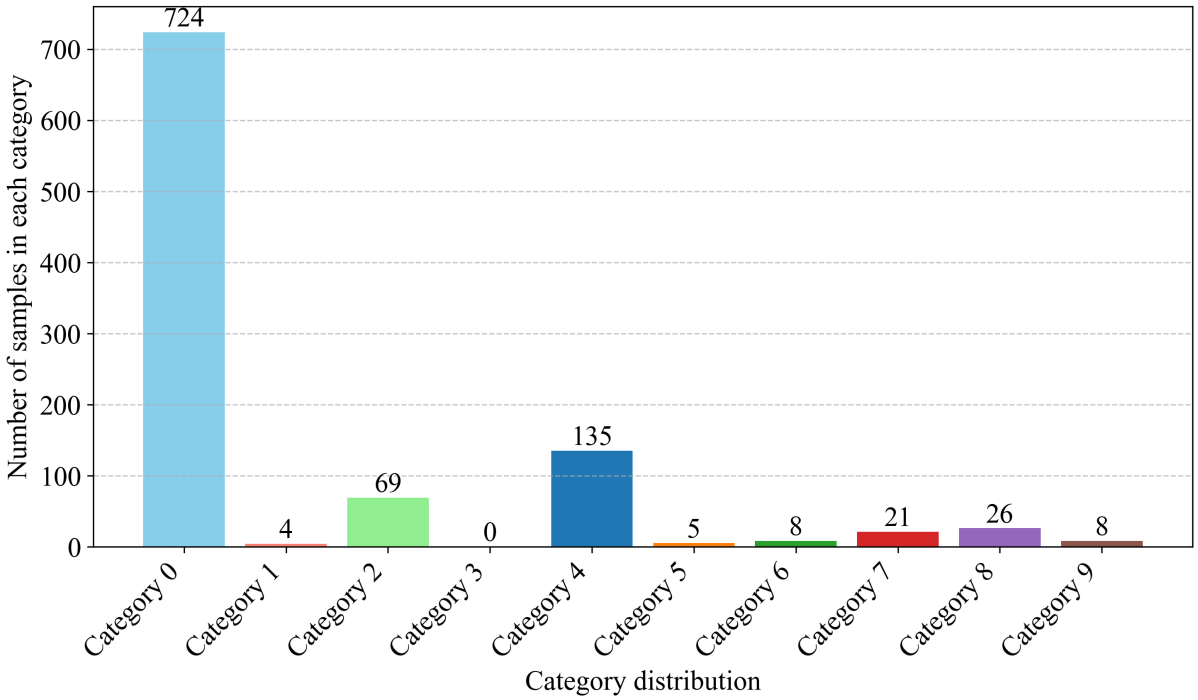


Figure 6. Distribution of data set after sampling

It can be seen that by comparing before and after sampling, the distribution is still seriously unbalanced, and further learning is needed for the processing of unbalanced data sets.

To simplify the problem, we choose first to investigate the imbalance issue within known categories. In

the H2T paper [8], a sampling strategy for labeled data is mentioned to address the long-tail problem of known categories. By adopting the sampling strategy of the H2T method, as illustrated in Figure 7, it is evident that it can effectively tackle the imbalance problem within the known category data in the training set. However, the imbalance issue for unknown categories still remains to be addressed.

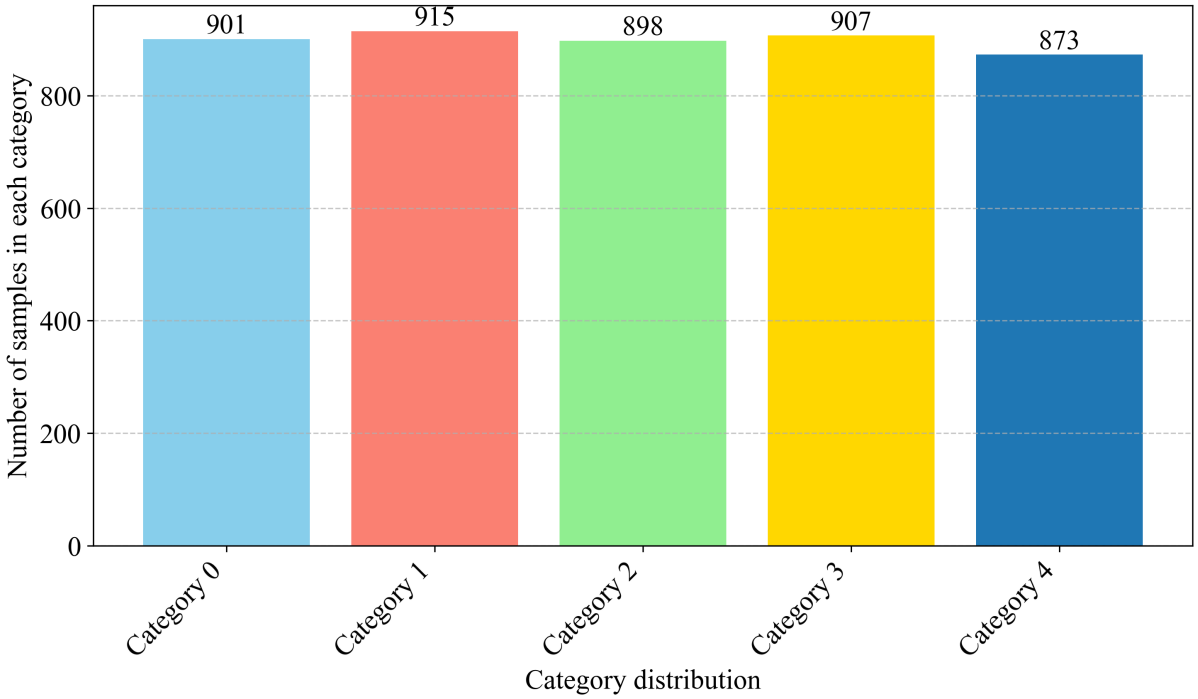


Figure 7. Known categories distribution after H2T sampling

## 6 Conclusion and future work

Generalized class discovery is an important study aimed at identifying known and unknown classes in a data set. This is particularly critical when quickly identifying new objects or phenomena in dynamic and changing environments. Our research is based on the existing SimGCD method and optimizes it by integrating techniques such as mutual information maximization and similarity loss. Experimental comparison shows that the overall performance of this optimization method is slightly better than the original SimGCD on some data sets. Its main advantage is more accurate identification of known classes. However, we also note performance degradation when processing unlabeled data, which highlights the challenge of fully mining unknown classes.

Additionally, we dive into strategies for dealing with the class imbalance problem prevalent in real-world data. Class imbalance often causes the model to be biased towards identifying classes with a large sample size and ignore those that are rarer. In order to solve this problem, we adopted a resampling strategy in order to enhance the recognition ability of unknown categories while maintaining high-precision recognition of known categories. Despite certain attempts, effectively handling imbalanced datasets is an area that needs to be deeply explored and improved in future research.

## References


[1] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1893–1902, 2015.

[2] Florent Chiaroni, Jose Dolz, Ziko Imtiaz Masud, Amar Mitiche, and Ismail Ben Ayed. Parametric information maximization for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1729–1739, 2023.

[3] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

[4] Yixin Fei, Zhongkai Zhao, Siwei Yang, and Bingchen Zhao. Xcon: Learning with experts for fine-grained category discovery. *arXiv preprint arXiv:2208.01898*, 2022.

[5] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8401–8409, 2019.

[6] Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Improving contrastive learning on imbalanced data via open-world sampling. *Advances in Neural Information Processing Systems*, 34:5997–6009, 2021.

[7] Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2020.

[8] Mengke Li, Zhikai Hu, Yang Lu, Weichao Lan, Yiu ming Cheung, and Hui Huang. Feature fusion from head to tail for long-tailed visual recognition. In *AAAI Conference on Artificial Intelligence*, 2024.

[9] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31, 2018.

[10] Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7579–7588, 2023.

[11] Subhankar Roy, Mingxuan Liu, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Class-incremental novel class discovery. In *European Conference on Computer Vision*, pages 317–333. Springer, 2022.

[12] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7492–7501, 2022.

[13] Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16590–16600, 2023.

[14] Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Shahbaz Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3479–3488, 2023.

[15] Bingchen Zhao and Oisin Mac Aodha. Incremental generalized category discovery. *arXiv preprint arXiv:2304.14310*, 2023.

[16] Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery. *arXiv preprint arXiv:2305.06144*, 2023.