

# A Hidden Stumbling Block in Generalized Category Discovery: Distracted Attention

Qiyu Xu<sup>1,3</sup> Zhanxuan Hu<sup>1†</sup> Yu Duan<sup>2</sup> Ercheng Pei<sup>3</sup> Yonghang Tai<sup>1†</sup>

<sup>1</sup>Yunnan Normal University, <sup>2</sup>Xidian University

<sup>3</sup>Xi’an University of Posts and Telecommunications

{graceafleve, zhanxuanhu, duanyuee}@gmail.com

ercheng.pei@xupt.edu.cn, taiyonghang@126.com

## Abstract

*Generalized Category Discovery (GCD) aims to classify unlabeled data from both known and unknown categories by leveraging knowledge from labeled known categories. While existing methods have made notable progress, they often overlook a hidden stumbling block in GCD: distracted attention. Specifically, when processing unlabeled data, models tend to focus not only on key objects in the image but also on task-irrelevant background regions, leading to suboptimal feature extraction. To remove this stumbling block, we propose Attention Focusing (AF), an adaptive mechanism designed to sharpen the model’s focus by pruning non-informative tokens. AF consists of two simple yet effective components: Token Importance Measurement (TIME) and Token Adaptive Pruning (TAP), working in a cascade. TIME quantifies token importance across multiple scales, while TAP prunes non-informative tokens by utilizing the multi-scale importance scores provided by TIME. AF is a lightweight, plug-and-play module that integrates seamlessly into existing GCD methods with minimal computational overhead. When incorporated into one prominent GCD method, SimGCD, AF achieves up to 15.4% performance improvement over the baseline with minimal computational overhead. The implementation code is provided in: <https://github.com/Afleve/AFGCD>.*

## 1. Introduction

The rapid advancement of deep learning has led to significant breakthroughs in object recognition, yet many real-world applications demand more than merely classifying data into pre-defined categories. In scenarios such as autonomous driving and medical imaging, models must be capable of discovering and learning from unseen classes. Generalized Category Discovery (GCD) addresses this

<sup>†</sup>Corresponding Authors

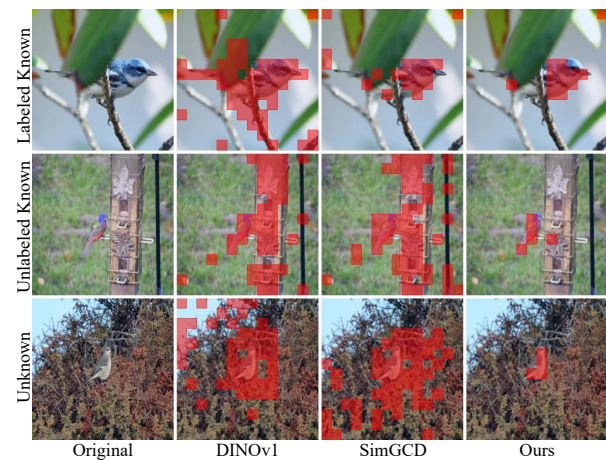


Figure 1. The masks obtained by thresholding the self-attention maps to retain 70% of the total mass. DINOv1 and SimGCD demonstrated substantial *distracted attention* on the unlabeled data, meaning it not only focuses on key objects within the image but also on task-irrelevant background regions. In contrast, our method effectively refines the model’s focus. More visualization results and analyses can be found in **Appendix C.1**.

challenge by leveraging knowledge from a set of labeled known categories to classify unlabeled data that may contain both known and unknown categories.

Most existing GCD methods follow a standardized learning paradigm: 1) employing a pre-trained Vision Transformer (ViT) as the foundational feature extraction backbone and 2) constructing task-specific GCD heads through the [CLS] token embeddings produced by the backbone. Despite notable progress, they often overlook a hidden stumbling block: *distracted attention*. Specifically, when processing unlabeled data, models tend to distribute their focus not only on key objects but also on irrelevant background regions. To investigate this, we examine one prominent GCD method, SimGCD [33], on a challenging dataset,

CUB [32]. As illustrated in Figure 1, visualization of self-attention scores in the final block of ViT shows that while the [CLS] tokens for labeled data consistently concentrate on foreground objects, those for unlabeled data, particularly from unknown categories, exhibit pronounced associations with background regions. This unintended capture of extraneous information degrades the quality of feature representations and, consequently, model performance.

We hypothesize that *distracted attention* arises partly from data augmentation. For labeled data, images within the same class often display varied backgrounds, prompting the model to concentrate on the key objects. In contrast, augmentations applied to unlabeled data typically introduce only minor variations in the background, enabling the model to exploit spurious correlations as shortcuts in unsupervised or self-supervised learning. Based on this assumption, a straightforward solution is to prune task-irrelevant tokens from the input image, ensuring that the model’s decision relies exclusively on tokens pertinent to the key object.

To this end, we propose Attention Focusing (AF), an adaptive mechanism designed to sharpen the model’s focus by pruning non-informative tokens. As shown in Figure 2, AF consists of two simple yet effective components: Token Importance Measurement (TIME) and Token Adaptive Pruning (TAP), working in a cascade. In practice, TIME introduces a task-specific query token in each ViT block to quantify token importance across multiple scales. Subsequently, TAP utilizes the multi-scale importance scores generated by TIME to prune non-informative tokens, mitigating the interference from task-irrelevant information.

Benefiting from its straightforward design, AF is a lightweight, plug-and-play module that integrates seamlessly into existing GCD methods with minimal computational overhead. In this paper, we integrate AF into SimGCD for two primary reasons. First, SimGCD employs an exceptionally simple architecture that effectively combines supervised and self-supervised learning, without introducing overly complex modules. Second, SimGCD has already demonstrated promising results across a wide range of datasets. To evaluate the effectiveness of AF, we extensively test the improved method on seven publicly available GCD datasets. The experimental results reveal that AF significantly boosts the performance of SimGCD, especially on fine-grained datasets with complex background information. Remarkably, these significant performance improvements are achieved with minimal computational overhead. This demonstrates that AF offers a highly efficient enhancement to the existing GCD framework. The main contributions of this work are summarized as follows:

1. *Novel perspective.* To the best of our knowledge, we are the first to investigate and quantify the harmful effects of *distracted attention* in GCD. This incredible finding provides a new direction toward improving this field.

2. *Novel method.* We propose AF, a simple yet effective module that provides the first generic solution for attention correction in GCD through token adaptive pruning.
3. *Promising results.* We evaluate the effectiveness and efficiency of AF across different settings. Experimental results demonstrate that AF can significantly improve performance with minimal computational overhead.

## 2. Related Work

### 2.1. Generalized Category Discovery

GCD extends the paradigms of Semi-Supervised Learning (SSL) [8, 16] and Novel Category Discovery (NCD) [7], which leverages knowledge of known categories within open-world settings to simultaneously identify both known and unknown classes from unannotated data. Most existing GCD methods can be broadly categorized into: 1) non-parametric methods; and 2) parametric methods.

Non-parametric methods [5, 23, 24, 29, 35, 38] typically involve training a feature extractor followed by the application of clustering techniques, such as semi-supervised K-means++ [29], to obtain the final classification results. For example, GCD [29] introduces a fundamental framework that utilizes traditional supervised and unsupervised contrastive learning to achieve effective representation learning. Similarly, DCCL [23] optimizes instance-level and concept-level contrastive objectives through dynamic conception generation and dual-level contrastive learning, exploiting latent relationships among unlabeled samples. Furthermore, GPC [37] integrates a Gaussian Mixture Model within an Expectation-Maximization framework to alternate between representation learning and category estimation, SelEx [25] introduces ‘self-expertise’ to enhance the model’s ability to recognize subtle differences. In addition, PromptCAL [35] utilizes visual prompt tuning to facilitate contrastive affinity learning within a two-stage framework, while CMS [5] incorporates Mean Shift clustering into the contrastive learning process to encourage tighter grouping of similar samples.

Parametric methods [30, 31, 33] integrate the optimization of a parametric classifier to directly yield prediction outcomes. For instance, SimGCD [33] jointly trains a prototype classifier alongside representation learning, establishing a robust baseline for this category of methods. SPT-Net [31] employs a two-stage framework that alternates between model refinement and prompt learning. Moreover, GET [30] leverages CLIP to generate semantic prompts for novel classes via text generation, thereby unlocking the potential of multimodal models for addressing the GCD task.

Indeed, most existing GCD methods primarily focus on how to leverage unsupervised or self-supervised learning techniques to enhance model performance on unlabeled data. Despite notable progress, they often overlook a hidden

stumbling block: *distracted attention*. Addressing this challenge is the core of this paper. It is worth noting that during the review process, we identified two representative works that also aim to mitigate background interference [21, 36]. Nevertheless, our method differs fundamentally in both its underlying motivation and methodological design.

## 2.2. High-Resolution Image Recognition

High-resolution recognition refers to the capability of computer vision systems to accurately identify and analyze objects in images characterized by a high pixel count and intricate details. Managing *distracted attention* is a critical challenge in this context, as the extensive spatial information often leads to inefficient feature extraction and model focus drift. A widely adopted strategy to address this issue is to partition high-resolution images into smaller patches, thereby increasing the relative proportion of key targets within each patch. For instance, IPS [1] iteratively processes individual patches and selectively retains those most relevant to the specific task. SPHINX [34] segments a high-resolution image into a set of low-resolution images and concatenates these with a downsampled version of the original image as the visual input. Monkey [17] employs a sliding window approach combined with a visual resampling mechanism to enhance image resolution, thereby improving content comprehension while reducing computational overhead. Furthermore, LLaVA-UHD [9] ensures both efficiency and fidelity in image processing by optimizing slice computation and scoring functions, effectively minimizing resolution variations. On one hand, these methods are specifically designed for supervised learning scenarios and cannot be directly applied to GCD tasks without significant modifications. On the other hand, we process the original images directly, achieving greater efficiency while preserving accuracy.

## 2.3. Token Pruning

Another issue closely related to this work is token pruning, which aims to enhance computational efficiency and reduce redundancy by selectively removing task-irrelevant patches while preserving most of the original image information. EVit [18] leverages the attention values between the [CLS] token and patch tokens in ViT to select the most informative patches. SPVit [12] and SVit [19] propose retaining pruned tokens from upper layers for subsequent use, rather than discarding them entirely. PS-ViT (T2T) [28] adopts a reverse approach by selecting tokens for pruning based on the final output features. ToMe [3] reduces the computational workload by merging tokens with high key similarity. While these methods have achieved notable advancements in improving inference efficiency, they often result in varying degrees of performance degradation. In the context of the GCD task, however, model accuracy is of

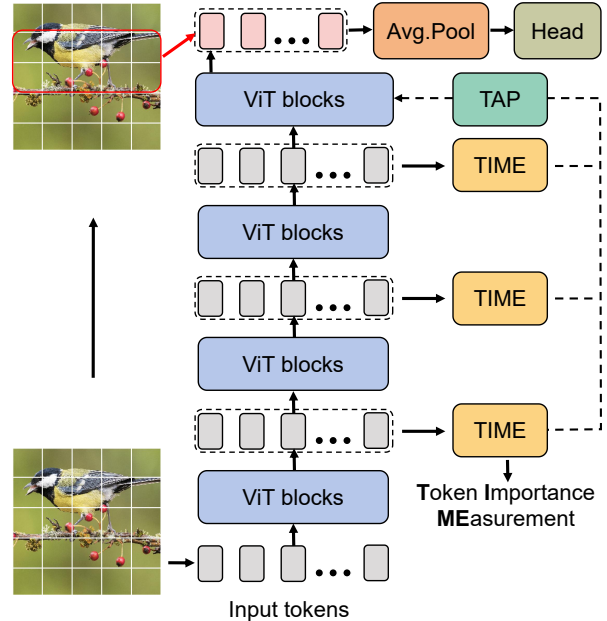


Figure 2. The pipeline of GCD with our proposed *Attention Focusing (AF)* mechanism. AF consists of two components: Token Importance Measurement (TIME) and Token Adaptive Pruning (TAP), working in a cascade. Here, the 'Head' can be inherited from any existing GCD model.

paramount importance. Additionally, many methods rely on the [CLS] token for pruning, but in the GCD task, the [CLS] token for unlabeled data tends to be of lower quality, making it susceptible to introducing misleading information (see **Appendix C.3**). The method most relevant to ours is Cropr [2], which prunes a fixed number of tokens at each ViT block. However, we adopted multi-scale adaptive pruning to address the diversity of image backgrounds, achieving better results (see Section 4.3).

## 3. Method

### 3.1. Problem Formulation

Generalized Category Discovery (GCD) addresses the problem of automatically clustering unlabeled data  $\mathcal{D}^u = \{(x_i, y_i^u) \in \mathcal{X} \times \mathcal{Y}_u\}$  in a partially labeled dataset  $\mathcal{D}^l = \{(x_i, y_i^l) \in \mathcal{X} \times \mathcal{Y}_l\}$ . Here,  $\mathcal{Y}_l$  represents the set of known classes, and  $\mathcal{Y}_u$  represents the set of all classes, with  $\mathcal{Y}_l \subset \mathcal{Y}_u$ . In different GCD approaches, the number of unknown classes  $|\mathcal{Y}_u|$  can be utilized as prior knowledge or estimated through established methods.

### 3.2. Overview

The currently popular GCD methods are primarily based on pre-trained ViT models. Specifically, given an image  $I \in \mathbb{R}^{h \times w \times 3}$ , ViT divides it into a sequence of non-overlapping patches, each of size  $P \times P$ . This sequence of

patches is then flattened and mapped into token embeddings  $\{\mathbf{x}_n \in \mathbb{R}^{1 \times D}, n = 1, 2, 3, \dots, N\}$  through a linear projection head, where  $N = H \times W, H = h/P, W = w/P$ , and  $D$  represents the dimensionality of the embedding space. After appending an additional [CLS] token to the patch tokens, the resulting token sequence  $\mathbf{X} \in \mathbb{R}^{(N+1) \times D}$  is passed sequentially through all transformer blocks. For simplicity, the batch size  $B$  and block number  $l$  are omitted from the description. Ultimately, the [CLS] token produced by the backbone network is passed into the task-specific GCD head. As illustrated in Figure 1, while the [CLS] tokens for labeled data consistently focus on foreground objects, those for unlabeled data, especially from unknown categories, show strong associations with background regions. This unintended capture of extraneous information degrades the quality of feature representations and, consequently, the performance of the GCD model.

To this end, we propose integrating a novel AF mechanism into the existing GCD model. As illustrated in Figure 2, the AF mechanism consists of two simple yet effective components: Token Importance Measurement (TIME) (Section 3.3) and Token Adaptive Pruning (TAP) (Section 3.4), which operate in a cascade. In practice, the TIME module is inserted into every block of the ViT, except for the last one. Each TIME module outputs a score vector that reflects the importance of each patch token. The TAP module then aggregates these multi-scale scores to prune the non-informative tokens. Finally, the remaining tokens are processed with average pooling and then used as input to the *Head*. It is important to note that the *Head* can be inherited from any existing GCD method. In this work, our primary experiment is based on SimGCD [33], a representative GCD method. Additionally, we integrate the AF mechanism into three representative methods, CMS [5], GET [30], and SelEx [25], to demonstrate its generalizability (see Section 4.3). Next, we will provide a detailed description of TIME and TAP, while further details on SimGCD can be found in the **Appendix A**.

### 3.3. Token Importance Measurement

As shown in Figure 3, TIME is trained exclusively on labeled data but is capable of generalizing to the entire training set. Specifically, given an image, TIME takes its tokens as input and produces a score vector  $\mathbf{s} \in \mathbb{R}^{1 \times (N+1)}$ , revealing the informativeness of the input tokens. Specifically, each TIME module consists of three key components: a *Measurer*, an *Aggregator*, and an *Auxiliary classifier*.

The *Measurer* assigns the score vector  $\mathbf{s} \in \mathbb{R}^{1 \times (N+1)}$  to each token by performing cross-attention between the tokens and a learnable query vector  $\mathbf{Q}$ . Specifically, the input tokens  $\mathbf{X}$  are treated as the key matrix  $\mathbf{K}$  and value matrix  $\mathbf{V}$ . The query vector  $\mathbf{Q}$  is then used to query  $\mathbf{K}$ , yielding attention results for each token. The scores between the query

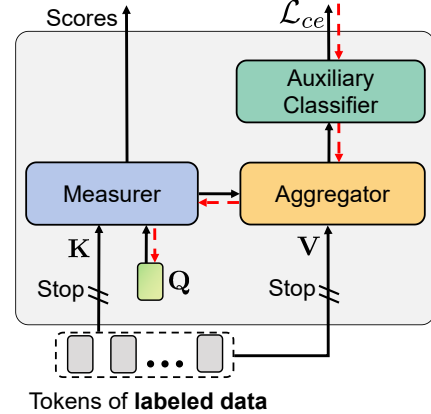


Figure 3. The internal pipeline of TIME. The red dashed lines represent the gradient propagation paths from the auxiliary classifier to the optimization of  $\mathbf{Q}$ . Besides, TIME is trained using only labeled data, but it works on both labeled and unlabeled data.

vector and the key matrix are computed as follows:

$$\mathbf{s}(\mathbf{Q}, \mathbf{K}) = \frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D}}, \quad (1)$$

where  $\sqrt{D}$  is a scaling factor to stabilize the attention values. To ensure the informativeness scores  $\mathbf{s}$  are properly utilized, the *Aggregator* leverages these scores to obtain an initial image representation. Specifically, the aggregated representation  $\mathbf{r}$  is computed as:

$$\mathbf{r} = \text{Softmax}(\mathbf{s})\mathbf{V}. \quad (2)$$

Furthermore, to increase the capacity of the *Aggregator*, we follow [2] and incorporate a transformer block’s Feed-Forward Network (FFN), which includes LayerNorm (LN) and an MLP with a residual connection. Mathematically,

$$\mathbf{r}' = \text{MLP}(\text{LayerNorm}(\mathbf{r})) + \mathbf{r}. \quad (3)$$

Next, the resulting representation  $\mathbf{r}'$  is passed through the *Auxiliary classifier*, producing a probability output  $\mathbf{p} \in \mathbb{R}^{1 \times |\mathcal{Y}_l|}$ , where  $|\mathcal{Y}_l|$  is the number of possible classes for labeled data. TIME is trained using a cross-entropy loss:

$$\mathcal{L}_{ce} = - \sum_{k=1}^{|\mathcal{Y}_l|} y^k \log p^k, \quad (4)$$

where  $y^k$  represents the ground truth label and  $p^k$  is the predicted probability.

In practice, the *Auxiliary classifier* aids in classifying labeled data, guiding the *Aggregator* to focus on the most informative features of the image that are crucial for classification. As training progresses, the query vector  $\mathbf{Q}$  dynamically adjusts the score vector  $\mathbf{s}$ , assigning progressively

higher importance to tokens with greater informativeness. This adaptive mechanism enables the model to prioritize the most relevant tokens for the task, improving its ability to capture critical information for accurate classification. Generally, unlabeled data and labeled data often share similar stylistic characteristics. Therefore, we hypothesize that the query vector  $\mathbf{Q}$ , learned from labeled data, generalizes well and can effectively assess the importance of patch tokens even in the case of unlabeled data.

Additionally, we apply a stop-gradient to isolate the *Auxiliary classifier* from the backbone, ensuring that conflicting gradients do not affect the encoder. During testing, the *Auxiliary classifier* is discarded, and only the query vector  $\mathbf{Q}$  is retained to process the test samples. This reduces computational overhead while maintaining the model’s capacity to evaluate token importance effectively.

### 3.4. Token Adaptive Pruning

The score vectors obtained from different TIME blocks represent the importance of patch tokens across different scales. TAP leverages these multi-scale importance scores to prune the input patch tokens. Specifically, given a set of score vectors  $\{\hat{\mathbf{s}}_l \in \mathbb{R}^{1 \times (N+1)}\}_{l=1}^{L-1}$ , where  $L$  denotes the number of ViT blocks, the multi-scale importance of patch tokens is computed as follows:

$$\mathbf{s}^m = \frac{1}{L-1} \sum_{l=1}^{L-1} \text{Softmax}(\hat{\mathbf{s}}_l), \quad (5)$$

where  $\hat{\mathbf{s}}_l \in \mathbb{R}^{1 \times N}$  represents a score vector that excludes the value associated with the [CLS] token. This exclusion is crucial because the [CLS] token aggregates high-level semantic information about the image, making it a meaningful token in itself. Next, for the patch tokens  $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$ , we prune the less informative tokens by applying an adaptive threshold  $\tau$ . Formally, we define the pruned patch tokens  $\mathbf{X}_p$  as:

$$\mathbf{X}_p = \{\mathbf{x}_i \mid i = 1, 2, \dots, t, \sum_{i=1}^N s_i^m \leq \tau\}, \quad (6)$$

where  $s_i^m$  is the  $i$ -th element of the multi-scale importance score vector  $\mathbf{s}^m$ , and the indices  $i = 1, 2, \dots, N$  are sorted in increasing order of  $s_i^m$ . The pruned patch tokens  $\mathbf{X}_p$  represent redundant information associated with task-irrelevant regions in the image. The remaining token sequence,  $\mathbf{X}_r$ , consisting of the residual patch tokens and the [CLS] token, is then passed through the final ViT block. Finally, the output token representations are processed using average pooling to form the final image representation, which is subsequently input into the GCD *Head*. The overall loss

function of our improved method is:

$$\mathcal{L} = \mathcal{L}_{gcd} + \lambda \sum_{l=1}^{L-1} \mathcal{L}_{ce}^l, \quad (7)$$

where  $\mathcal{L}_{gcd}$  denotes the loss function of the selected GCD baseline model,  $\lambda$  is a balancing parameter.

### 3.5. Discussion

During the training process of GCD, each instance is typically augmented with two distinct views, raising an important question: *Should we adopt single-view TAP or multi-view TAP?* The former applies TAP to only one of these views, while the latter applies TAP to both augmented views simultaneously. In this work, we opt for single-view TAP for two main reasons. First, TAP can be seen as a form of non-regular image cropping augmentation, where single-view TAP is particularly effective in helping the model focus on key objects of interest. By pruning unnecessary tokens in a single view, the model can retain critical information, improving its ability to extract meaningful features from the complex image. Second, multi-view TAP effectively forces the model to train without the interference of background information across both views. Although this may appear beneficial in theory by reducing noise, it can inadvertently hinder the model’s ability to generalize (as shown in [Appendix C.2](#)).

## 4. Experiments

### 4.1. Experimental Setup

**Dataset.** In this study, we primarily incorporate AF into SimGCD [33] and evaluate the effectiveness using three challenging fine-grained datasets from the Semantic Shift Benchmark [26]: CUB [32], Stanford Cars [13], and FGVC-Aircraft [20]. Additionally, we apply our method to three more generic classification datasets, namely CIFAR10/100 [14] and ImageNet-100 [6], as well as the large-scale fine-grained dataset Herbarium-19 [27]. As discussed in the [Appendix B.1](#), the former often includes complex background information, while the latter exhibits relatively minimal background interference. To ensure the fairness of the experiments, all other settings are kept consistent with SimGCD. More details can be found in the [Appendix A](#).

**Evaluation.** Following established practice [33], we utilize clustering accuracy (ACC) to evaluate the model performance. Prior to comparing the ground truth with the predicted labels, we employ the Hungarian algorithm [15] to align the labels of the *Unknown* category, followed by calculating the accuracy (ACC) using  $\frac{1}{M} \sum_{i=1}^M \mathbb{1}(y_i^* = p(\hat{y}_i))$  where  $M = |D_U|$ , and  $p$  denotes the optimal permutation.

For clarity and convenience, the accuracy metrics are reported for 'All' unlabeled data, along with the subsets cor-

responding to known and unknown classes, labeled as 'Old' and 'New' in the tables, respectively.

## 4.2. Main Results

**Evaluation on challenging fine-grained datasets.** Table 1 presents a comparison between SimGCD and several state-of-the-art methods on three challenging fine-grained datasets, where ' $\Delta$ ' denotes the performance improvements over the baseline model, SimGCD. Clearly, SimGCD serves as a robust baseline model, achieving competitive results in the vast majority of settings, despite its simple network architecture. Comparing with SimGCD+AF, we observe that the AF module significantly enhances the model's performance, underscoring its effectiveness in addressing the *distracted attention* issue in SimGCD. Compared to other state-of-the-art methods, SimGCD+AF consistently achieves the best or near-best performance across various datasets. On the CUB dataset, the performance of InfoSieve and CMS is comparable to that of SimGCD+AF. However, SimGCD+AF demonstrates a clear advantage on the other two datasets, particularly on Stanford Cars, where the performance improvement on 'All' reaches up to 10.1%. While SPTNet and SimGCD+AF perform similarly on FGVC-Aircraft, SPTNet's performance on Stanford Cars is notably weaker than that of SimGCD+AF. Additionally, SPTNet employs an alternating training strategy, resulting in a higher computational cost compared to SimGCD+AF. Both MOS and AptGCD also focus on mitigating the interference of background information and achieve results comparable to SimGCD+AF. However, AF is relatively simpler in module design and does not rely on any external models.

Datasets	CUB			Stanford Cars			FGVC-Aircraft		
	All	Old	New	All	Old	New	All	Old	New
RankStats [11]	33.3	51.6	24.2	28.3	61.8	12.1	26.9	36.4	22.2
UNO+ [7]	35.1	49.0	28.1	35.5	70.5	18.6	40.3	56.4	32.2
ORCA [10]	35.3	45.6	30.2	23.5	50.1	10.7	22.0	31.8	17.1
GCD [29]	51.3	56.6	48.7	39.0	57.6	29.9	45.0	41.1	46.9
DCCL [22]	63.5	60.8	64.9	43.1	55.7	36.2	-	-	-
GPC [38]	55.4	58.2	53.1	42.8	59.2	32.8	46.3	42.5	47.9
PIM [4]	62.7	75.7	56.2	43.1	66.9	31.6	-	-	-
InfoSieve [24]	69.4	<b>77.9</b>	65.2	55.7	74.8	46.4	56.3	63.7	52.5
CMS [5]	68.2	<u>76.5</u>	64.0	56.9	76.1	47.6	56.0	63.4	52.3
SPTNet [31]	65.8	68.8	65.1	59.0	79.2	49.3	59.3	61.8	58.1
AptGCD [36]	<b>70.3</b>	74.3	<b>69.2</b>	62.1	79.7	53.6	<b>61.1</b>	65.2	<b>59.0</b>
MOS [21]	<u>69.6</u>	72.3	<u>68.2</u>	<u>64.6</u>	<b>80.9</b>	<u>56.7</u>	<b>61.1</b>	<u>66.9</u>	<u>58.2</u>
SimGCD [33]	60.3	65.6	57.7	53.8	71.9	45.0	54.2	59.1	51.8
SimGCD+AF	69.0	74.3	66.3	<b>67.0</b>	<u>80.7</u>	<b>60.4</b>	<u>59.4</u>	<b>68.1</b>	55.0
$\Delta$	<b>+8.7</b>	<b>+8.7</b>	<b>+8.6</b>	<b>+13.2</b>	<b>+8.8</b>	<b>+15.4</b>	<b>+5.2</b>	<b>+9.0</b>	<b>+3.2</b>

Table 1. Comparison with several state-of-the-art methods on fine-grained datasets. The best results are highlighted in **bold**, and the second-best results are highlighted in underline. ' $\Delta$ ' refers to the performance improvement compared to SimGCD [33].

Datasets	CIFAR10			CIFAR100			ImageNet-100		
	All	Old	New	All	Old	New	All	Old	New
RankStats [11]	46.8	19.2	60.5	58.2	77.6	19.3	37.1	61.6	24.8
UNO+ [7]	68.6	<b>98.3</b>	53.8	69.5	80.6	47.2	70.3	95.0	57.9
ORCA [10]	96.9	95.1	97.8	69.0	77.4	52.0	73.5	92.6	63.9
GCD [29]	91.5	97.9	88.2	73.0	76.2	66.5	74.1	89.8	66.3
DCCL [22]	96.3	96.5	96.9	75.3	76.8	70.2	80.5	90.5	76.2
GPC [38]	92.2	<u>98.2</u>	89.1	77.9	<u>85.0</u>	63.0	76.9	94.3	71.0
PIM [4]	94.7	97.4	93.3	78.3	84.2	66.5	83.1	95.3	77.0
InfoSieve [24]	94.8	97.7	93.4	78.3	82.2	70.5	80.5	93.8	73.8
CMS [5]	-	-	-	<u>82.3</u>	<b>85.7</b>	75.5	84.7	<b>95.6</b>	79.2
SPTNet [31]	<u>97.3</u>	95.0	98.6	81.3	84.3	75.6	85.4	93.2	81.4
AptGCD [36]	<u>97.3</u>	95.8	<u>98.7</u>	<b>82.8</b>	81.8	<b>85.5</b>	<b>87.8</b>	<u>95.4</u>	<b>84.3</b>
SimGCD [33]	97.1	95.1	98.1	80.1	81.2	<u>77.8</u>	83.0	93.1	77.9
SimGCD+AF	<b>97.8</b>	95.9	<b>98.8</b>	82.2	85.0	76.5	<u>85.4</u>	94.6	80.8
$\Delta$	<b>+0.7</b>	<b>+0.8</b>	<b>+0.7</b>	<b>+2.1</b>	<b>+3.8</b>	<b>-1.3</b>	<b>+2.4</b>	<b>+1.5</b>	<b>+2.9</b>

Table 2. Comparison with several state-of-the-art methods on three generic datasets.

Datasets	Herbarium-19		
	All	Old	New
GCD [29]	35.4	51.0	27.0
PIM [4]	42.3	56.1	34.8
InfoSieve [24]	41.0	55.4	33.2
CMS [5]	36.4	54.9	26.4
SPTNet [31]	43.4	<u>58.7</u>	35.2
SimGCD [33]	<u>44.0</u>	58.0	<u>36.4</u>
SimGCD+AF	<b>45.5</b>	<b>59.0</b>	<b>38.3</b>
$\Delta$	<b>+1.5</b>	<b>+1.0</b>	<b>+1.9</b>

Table 3. Comparison with several state-of-the-art methods on Herbarium-19.

**Evaluation on generic datasets.** Table 2 presents the results on generic datasets. We observed that the improvement brought by AF on these datasets is less pronounced than on the fine-grained datasets. We attribute this to two main factors. First, the SimGCD model has already achieved excellent performance on these datasets, such as nearly 100% accuracy on CIFAR-10. Second, the backgrounds of these datasets are relatively simple, leading to minimal interference. For example, on CIFAR-100, due to the lack of complex backgrounds, AF even resulted in a performance decrease for the new classes. In contrast, for ImageNet100, a dataset with more complex backgrounds, AF provided a more noticeable performance improvement. Compared to other methods, SimGCD+AF also achieves competitive results, but it typically involves lower computational cost.

**Evaluation on more challenging datasets.** Compared to the above three fine-grained datasets, Herbarium-19 has a simpler background, and as a result, the performance gain brought by AF is also relatively limited. This highlights a limitation of our method AF: while it effectively suppresses interference from background information, it does not significantly improve the model's ability to extract information from the key objects themselves.

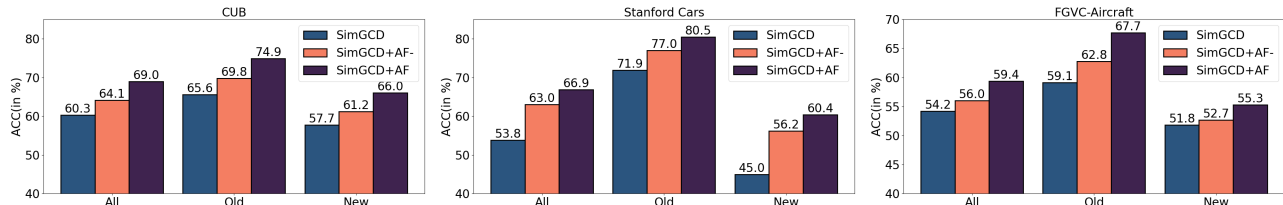


Figure 4. Investigation of Multi-scale token importance measurement. "SimGCD+AF-" refers to a setting where only the query from the penultimate block is used as the basis for token pruning within TAP.



Figure 5. The results of token pruning using query vectors from each layer. Specifically, the last column illustrates the multi-scale token importance measurement used in AF.

### 4.3. Discussion on the design of AF

**Is AF effective for other GCD models?** As mentioned above, AF is a plug-and-play module that can be seamlessly integrated into existing GCD methods without requiring extensive modifications. To further assess the generalizability and effectiveness of AF, we incorporated it into three additional GCD methods, CMS [5], SelEx [25], and GET [30]. The results, as displayed in the Table 4, reveal a substantial improvement in performance across various datasets, with particularly notable enhancements observed in the Stanford Cars and FGVC-Aircraft datasets. These findings provide strong evidence of AF's ability to significantly boost the performance of baseline models, highlighting its broad applicability and compatibility with different GCD approaches.

**Is multi-scale token importance measurement necessary?** In this work, TAP prunes less informative tokens

Datasets	CUB			Stanford Cars			FGVC-Aircraft		
	All	Old	New	All	Old	New	All	Old	New
CMS	67.3	75.6	63.1	53.1	73.0	43.5	54.2	63.2	49.8
<b>CMS+AF</b>	68.2	75.9	64.3	61.8	76.3	54.8	57.5	62.7	54.9
SelEx	73.4	73.9	73.2	58.9	78.6	49.4	57.2	66.3	52.6
<b>SelEx+AF</b>	79.2	76.3	80.6	61.2	80.1	52.0	62.8	66.5	60.9
GET	75.2	77.9	73.9	78.3	86.0	74.6	57.4	59.6	54.7
<b>GET+AF</b>	77.3	77.1	77.4	81.5	90.6	77.1	59.5	67.0	55.8

Table 4. Results of incorporating AF into three additional methods: CMS [5], SelEx [25] GET [30]. Notably, CMS did not perform mean shift clustering during testing.

by aggregating importance scores across multiple scales. Figure 5 illustrates the selected patches at different ViT blocks. As shown, the patches selected by the model vary significantly across different layers, primarily due to the differences in the feature scales at each layer. This vari-

ability underscores the need for a multi-scale approach, as it enables the model to capture a broader range of key object information, leading to a more robust and comprehensive understanding of the image. Besides, we explored using only the query from the penultimate block as the basis for token pruning in TAP. While this approach still results in some performance improvements for the baseline model SimGCD, as depicted in the Figure 4, the model’s performance degrades substantially when compared to SimGCD+AF. This result highlights the necessity of integrating multi-scale token importance measurement.

**Learn queries from only labeled data or all training data?** To empower the *queries* with the capability of selectively attending to informative image tokens, the learnable *queries* in AF are exclusively trained on labeled data. This design choice is motivated by two critical considerations. First, in the absence of supervisory signals, the model struggles to accurately identify and focus on the true key objects within unlabeled images, as the background clutter and irrelevant regions may dominate the feature representation. Second, and more importantly, the self-distillation loss, which is commonly employed in unlabeled data, can inadvertently introduce noise and bias into the learning process of *queries*, thereby deteriorating their ability to distinguish between informative and non-informative patches. This phenomenon is empirically validated in Table 5, where we observe that training the *queries* on the entire dataset (including both labeled and unlabeled samples) results in a substantial performance drop across all benchmarks. This degradation underscores the importance of leveraging clean, supervised signals for learning robust and discriminative *queries* that can effectively guide the model’s attention towards task-relevant tokens.

Datasets	CUB			Stanford Cars			FGVC-Aircraft		
	All	Old	New	All	Old	New	All	Old	New
SimGCD	60.1	69.7	55.4	55.7	73.3	47.1	53.7	64.8	48.2
+AF(all)	67.4	73.9	64.1	63.0	81.5	54.1	54.6	60.5	51.6
<b>AF</b>	69.0	74.3	66.3	67.0	80.7	60.4	59.4	68.1	55.0

Table 5. Investigation of *Query learning*. ‘AF(all)’ refers to a setting where *Query learning* is based on the entire training dataset.

**How important is token adaptive pruning?** Considering the inherent variability in background information across different images, we adopt a token-adaptive pruning strategy in TAP instead of employing a fixed pruning approach. To demonstrate the superiority of TAP, we conduct a comparative experiment using fixed pruning, where a predetermined number of  $k$  patches are uniformly removed from training images. As illustrated in Table 6, while the model’s performance exhibits some improvement as the number of removed patches increases within a limited range, it consistently falls short of the performance achieved by TAP.

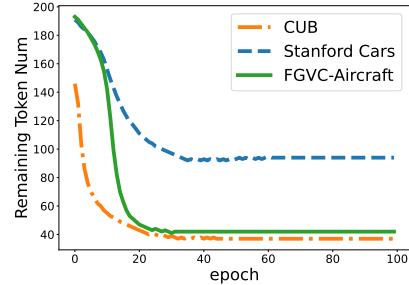


Figure 6. The dynamic change of the number of retaining patches during the training process.

Notably, when  $K = 128$ , the model’s performance on the Stanford Cars degrades compared to  $K = 64$ , likely due to the excessive removal of informative patches, which undermines the model’s ability to capture essential features. This observation is further corroborated by Figure 6, which reveals that TAP retains a higher proportion of patches on the Stanford Cars dataset compared to CUB and FGVC-Aircraft. These findings underscore the importance of a dynamic, image-specific pruning strategy, as implemented in TAP, to effectively balance the removal of non-informative background patches while preserving critical visual information.

Datasets	CUB			Stanford Cars			FGVC-Aircraft		
	All	Old	New	All	Old	New	All	Old	New
SimGCD	60.1	69.7	55.4	55.7	73.3	47.1	53.7	64.8	48.2
$k = 16$	65.1	74.1	60.5	60.4	75.2	53.3	54.1	64.7	48.8
$k = 64$	67.1	72.3	64.5	63.5	79.8	55.6	54.3	61.3	50.7
$k = 128$	67.0	75.0	63.0	62.4	82.8	52.6	55.5	64.9	50.7
<b>TAP</b>	69.0	74.3	66.3	67.0	80.7	60.4	59.4	68.1	55.0

Table 6. Investigation of *Token Adaptive Pruning*. ‘ $k$ ’ refers to a setting where a predetermined number of  $k$  patches are uniformly removed from training images.

## 5. Conclusion

In this work, we introduced AF, a simple yet powerful mechanism designed to address the issue of distracted attention in GCD. By pruning non-informative tokens, AF refines the model’s focus on the key objects in the image, resulting in enhanced performance across both known and unknown categories. Extensive experiments show that when integrated with existing GCD methods, such as SimGCD, AF leads to substantial performance gains while maintaining minimal computational overhead. However, while AF effectively mitigates background interference, it does not significantly improve the model’s ability to extract more discriminative features from the key objects themselves. This limitation points to an avenue for future research: developing methods that can further enhance the model’s ability to focus on the most relevant features of the key objects.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62201453), the Basic Research Project of Yunnan Province (No.202501CF070004), and the Xingdian Talent Support Program.

## References

- [1] Benjamin Bergner, Christoph Lippert, and Aravindh Mahendran. Iterative patch selection for high-resolution image recognition. In *International Conference on Learning Representations*, 2022. 3
- [2] Benjamin Bergner, Christoph Lippert, and Aravindh Mahendran. Token crop: Faster vits for quite a few tasks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9740–9750, 2025. 3, 4
- [3] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023. 3
- [4] 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. 6
- [5] Sua Choi, Dahyun Kang, and Minsu Cho. Contrastive mean-shift learning for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 4, 6, 7
- [6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 5
- [7] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9284–9292, 2021. 2, 6
- [8] Enrico Fini, Pietro Astolfi, Karteek Alahari, Xavier Alameda-Pineda, Julien Mairal, Moin Nabi, and Elisa Ricci. Semi-supervised learning made simple with self-supervised clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3187–3197, 2023. 2
- [9] Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. LLaVA-UHD: an lmm perceiving any aspect ratio and high-resolution images. In *ECCV*, 2024. 3
- [10] 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
- [11] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel: Automatically discovering and learning novel visual categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6767–6781, 2021. 6
- [12] Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Wei Niu, Mengshu Sun, Xuan Shen, Geng Yuan, Bin Ren, Hao Tang, et al. Spvit: Enabling faster vision transformers via latency-aware soft token pruning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XI*, pages 620–640. Springer, 2022. 3
- [13] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 554–561, 2013. 5
- [14] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5
- [15] Harold W Kuhn. The hungarian method for the assignment problem. In *Naval research logistics quarterly*, 1955. 5
- [16] Junnan Li, Caiming Xiong, and Steven CH Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9475–9484, 2021. 2
- [17] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 3
- [18] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations*, 2022. 3
- [19] Yifei Liu, Mathias Gehrig, Nico Messikommer, Marco Cannici, and Davide Scaramuzza. Revisiting token pruning for object detection and instance segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. 3
- [20] Subhansu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5
- [21] Zhengyuan Peng, Jinpeng Ma, Zhimin Sun, Ran Yi, Haichuan Song, Xin Tan, and Lizhuang Ma. Mos: Modeling object-scene associations in generalized category discovery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15118–15128, 2025. 3, 6
- [22] Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptual contrastive learning for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7579–7588, 2023. 6
- [23] Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptual contrastive learning for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7579–7588, 2023. 2
- [24] Sarah Rastegar, Hazel Doughty, and Cees Snoek. Learn to categorize or categorize to learn? self-coding for generalized category discovery. In *Advances in Neural Information Processing Systems*, 2023. 2, 6

- [25] Sarah Rastegar, Mohammadreza Salehi, Yuki M Asano, Hazel Doughty, and Cees GM Snoek. Selex: Self-expertise in fine-grained generalized category discovery. In *European Conference on Computer Vision*, pages 440–458. Springer, 2024. [2](#), [4](#), [7](#)
- [26] Andrea Vedaldi Sagar Vaze, Kai Han and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? In *arXiv preprint arXiv:2110.06207*, 2021. [5](#)
- [27] Kiat Chuan Tan, Yulong Liu, Barbara Ambrose, Melissa Tulig, and Serge Belongie. The herbarium challenge 2019 dataset. *arXiv preprint arXiv:1906.05372*, 2019. [5](#)
- [28] Yehui Tang, Kai Han, Yunhe Wang, Chang Xu, Jianyuan Guo, Chao Xu, and Dacheng Tao. Patch slimming for efficient vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12165–12174, 2022. [3](#)
- [29] 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. [2](#), [6](#)
- [30] Enguang Wang, Zhimao Peng, Zhengyuan Xie, Fei Yang, Xialei Liu, and Ming-Ming Cheng. Get: Unlocking the multi-modal potential of clip for generalized category discovery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20296–20306, 2025. [2](#), [4](#), [7](#)
- [31] Hongjun Wang, Sagar Vaze, and Kai Han. Sptnet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. In *International Conference on Learning Representations (ICLR)*, 2024. [2](#), [6](#)
- [32] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010. [2](#), [5](#)
- [33] 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. [1](#), [2](#), [4](#), [5](#), [6](#)
- [34] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient finetuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. [3](#)
- [35] 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. [2](#)
- [36] Wei Zhang, Baopeng Zhang, Zhu Teng, Wenxin Luo, Junnan Zou, and Jianping Fan. Less attention is more: Prompt transformer for generalized category discovery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30322–30331, 2025. [3](#), [6](#)
- [37] Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16623–16633, 2023. [2](#)
- [38] Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16623–16633, 2023. [2](#), [6](#)