GET: Unlocking the Multi-modal Potential of CLIP for Generalized Category Discovery

Enguang Wang[1], Zhimao Peng[1], Zhengyuan Xie[1], Fei Yang[2,1]*, Xialei Liu[2,1], Ming-Ming Cheng[2,1]

[1]VCIP, CS, Nankai University [2]NKIARI, Shenzhen Futian

{enguangwang,zhimao796,xiezhengyuan}@mail.nankai.edu.cn

{feiyang,xialei,cmm}@nankai.edu.cn

Abstract

Given unlabelled datasets containing both old and new categories, generalized category discovery (GCD) aims to accurately discover new classes while correctly classifying old classes. Current GCD methods only use a single visual modality of information, resulting in a poor classification of visually similar classes. As a different modality, text information can provide complementary discriminative information, which motivates us to introduce it into the GCD task. However, the lack of class names for unlabelled data makes it impractical to utilize text information. To tackle this challenging problem, in this paper, we propose a Text Embedding Synthesizer (TES) to generate pseudo text embeddings for unlabelled samples. Specifically, our TES leverages the property that CLIP can generate aligned vision-language features, converting visual embeddings into tokens of the CLIP's text encoder to generate pseudo text embeddings. Besides, we employ a dual-branch framework, through the joint learning and instance consistency of different modality branches, visual and semantic information mutually enhance each other, promoting the interaction and fusion of visual and text knowledge. Our method unlocks the multi-modal potentials of CLIP and outperforms the baseline methods by a large margin on all GCD benchmarks, achieving new state-of-the-art. Our code is available at: https://github.com/enguangW/GET

1. Introduction

Deep neural networks trained on large amounts of labeled data have shown powerful visual recognition capabilities [23]. Although this is heartening, the close-set assumption severely hinders the deployment of the model in practical application scenarios. Recently, novel class discovery (NCD) [15] has been proposed to categorize unknown classes of unlabelled data, leveraging knowledge learned from labeled data. As a realistic extension to NCD, generalized category discovery (GCD) [41] assumes that the unlabelled data come from both known and unknown classes, rather than just unknown classes as in NCD. The model needs to accurately discover unknown classes while correctly classifying known classes of the unlabelled data, breaking the close-set limitation, making GCD a challenging and meaningful task.

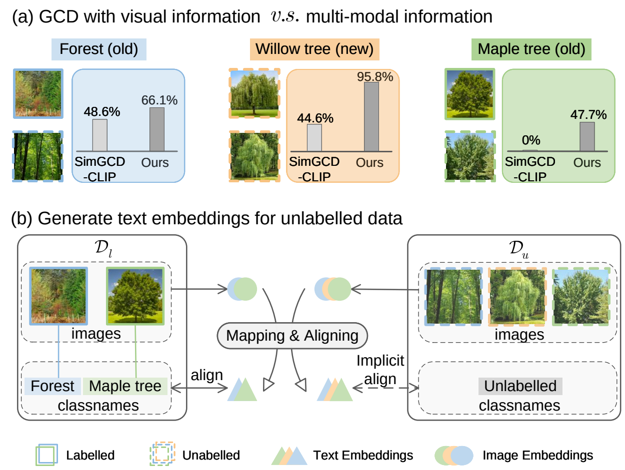

Figure 1. The motivation of our method. (a) Current GCD methods [45] rely on single visual features, resulting in poor classification of visually similar classes; Our approach introduces text information, improving the discriminative capabilities of the model. (b) Our proposed method maps image embeddings to text embeddings while simultaneously achieving modal alignment.

Previous GCD methods [35, 41, 45, 48, 51] utilize a DINO [5] pre-trained ViT as the backbone network to expect good initial discrimination ability of the model, thereby facilitating fine-tuning on the training data. Although promising results have been achieved, these representations derived from a single visual backbone often struggle with visually similar categories, such as the classes in all fine-grained datasets and some super-class subsets of generic datasets. As shown in Fig. 1 (a), replacing the backbone of

*Corresponding author.

the parametric baseline [45] with the powerful CLIP [36] visual encoder still struggles to generalize certain visual concepts, leading to sub-optimal results. Inspired by the idea that the textual modality can provide complementary discriminative information, we decided to introduce text information into the GCD task to compensate for the insufficient discriminative of visual concepts. However, the lack of class names for unlabelled data in GCD makes it impractical to use the text encoder, thus locking the multi-modal potential of CLIP for GCD.

In order to tackle this challenging problem, in this paper, we propose a generative-based method to **GE**nerate pseudo **T**ext embeddings for unlabelled data, dubbed **GET**. In particular, we first introduce a *Text Embedding Synthesizer* (TES) module based on the vision-language alignment property of CLIP, producing reliable and modality-aligned pseudo text features. As shown in Fig. 1 (b), TES learns a mapping that transforms image embeddings into text embeddings. Specifically, TES converts visual embeddings into tokens for the text encoder, eliminating the need for textual input. To mitigate the gap between generated pseudo-text embeddings and real text embeddings, TES distills knowledge from real text embeddings corresponding to labeled data. Additionally, TES aligns text and image for the same instance, enforcing consistency between language and vision while preventing overfitting to known classes. This training approach renders TES equivalent to a special finetuned text encoder with only visual input. From another perspective, our TES can be considered as performing an image captioning task [31].

To leverage such multi-modal features in the GCD task, we propose a dual-branch multi-modal joint training strategy with a cross-modal instance consistency objective. One branch focuses on visual information, while the other branch supplements it with text information. Through joint learning on the GCD task, visual and semantic information aspects mutually enhance each other. Furthermore, our cross-modal instance consistency objective enforces the instance have the same relationship in both visual and text modalities with anchors constructed by labeled instances, promoting the interaction and alignment of visual and text embedding space. With the supplementation of text embeddings generated by TES and an appropriate dual-branch training strategy, the multi-modal features correct the classification hyperplane, enhancing discriminative ability while reducing bias issues.

To summarize, our contributions are as follows:

- To tackle the problem that the text encoder can not be used on the unlabelled data, we propose a TES module converting visual embeddings into tokens of the CLIP's text encoder to generate pseudo text embeddings.
- Through the proposed cross-modal instance consistency objective in our dual-branch framework, information of different modalities mutually enhances each other, producing more discriminative classification prototypes.
- Our method achieves state-of-the-art results on multiple benchmarks, providing GCD a multi-modal paradigm.

2. Related Works

Novel Class Discovery (NCD). NCD can be traced back to KCL [18], where pairwise similarity generated by a similarity prediction network guides clustering reconstruction, offering a constructive approach for transfer learning across tasks and domains. Early methods are based on two objectives: pretraining on labeled data and clustering on unlabelled data. RS [16] performs a self-supervised pretraining on both labeled and unlabelled data, alleviating the model's bias towards known classes. Simultaneously, RS proposes knowledge transfer through rank statistics, which has been widely adopted in subsequent research. [49] proposes a two-branch learning framework with dual ranking statistics, exchanging information through mutual knowledge distillation, which is similar to our approach to some extent. Differently, our two branches focus on semantic and visual information rather than local and global characteristics in [49]. In order to simplify NCD approaches, UNO [13] recommends optimizing the task with a unified cross-entropy loss using the multi-view SwAV [4] exchange prediction strategy, which sets a new paradigm.

Generalized Category Discovery (GCD). Recently, GCD [41] extends NCD to a more realistic scenario, where unlabelled data comes from both known and unknown classes. GCD [41] employs a pre-trained vision transformer [11] to provide initial visual representations, finetuning the backbone through supervised and self-supervised contrastive learning on the labeled and the entire data. Once the model learns discriminative representations, semi-supervised k-means are used for classification by constraining the correct clustering of labeled samples. As an emerging and realistic topic, GCD is gradually gaining attention. PromptCAL [48] propose a two-stage framework to tackle the class collision issue caused by false negatives while enhancing the adaptability of the model on downstream datasets. SimGCD [45] introduces a parametric classification approach, addressing the computational overhead of GCD clustering while achieving remarkable improvements. Specifically, SimGCD adds a classifier on top of GCD and jointly learns self-distillation and supervised training strategies. μGCD [43] examines the taxonomy bias in previous methods by introducing the Clevr4 dataset and employs the mean-teacher technique and a more efficient training strategy thus achieving remarkable performance improvements. CLIP-GCD [33] mines text descriptions from a large text corpus to use the text encoder and simply concatenates visual and text features for classification. In contrast, our method focuses on the dataset itself, without introducing additional corpus. Most recently, TextGCD [52] collects many text tags from multiple benchmarks and leverages LLMs to enhance these tags, constructing a visual lexicon. It then generates textual descriptions for each sample based on the similarity between the visual lexicon and visual feature. Different from these methods, our GET employs CLIP to introduce multi-modal information into the task without relying on any additional databases or large models.

Vision-Language Pre-training. Vision-Language pre-training [6, 7, 9, 12, 14, 25] aims to train a large-scale model on extensive image-text data, which, through fine-tuning, can achieve strong performance on a range of downstream visual-language tasks. Some studies [8, 24, 27, 28, 39] achieve improved performance in various image-language tasks by modeling image-text interactions through a fusion approach. However, the need to encode all image-text pairs in the fusion approach makes the inference speed in image-text retrieval tasks slow. Consequently, some studies [19, 36] propose a separate encoding of images and texts, and project image and text embeddings into a joint embedding space through contrastive learning. CLIP [36] uses contrastive training on large-scale image-text pairs, minimizing the distance between corresponding images and texts while simultaneously maximizing the distance between non-corresponding pairs. The strong generalization capabilities and multi-modal properties of CLIP prompt us to introduce it to the GCD Task.

3. Preliminaries

3.1. Problem formulation

In the context of GCD, the training data $\mathcal{D}$ is divided into a labeled dataset $\mathcal{D}_l = \left\{(\boldsymbol{x}_i^l, y_i^l)\right\}_{i=1}^N \in \mathcal{X} \times \mathcal{Y}_l$ and an unlabelled dataset $\mathcal{D}_u = \left\{(\boldsymbol{x}_i^u, y_i^u)\right\}_{i=1}^M \in \mathcal{X} \times \mathcal{Y}_u$, where $\mathcal{Y}_l$ and $\mathcal{Y}_u$ represent the label space while $\mathcal{Y}_l \subset \mathcal{Y}_u$, and $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$. $|\mathcal{Y}_l|$ and $|\mathcal{Y}_u|$ represent the number of categories for labeled samples and unlabelled samples, respectively. Following the setting in [41, 45], we assume the class number of new classes $|\mathcal{Y}_u \backslash \mathcal{Y}_l|$ is known, or it can be estimated through some off-the-shelf methods [15, 41]. The goal of GCD is to correctly cluster unlabelled samples with the help of labeled samples.

3.2. Parametric GCD method (SimGCD)

In this paper, we tackle the GCD problem in a parametric way which is proposed by SimGCD [45]. It trains a unified prototypical classification head for all new/old classes to perform GCD through a DINO-like form of self-distillation. Specifically, it includes two types of loss functions: representation learning and parametric classification. For representation learning, it performs supervised representation learning [20] $\mathcal{L}_{scon}$ on all labeled data and self-supervised contrastive learning $\mathcal{L}_{con}$ on all training data, the loss functions are as follows:

$$\mathcal{L}_{scon} = -\frac{1}{|B_l|}\sum_{i \in B_l}\frac{1}{|\mathcal{N}_i|}\sum_{q \in \mathcal{N}_i}\log\frac{\exp\left(\frac{\boldsymbol{h}_i{}^\top \boldsymbol{h}_q{}'}{\tau_{sc}}\right)}{\sum_{n \in B_l}^{n \neq i}\exp\left(\frac{\boldsymbol{h}_i{}^\top \boldsymbol{h}_n{}'}{\tau_{sc}}\right)}, \tag{1}$$

$$\mathcal{L}_{con} = -\frac{1}{|B|}\sum_{i \in B}\log\frac{\exp\left(\frac{\boldsymbol{h}_i{}^\top \boldsymbol{h}_i{}'}{\tau_c}\right)}{\sum_{n \in B}^{n \neq i}\exp\left(\frac{\boldsymbol{h}_i{}^\top \boldsymbol{h}_n{}'}{\tau_c}\right)}, \tag{2}$$

where $\mathcal{N}_i$ denotes the indices of other images with the same semantic label as $\boldsymbol{x}_i$ in a batch, B_l corresponds to the labeled subset of the mini-batch B, τ_{sc} and τ_c are temperature values. For visual embeddings $\boldsymbol{z}_i$ and $\boldsymbol{z}_i'$ of two views $\boldsymbol{x}_i$ and $\boldsymbol{x}_i'$ generated by the image encoder, an MLP layer $g(\cdot)$ is used to map $\boldsymbol{z}_i$ and $\boldsymbol{z}_i'$ to high-dimensional embeddings $\boldsymbol{h}_i = g(\boldsymbol{z}_i)$ and $\boldsymbol{h}_i' = g(\boldsymbol{z}_i')$. For parametric classification, all labeled data are trained by a cross-entropy loss $\mathcal{L}_{\text{cls}}^s$ and all training data are trained by a self-distillation loss $\mathcal{L}_{\text{cls}}^u$:

$$\mathcal{L}_{\text{cls}}^s = \frac{1}{|B_l|}\sum_{i \in B^l}\mathcal{H}\left(y_i, \sigma(\boldsymbol{p}_i, \tau_s)\right), \tag{3}$$

$$\mathcal{L}_{\text{cls}}^u = \frac{1}{|B|}\sum_{i \in B}\mathcal{H}\left(\sigma(\boldsymbol{p}_i{}', \tau_t), \sigma(\boldsymbol{p}_i, \tau_s)\right), \tag{4}$$

where $\sigma(\cdot)$ is the softmax function, $\boldsymbol{p}_i$ and $\boldsymbol{p}_i{}'$ are the outputs of two views x_i and x_i' on the prototypical classifier, respectively. τ_s is a temperature parameter and τ_t is a sharper version. $\mathcal{H}(\cdot)$ denotes the cross-entropy function, y_i is the corresponding ground truth of x_i, and $\sigma(\boldsymbol{p}_i{}', \tau_t)$ is the soft pseudo-label of x_i.

In addition, SimGCD also introduces a mean-entropy maximization regularization term $H(\overline{\boldsymbol{p}})$ to prevent trivial solutions, where $H(\cdot)$ is the entropy of predictions [37], $\overline{\boldsymbol{p}} = \frac{1}{2|B|}\sum_{i \in B}\left(\sigma(\boldsymbol{p}_i{}', \tau_s) + \sigma(\boldsymbol{p}_i, \tau_s)\right)$ is the mean softmax probability of a batch. By using the above loss functions and regularization term to train the model, SimGCD has achieved significant improvements, however, it struggles with performance on visually similar categories due to the use of single visual modality information.

4. Our Method

In this paper, we propose GET, which addresses the GCD task in a multi-modal paradigm. As shown in Fig. 2, our GET contains two stages. In the first stage, we learn a text embedding synthesizer (TES, in Sec. 4.1) to generate pseudo text embeddings for each sample. In the second stage, a dual-branch multi-modal joint training strategy with cross-modal instance consistency (in Sec. 4.2) is introduced to fully leverage multi-modal features.

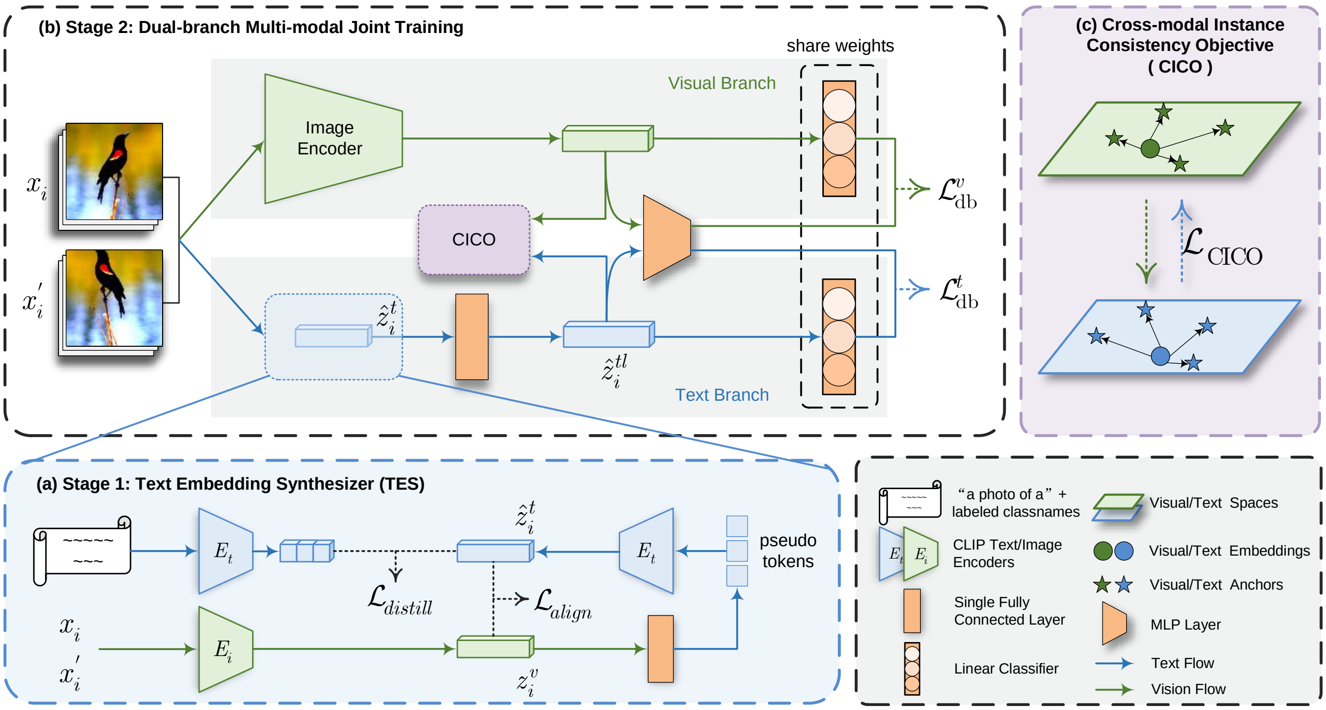

Figure 2. Overview of our GET framework. (a) In the first stage, we introduce a text embedding synthesizer that generates pseudo text embeddings for unlabelled data. TES learns a linear mapping that transforms image features into input tokens for the text encoder. The resulting pseudo text embeddings are then used for joint training in the second stage. (b) We proposed a dual-branch multi-modal joint training framework with a cross-modal instance consistency objective in the second stage. Two branches utilize the same parameterized training strategy [45] while focusing on text and visual information, respectively. (c) Our cross-modal instance consistency objective makes visual and text information exchange and benefit from each other.

4.1. Text embedding synthesizer

The absence of natural language class names for unlabelled data makes it challenging to introduce text information into the GCD task. In this paper, we attempt to generate pseudo text embeddings aligned with visual embeddings for each image from a feature-based perspective.

Inspired by BARON [46], which treats embeddings within bounding boxes as embeddings of words in a sentence to solve the open-vocabulary detection task, we propose a text embedding synthesizer (TES). Specifically, our TES leverages the property that CLIP can generate aligned vision-language features, converting visual embeddings into tokens of the CLIP's text encoder to generate pseudo text embeddings for each sample. The architecture of TES is shown in Fig. 2 (a). For each image $\boldsymbol{x}_i$ in a mini-batch, we use CLIP's image encoder to obtain its visual embedding $\boldsymbol{z}_i^v$. A single fully connected layer l is used to map the visual embedding to pseudo tokens that serve as input to the CLIP's text encoder, thus generating corresponding pseudo text embeddings $\hat{\boldsymbol{z}}_i^t$.

The objective of TES contains an align loss on all samples and a distill loss on labeled samples. To align the generated pseudo text embeddings $\hat{\boldsymbol{z}}_i^t$ with their corresponding visual features $\boldsymbol{z}_i^v$, our align loss leverages the modality alignment property of CLIP's encoders, pulling correct visual-text embedding pairs closer while pushing away the incorrect ones. The align loss consists of symmetric components $\mathcal{L}_{align}^v$ and $\mathcal{L}_{align}^t$, calculated as:

$$\mathcal{L}_{align}^v = -\frac{1}{|B|}\sum_{i\in B}\log\frac{\exp\left({\boldsymbol{z}_i^v}^\top \hat{\boldsymbol{z}}_i^t/\tau_a\right)}{\sum_{j\in B}\exp\left({\boldsymbol{z}_i^v}^\top \hat{\boldsymbol{z}}_j^t/\tau_a\right)}, \tag{5}$$

$$\mathcal{L}_{align}^t = -\frac{1}{|B|}\sum_{i\in B}\log\frac{\exp\left({\hat{\boldsymbol{z}}_i^t}^\top \boldsymbol{z}_i^v/\tau_a\right)}{\sum_{j\in B}\exp\left({\hat{\boldsymbol{z}}_i^t}^\top \boldsymbol{z}_j^v/\tau_a\right)}, \tag{6}$$

where $\hat{\boldsymbol{z}}_i^t$ and $\boldsymbol{z}_i^v$ are ℓ_2-normalised, and τ_a is a temperature parameter. Thus, the align loss is $\mathcal{L}_{align} = \mathcal{L}_{align}^v + \mathcal{L}_{align}^t$.

To ensure that our generated pseudo text features reside in the same embedding space as real text features and maintain consistency, we introduce a distill loss $\mathcal{L}_{distill}$:

$$\begin{aligned}\mathcal{L}_{distill} = &-\frac{1}{|B_l|}\sum_{i\in B_l}\log\frac{\exp\left({\hat{\boldsymbol{z}}_i^t}^\top T(n_i)\right)}{\sum_{j=0}^{|\mathcal{Y}_l|}\mathbb{1}_{[j\neq n_i]}\exp\left({\hat{\boldsymbol{z}}_i^t}^\top T(j)\right)}\\ &+\frac{1}{|B_l|}\sum_{i\in B_l}\left(\hat{\boldsymbol{z}}_i^t - T(n_i)\right)^2,\end{aligned} \tag{7}$$

where $T \in |\mathcal{Y}_l| \times dim$ are the real text embeddings of $|\mathcal{Y}_l|$ semantic labels, $n_i \in \mathcal{Y}_l$ indexes the corresponding class name of $\boldsymbol{x}_i$ among all known class names, $T(j)$ denotes the j-th real text embeddings of all known class names and $\mathbb{1}_{[\cdot]}$ is the indicator function. Each vector in T is produced by the text encoder using the prompt "a photo of a {CLS}" where {CLS} denotes the corresponding class name.

The overall objective of our text embedding synthesizer is $\mathcal{L}_{TES} = \mathcal{L}_{align} + \mathcal{L}_{distill}$. The distill loss is used to guide pseudo text embeddings of the network's output towards the real semantic corresponding space and adapt the model to the distribution of the dataset, while the align loss prevents overfitting to known classes and enforces the consistency between two modalities. Moreover, we introduce a multi-view strategy for TES. Specifically, we calculate both the align loss $\mathcal{L}_{align}$ and distill loss $\mathcal{L}_{distill}$ for two different views $\boldsymbol{x}_i$ and $\boldsymbol{x}_i{}'$ of the same image in a mini-batch. This further implicitly enhances the instance discriminative nature [47] of our TES training, allowing different views of the same labeled image to generate identical pseudo text embeddings. The generated pseudo text embeddings $\hat{\boldsymbol{z}}_i^t$ are then used for joint training in the second stage.

4.2. Dual-branch multi-modal joint training

Intuitively, the introduction of multi-modal information can have a positive impact on the GCD task. Textual information can serve as an effective complement to visual information, enhancing the model's discriminative capabilities. However, how to effectively utilize visual and text information in the GCD task and make the most of their respective roles remains challenging. In this paper, we propose a dual-branch architecture as illustrated in Fig. 2 (b), which focuses on semantic and visual information, respectively. We employ the same parametric training strategy (in Sec. 3.2) for each branch to promote that the model has aligned and complementary discriminative capabilities for visual and text features of the same class. Furthermore, we introduce a cross-modal instance consistency loss, which constrains the instance relationships of samples in both visual and text spaces, enabling the two branches to learn from each other. We use the indicator v to represent the visual concept while t for the text concept.

Visual-branch. The objective of the visual branch contains a representation learning part and a parametric classification part. Given a visual embedding $\boldsymbol{z}_i^v$ of image $\boldsymbol{x}_i$ generated by the image encoder, an MLP layer $g(\cdot)$ is used to map $\boldsymbol{z}_i^v$ to a high-dimensional embedding $\boldsymbol{h}_i^v = g(\boldsymbol{z}_i^v)$. Meanwhile, we employ a prototypical classifier $\eta(\cdot)$ to generate classification probability distribution $\boldsymbol{p}_i^v = \eta(\boldsymbol{z}_i^v)$. Simply replace all the high-dimensional embedding $\boldsymbol{h}$ (the subscript is omitted for brevity) in Eq. (1) and Eq. (2) with its corresponding visual branch version $\boldsymbol{h}^v$ can obtain the the supervised contrastive loss $\mathcal{L}_{scon}^v$ and the self-supervised contrastive loss $\mathcal{L}_{ucon}^v$. The overall representation learning loss is balanced with λ, written as:

$$\mathcal{L}_{\text{rep}}^v = (1-\lambda)\mathcal{L}_{ucon}^v + \lambda\mathcal{L}_{scon}^v\,. \tag{8}$$

For the parametric classification part, just replace $\boldsymbol{p}_i$ and $\boldsymbol{p}_i{}'$ of Eq. (3) and Eq. (4) with $\boldsymbol{p}_i^v$ and $\boldsymbol{p}_i^{v\prime}$ can obtain the cross-entropy loss $\mathcal{L}_{\text{cls-v}}^s$ and the self-distillation loss $\mathcal{L}_{\text{cls-v}}^u$. Thus, the classification loss is $\mathcal{L}_{\text{cls}}^v = (1-\lambda)\mathcal{L}_{\text{cls-v}}^u + \lambda\mathcal{L}_{\text{cls-v}}^s$.

The overall objective of the visual branch is as follows:

$$\mathcal{L}_{\text{db}}^v = \mathcal{L}_{\text{rep}}^v + \mathcal{L}_{\text{cls}}^v\,. \tag{9}$$

Text-branch. Our text branch simply adopts the same training strategy as the visual branch. That is, in particular, given a text embedding $\hat{\boldsymbol{z}}_i^t$ generated by TES, we first input it into a fully connected layer to gain a learnable text embedding $\hat{\boldsymbol{z}}_i^{tl}$ while change its dimension. Simply replace $\boldsymbol{h}_i^v$ in the representation learning objective $\mathcal{L}_{\text{rep}}^v$ with $\boldsymbol{h}_i^t = g(\hat{\boldsymbol{z}}_i^{tl})$ and replace $\boldsymbol{p}_i^v$ in classification parts $\mathcal{L}_{\text{cls}}^v$ with $\boldsymbol{p}_i^t = \eta(\hat{\boldsymbol{z}}_i^{tl})$ can yield the corresponding text objectives $\mathcal{L}_{\text{rep}}^t$ and $\mathcal{L}_{\text{cls}}^t$. In other words, changing the visual conception indicator v into text conception indicator tl can get the corresponding objectives for the text branch. Thus, the overall objective of our text branch can be formally written as $\mathcal{L}_{\text{db}}^t = \mathcal{L}_{\text{rep}}^t + \mathcal{L}_{\text{cls}}^t$

To mitigate the bias between old and new classes, we extended the mean-entropy regularization [1] to a multi-modal mean entropy regularization $H_{mm} = H(\overline{p}_{mm}, \overline{p}_{mm})$, here $\overline{p}_{mm}$ can calculate by $\overline{p}_{mm} = \frac{1}{2|B|}\sum_{i\in B}(\sigma(\boldsymbol{p}_i^v, \tau_s) + \sigma(\boldsymbol{p}_i^t, \tau_s))$. In this way, the prediction probabilities in different modalities for each prototype are constrained to be the same, preventing trivial solutions.

Cross-modal instance consistency objective. In order to enable the two branches to learn from each other while encouraging agreement between two different modes, we propose a cross-modal instance consistency objective (CICO), shown in Fig. 2 (c). Our CICO has the same form of mutual knowledge distillation as [49], but we distill the instance consistency between the two branches. For each mini-batch B, we choose its labeled subset B_l containing K categories as anchor samples, calculate the visual and text prototypes for K categories as visual anchors $\mathcal{P}_v$ and text anchors $\mathcal{P}_t$, respectively. We define the instance relationships in visual and text branches as:

$$\begin{aligned} s_i^v &= \sigma({\boldsymbol{z}_i^v}^\top \mathcal{P}_v)\,, \\ s_i^t &= \sigma({\hat{\boldsymbol{z}}_i^{tl}}{}^\top \mathcal{P}_t)\,. \end{aligned} \tag{10}$$

Thus our CICO can formally written as:

$$\mathcal{L}_{\text{CICO}} = \frac{1}{2|B|}\sum_{i\in B}\left(D_{KL}(s_i^t||s_i^v) + D_{KL}(s_i^v||s_i^t)\right)\,, \tag{11}$$

where D_{KL} is the Kullback-Leibler divergence. Mutual knowledge distillation on instance relationships for two

	CUB			Stanford Cars			FGVC-Aircraft			CIFAR10			CIFAR100			ImageNet-100		
Method	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
k-means [29]	34.3	38.9	32.1	12.8	10.6	13.8	16.0	14.4	16.8	83.6	85.7	82.5	52.0	52.2	50.8	72.7	75.5	71.3
RS+ [16]	33.3	51.6	24.2	28.3	61.8	12.1	26.9	36.4	22.2	46.8	19.2	60.5	58.2	77.6	19.3	37.1	61.6	24.8
UNO+ [13]	35.1	49.0	28.1	35.5	70.5	18.6	40.3	56.4	32.2	68.6	**98.3**	53.8	69.5	80.6	47.2	70.3	95.0	57.9
ORCA [2]	35.3	45.6	30.2	23.5	50.1	10.7	22.0	31.8	17.1	81.8	86.2	79.6	69.0	77.4	52.0	73.5	92.6	63.9
GCD [41]	51.3	56.6	48.7	39.0	57.6	29.9	45.0	41.1	46.9	91.5	97.9	88.2	73.0	76.2	66.5	74.1	89.8	66.3
GPC [50]	55.4	58.2	53.1	42.8	59.2	32.8	46.3	42.5	47.9	92.2	98.2	89.1	77.9	85.0	63.0	76.9	94.3	71.0
DCCL [34]	63.5	60.8	64.9	43.1	55.7	36.2	-	-	-	96.3	96.5	96.9	75.3	76.8	70.2	80.5	90.5	76.2
PromptCAL [48]	62.9	64.4	62.1	50.2	70.1	40.6	52.2	52.2	52.3	**97.9**	96.6	98.5	81.2	84.2	75.3	83.1	92.7	78.3
SimGCD [45]	60.3	65.6	57.7	53.8	71.9	45.0	54.2	59.1	51.8	97.1	95.1	98.1	80.1	81.2	77.8	83.0	93.1	77.9
μGCD [43]	65.7	68.0	64.6	56.5	68.1	50.9	53.8	55.4	53.0	-	-	-	-	-	-	-	-	-
LegoGCD [3]	63.8	71.9	59.8	57.3	75.7	48.4	55.0	61.5	51.7	97.1	94.3	98.5	81.8	81.4	**82.5**	86.3	94.5	82.1
GCD-CLIP	57.6	65.2	53.8	65.1	75.9	59.8	45.3	44.4	45.8	94.0	97.3	92.3	74.8	79.8	64.6	75.8	87.3	70.0
SimGCD-CLIP	71.7	76.5	69.4	70.0	83.4	63.5	54.3	58.4	52.2	97.0	94.2	98.4	81.1	85.0	73.3	90.8	95.5	88.5
GET (Ours)	**77.0**	**78.1**	**76.4**	**78.5**	**86.8**	**74.5**	**58.9**	**59.6**	**58.5**	97.2	94.6	**98.5**	**82.1**	**85.5**	75.5	**91.7**	**95.7**	**89.7**

Table 1. Results (%) on fine-grained and generic datasets. The best results are highlighted in **bold**.

modalities makes visual and text flows exchange and benefit from each other, thus the two branches can serve as complementary discriminative aids to each other.

The overall optimization objective of our method is:

$$\mathcal{L}_{\text{Dual}} = \mathcal{L}_{\text{db}}^{v} + \mathcal{L}_{\text{db}}^{t} - \epsilon H_{mm} + \lambda_c \mathcal{L}_{\text{CICO}} \,. \tag{12}$$

Since information from different modalities is exchanged and learned through CICO and injected into the visual backbone, we utilize the last-epoch visual branch for inference.

5. Experiments

5.1. Experimental setup

Datasets. We evaluate our method on multiple benchmarks, including three image classification generic datasets (*i.e.*, CIFAR 10/100 [22] and ImageNet-100 [10]), three fine-grained datasets from Semantic Shift Benchmark [42] (*i.e.*, CUB [44], Stanford Cars [21] and FGVC-Aircraft [30]), and three challenging datasets (*i.e.*, Herbarium 19 [40], ImageNet-R [17] and ImageNet-1K [10]). We are the first to introduce ImageNet-R into the GCD task, which contains various renditions of 200 ImageNet classes, thus challenging the GCD's assumption that the data comes from the same domain. The data splits are reported in *Supp*.

Evaluation and implementation details. Following standard evaluation protocol in [41, 45], we evaluate the performance with clustering accuracy (ACC). We use a CLIP [36] pre-trained ViT-B/16 [11] as the image and text encoder. In the first stage, we train a fully connected layer. In the second stage, we remove the projector of the image encoder, resulting in features with a dimension of 768. Other details and the pseudo-code can be found in *Supp*.

5.2. Comparison with state of the arts

In this section, we compare GET with several state-of-the-art methods. GCD [41] and SimGCD [45] provide paradigms for non-parametric and parametric classification, thus we replace their backbone with CLIP for a fair comparision, denoted as GCD-CLIP and SimGCD-CLIP.

Evaluation on fine-grained and generic datasets. As shown in Tab. 1, our method achieves consistently remarkable success on all three fine-grained datasets. Specifically, we surpass SimGCD-CLIP by 5.3%, 8.5%, and 4.6% on 'All' classes of CUB, Stanford Cars, and Aircraft, respectively. In fine-grained datasets, the visual conceptions of distinct classes exhibit high similarity, making it challenging for classification based solely on visual information. However, their text information can provide additional discriminative information. Consequently, our GET significantly enhances classification accuracy through the reciprocal augmentation of text and visual information flows. In Tab. 1, we also present the performances for three generic datasets. Due to the low resolution of the CIFAR dataset and model biases (CLIP itself performs poorly on CIFAR100, with a zero-shot performance of 68.7), the results for novel classes are inferior compared to the DINO backbone. However, despite the inherent limitations in the discriminative capability of CLIP itself, our method still achieves an improvement of 0.4% on 'Old' classes of CIFAR10 and 2.2% on 'New' classes of CIFAR100, compared to SimGCD-CLIP. For ImageNet-100, SimGCD-CLIP has achieved an exceptionally saturated result of 90.8% on 'All' classes, further advancements pose considerable challenges. However, leveraging the additional modality information, GET elevates the performance ceiling to an impressive 91.7%.

Method	Herbarium 19			ImageNet-1K			ImageNet-R		
	All	Old	New	All	Old	New	All	Old	New
k-means [29]	13.0	12.2	13.4	-	-	-	-	-	-
RS+ [16]	27.9	55.8	12.8	-	-	-	-	-	-
UNO+ [13]	28.3	53.7	14.7	-	-	-	-	-	-
ORCA [2]	20.9	30.9	15.5	-	-	-	-	-	-
μGCD [43]	45.8	61.9	37.2	-	-	-	-	-	-
LegoGCD [3]	45.1	57.4	38.4	62.4	**79.5**	53.8	-	-	-
GCD [41]	35.4	51.0	27.0	52.5	72.5	42.2	32.5	58.0	18.9
SimGCD [45]	44.0	58.0	36.4	57.1	77.3	46.9	29.5	48.6	19.4
GCD-CLIP	37.3	51.9	29.5	55.0	65.0	50.0	44.3	79.0	25.8
SimGCD-CLIP	48.9	**64.7**	40.3	61.0	73.1	54.9	54.9	72.8	45.3
GET (Ours)	**49.7**	64.5	**41.7**	**62.4**	74.0	**56.6**	**58.1**	**78.8**	**47.0**

Table 2. Results (%) on more challenging datasets.

	TES	Dual-branch	CICO	Stanford Cars			CIFAR100		
				All	Old	New	All	Old	New
(1)	✗	✗	✗	70.0	83.4	63.5	81.1	85.0	73.3
(2)	✓	✓	✗	76.2	85.3	71.7	81.0	85.3	72.3
(3)	✓	✓	✓	**78.5**	**86.8**	**74.5**	**82.1**	**85.5**	**75.5**

Table 3. Ablation study of different components.

Evaluation on more challenging datasets. As shown in Tab. 2, GET outperforms all other methods for both 'All' and 'New' classes on Herb19 and ImageNet-1K datasets. In particular, our method achieves a notable improvement of 1.4% and 1.7% on 'New' classes of Herb19 and ImageNet-1K, respectively. Furthermore, the suboptimal performance of GCD and SimGCD with the DINO backbone on the ImageNet-R dataset highlights the difficulty of DINO in discovering new categories with multiple domains. Despite multiple domains for images of the same category, their textual information remains consistent. Our method effectively integrates text information, resulting in a substantial improvement of 3.2% and 6.0% over the sota for 'All' classes and 'Old' classes, respectively. It is worth noting that, owing to the text consistency within the same category, our text branch achieves remarkable 62.6% and 63.5% accuracy for 'All' classes of Imagenet-R and ImageNet-1K, respectively.

5.3. Ablation study and analysis

Effectiveness of different components. To evaluate the effectiveness of different components, we conduct an ablation study on SCars and CIFAR100 in Tab. 3. Comparing (2) with (1), leverage the text features generated by TES, resulting in a 6.2% improvement on SCars' 'All' classes and a 0.3% improvement on CIFAR100's 'Old' classes. Furthermore, comparing (3) with (1), CICO enables the two branches to exchange information and mutually benefit from each other, resulting in remarkable improvements of 11% on SCars' 'New' and 2.2% on CIFAR100's 'New'.

	Dual-branch	Concat	Mean	Stanford Cars			CIFAR100		
				All	Old	New	All	Old	New
(1)	✗	✓	✗	68.9	79.1	64.0	79.9	85.5	68.7
(2)	✗	✗	✓	72.0	85.0	65.6	81.1	84.3	74.8
(3)	✓	✗	✗	**78.5**	**86.8**	**74.5**	**82.1**	**85.5**	**75.5**

Table 4. Comparison with different fusion methods.

Method		Total Params	All	Old	New
Baseline	SimGCD-CLIP	92.2M	71.7	76.5	69.4
Text-Retrieval	WordNet	155.8M	69.8	77.1	66.2
	CC3M	155.8M	72.3	**79.1**	68.9
VQA	BLIP (ViT-L)	625.8M	67.1	74.3	63.5
	BLIP-2 (opt2.7b)	3.9B	71.3	73.5	70.2
Captioning	BLIP (ViT-L)	625.8M	40.5	54.6	33.4
	BLIP-2 (opt2.7b)	3.9B	42.6	56.1	35.8
Feature-Generation	TES (Ours)	165.1M	**77.0**	78.1	**76.4**
	TES w/o $\mathcal{L}_{align}$	165.1M	74.7	76.9	73.5
	TES w/o $\mathcal{L}_{distill}$	165.1M	75.3	77.5	74.2

Table 5. Experiments on different pseudo text embeddings.

Comparison with different fusion methods. In Tab. 4, we compare our dual-branch strategy with other modality fusion methods, including concatenation and mean. Although they may show improvements due to the multi-modal information, we demonstrate that joint learning of the two branches is more effective as it encourages the model to have complementary and aligned discriminative capabilities for visual and text features of the same class, leading to more discriminative multi-modal prototypes.

Effectiveness of TES. To demonstrate the superiority of TES, we conduct experiments that replaced the embeddings generated by TES with embeddings obtained through *Text-Retrieval*, *VQA*, and *Captioning* on the CUB dataset. For *Text-Retrieval*, we retrieve the most similar text for each image from two corpora (WordNet [32] and CC3M [38]) based on the cosine similarity between the image and text embeddings. We use BLIP [25] and BLIP-2 [26] to perform the *Captioning* and use the question "What's the name of the bird in the image?" to perform *VQA*. As in Tab. 5, due to the high visual similarity in fine-grained images, category names retrieved or generated through VQA are often imprecise, making them less effective. Meanwhile, captioning methods tend to describe object poses and scenes rather than class-specific information, leading to varied captions for samples of the same class, which significantly harms category discovery. Our method achieves the best performance with moderate parameters. We provide additional experiments about TES in *Supp*, including its architectural design, feature distribution, and flexibility.

Results using different prompts. In our method, we use a simple prompt: "a photo of a {CLS}." Tab. 6 presents experimental results exploring the use of alternative prompts:

Prompts	CUB			Stanford Cars		
	All	Old	New	All	Old	New
(1)	77.0	78.1	76.4	78.5	86.8	**74.5**
(2)	76.3	78.2	75.4	78.5	88.2	73.8
(3)	76.8	**78.7**	75.8	78.6	**90.4**	72.9
(4)	**78.3**	77.6	**78.7**	**79.1**	88.8	74.3

Table 6. Results using different prompts.

Methods	NEV			TV-100		
	All	Old	New	All	Old	New
CLIP(zero-shot)	10.7	-	-	1.93	-	-
SimGCD	54.7	88.0	38.0	35.2	50.3	29.2
SimGCD-CLIP	79.1	**96.7**	70.3	55.7	75.8	47.8
GET (Ours)	**85.3**	96.0	**80.0**	**57.1**	**77.3**	**49.2**

Table 7. The results on the NEV and TV-100 datasets.

(1) "a photo of a {CLS}," (2) "a photo of a {CLS}, which is a type of bird/car," (3) descriptions of {CLS} generated by LLM (GPT4o-mini), and (4) average the textual features of (1) to (3). Our prompt is simple yet effective, while more finely designed prompts can further improve performance.

Discussion about using CLIP in GCD. A key purpose of GCD is to discover novel classes, which highly rely on the initial representation discrimination provided by the backbone model. Due to the strong generalization ability of CLIP, it can encode more discriminative features, making it a natural idea to introduce CLIP into the GCD task. One concern is that CLIP may have seen unknown classes or class names in GCD. Therefore, we discuss the significance of using CLIP in GCD from three perspectives. **Methodological Significance:** Even CLIP is pre-trained on a vast dataset with potential overlap, its knowledge is implicit and unstructured. Effectively using it for GCD demands novel methodologies, particularly in exploring how to use the text encoder for unlabelled data. The improvements of our method over SimGCD-CLIP validate this significance. **Forward-looking Significance:** To evaluate the ability of category discovery methods in scenarios unseen by CLIP, we constructed a small fine-grained dataset of new energy vehicles (NEV) introduced in 2023, where CLIP lacks prior knowledge. As shown in Tab. 7, experiments on NEV and TV-100 [53](a TV series dataset that the pre-trained CLIP model has not been exposed to) demonstrate that even for categories not seen by CLIP, leveraging the text modality remains crucial for effective category discovery. This serves as a forward-looking exploration of how CLIP's generalization can address future GCD tasks involving truly novel data. **Practical Implications:** Exploring CLIP's potential in realistic scenarios is meaningful, thus, we present experiments on medical and ultra-fine-grained datasets in *SSSupp*. Our work lays the foundation for leveraging CLIP to address challenging GCD applications.

5.4. Qualitative results

Attention map visualization. As in Fig. 3, compared to SimGCD-CLIP, our method additionally focuses on the feather texture of birds, which is crucial for distinguishing visually similar bird species. With the assistance of text information, the attention maps of our visual branch become more refined, focusing on more discriminative regions.

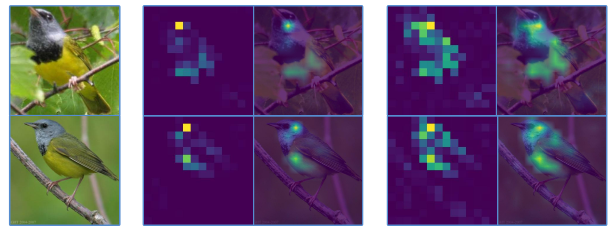

Figure 3. Attention map visualization of class tokens.

The t-SNE visualization. Fig. 4 shows the t-SNE visualization of visual and text features on the randomly sampled 20 classes of the CUB dataset. Both the visual and text features of our method exhibit clearer and compacter clusters. We provide more visualizations and cluster results in *Supp*.

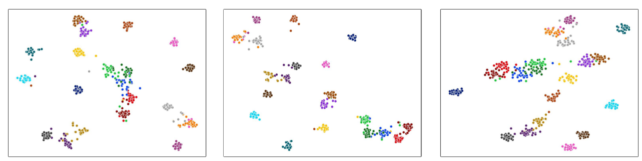

Figure 4. The t-SNE visualizations.

6. Conclusions

In this work, we propose to leverage multi-modal information to solve the GCD task. In particular, we introduce a text embedding synthesizer to generate pseudo text embeddings for unlabelled data. Our text embedding synthesizer module makes it possible to use CLIP's text encoder, thus unlocking the multi-modal potential for the GCD task. Meanwhile, we use a dual-branch training strategy with a cross-modal instance consistency objective, which facilitates collaborative action and mutual learning between the different modalities. Our research extending the GCD to a multi-modal paradigm and the superior performance on multiple benchmarks demonstrates the effectiveness of our method.

7. Acknowledgements

This work is funded by NSFC (NO. 62206135, 62225604), Young Elite Scientists Sponsorship Program by CAST (2023QNRC001), the Fundamental Research Funds for the Central Universities (Nankai Universitiy, 070-63233085), and "Science and Technology Yongjiang 2035" key technology breakthrough plan project (2024Z120). Computation is supported by the Supercomputing Center of Nankai University.

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 *European Conference on Computer Vision*, pages 456–473. Springer, 2022. 5

[2] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. In *International Conference on Learning Representations*, 2022. 6, 7

[3] Xinzi Cao, Xiawu Zheng, Guanhong Wang, Weijiang Yu, Yunhang Shen, Ke Li, Yutong Lu, and Yonghong Tian. Solving the catastrophic forgetting problem in generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16880–16889, 2024. 6, 7

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1

[6] Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023. 3

[7] Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023. 3

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 3

[9] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. 3

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jua Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[11] 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. In *International Conference on Learning Representations*, 2021. 2, 6

[12] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*, 2022. 3

[13] Enrico Fini, Enver Sangineto, Stéphane Lathuiliere, 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, 7

[14] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance. *Visual Intelligence*, 2(1):1–17, 2024. 3

[15] 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. 1, 3

[16] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *International Conference on Learning Representations (ICLR)*, 2020. 2, 6, 7

[17] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 6

[18] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In *International Conference on Learning Representations*, 2018. 2

[19] 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 *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 3

[20] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 3

[21] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, 2013. 6

[22] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical report*, 2009. 6

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 1

[24] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 3

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified

vision-language understanding and generation. In *ICML*, 2022. 3, 7

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 7

[27] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3

[28] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3

[29] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pages 281–297. Oakland, CA, USA, 1967. 6, 7

[30] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6

[31] Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*, 2022. 2

[32] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 1995. 7

[33] Rabah Ouldnoughi, Chia-Wen Kuo, and Zsolt Kira. Clipgcd: Simple language guided generalized category discovery. *arXiv preprint arXiv:2305.10420*, 2023. 2

[34] 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 (CVPR)*, pages 7579–7588, 2023. 6

[35] 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. 1

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 6

[37] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948. 3

[38] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018. 7

[39] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 3

[40] Kiat Chuan Tan, Yulong Liu, Barbara Ambrose, Melissa Tulig, and Serge Belongie. The herbarium challenge 2019 dataset. In *Workshop on Fine-Grained Visual Categorization*, 2019. 6

[41] 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. 1, 2, 3, 6, 7

[42] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: a good closed-set classifier is all you need? In *International Conference on Learning Representations*, 2022. 6

[43] Sagar Vaze, Andrea Vedaldi, and Andrew Zisserman. No representation rules them all in category discovery. *Advances in Neural Information Processing Systems 37*, 2023. 2, 6, 7

[44] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 6

[45] 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, 3, 4, 6, 7

[46] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *CVPR*, 2023. 4

[47] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 5

[48] 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. 1, 2, 6

[49] Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 2, 5

[50] 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. 6

[51] Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery. *arXiv preprint arXiv:2305.06144*, 2023. 1

[52] Haiyang Zheng, Nan Pu, Wenjing Li, Nicu Sebe, and Zhun Zhong. Textual knowledge matters: Cross-modality co-teaching for generalized visual class discovery. In *European Conference on Computer Vision*, pages 41–58. Springer, 2024. 3

[53] Da-Wei Zhou, Zhi-Hong Qi, Han-Jia Ye, and De-Chuan Zhan. Tv100: A tv series dataset that pre-trained clip has not seen. *Frontiers of Computer Science*, 18(5):185349, 2024. 8