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

Supplementary Material

This Supplementary Material provides the following sections:

- Experimental setup (Sec. 8)
- Discussion about Using CLIP in GCD (Sec. 9)
- Additional experiments and analysis (Sec. 10)
- Pseudo-code (Sec. 11)
- Different pre-trained models (Sec. 12)
- Cluster results of GET (Sec. 13)
- Limitations and broader impact (Sec. 14)

8. Experimental setup

Datasets. We evaluate our method on multiple benchmarks, including three image classification generic datasets (*i.e.*, CIFAR 10/100 [9] and ImageNet-100 [3]), three fine-grained datasets from Semantic Shift Benchmark [18] (*i.e.*, CUB [21], Stanford Cars [8] and FGVC-Aircraft [11]), and three challenging datasets (*i.e.*, Herbarium 19 [16], ImageNet-R [6] and ImageNet-1K [3]). Notice that 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. For ImageNet-R, we subsample the first 100 classes as old classes, leaving the rest as new classes; the labeled dataset $\mathcal{D}_l$ consists of half of the old class samples, while the other half and all the new class samples are used to construct unlabelled dataset $\mathcal{D}_u$. Furthermore, we conduct experiments on the TV100 dataset [26], a TV series dataset that the pre-trained CLIP model has not been exposed to. We use the first 50 classes as old categories and the remaining 50 classes as new categories. As for other benchmarks, we follow the previous [17, 23] to sample $\mathcal{D}_l$ and $\mathcal{D}_u$. The details of the standard datasets we evaluate on are shown in Tab. 8.

	Labelled		Unlabelled	
Dataset	Images	Classes	Images	Classes
CIFAR10 [9]	12.5K	5	37.5K	10
CIFAR100 [9]	20.0K	80	30.0K	100
ImageNet-100 [3]	31.9K	50	95.3K	100
CUB [21]	1.5K	100	4.5K	200
Stanford Cars [8]	2.0K	98	6.1K	196
FGVC-Aircraft [11]	1.7K	50	5.0K	100
Herbarium 19 [16]	8.9K	341	25.4K	683
ImageNet-R [6]	7.7K	100	22.3K	200
ImageNet-1K [3]	321K	500	960K	1000

Table 8. The details of the standard datasets we evaluate on.

The NEV dataset As mentioned in the main paper, we conduct a toy experiment to prove that our TES can deal with a scenario where CLIP lacks information on a specific category class. Since CLIP saw most of the visual concepts and corresponding texts before 2022, we constructed a small dataset of new energy vehicles (NEV) that appeared in 2023. As in Tab. 9, the NEV dataset contains 12 categories, each with 50 images from the Internet, and the classnames of the dataset consist of the brand and model of the car. We split them in the same way as standard benchmarks.

Old classes	New classes
BMW_xDrive_M60	Geely_Jiyue_01
BYD_Seagull	Geely_Zeeker_X
BYD_Song_L	Mercedes-Benz_EQE_SUV
BYD_Yangwang_U8	SAIC-Motor_MG_Cyberster
GAC-Motor_Trumpchi_ES9	SAIC-Motor_Rising_F7
Geely_Galaxy_E8	XPeng_X9

Table 9. The class names for the NEV dataset.

Implementation details. We use a CLIP [14] pre-trained ViT-B/16 [4] as the image and text encoder. In the first stage, we train a fully connected layer to transfer image embeddings into pseudo-tokens. In the second stage, the projector of the image encoder is removed, resulting in features with a dimension of 768. The exception is ImageNet-1K, we remain and fine-tune the last projection layer, which avoids gradient explosion and improves results with lower computational cost, resulting in features with a dimension of 512. We use a single linear layer to turn pseudo text embeddings generated by TES into learnable embeddings while changing their dimensions (512 to 768) to match those of the visual features. The batch size is fixed to 128 for training and 256 for testing. Training is done with an SGD optimizer and an initial learning rate of 0.1 decayed by a cosine annealing rule. We train for 200 epochs on each dataset in both two stages. In the first stage, we set the number of pseudo text tokens to 7. The balance coefficient λ is set to 0.35 as [17], and λ_c is set to 1. The temperature value τ_a is set to 0.01 while other temperature values τ_{sc}, τ_c, τ_s , τ_t and the balanced value ϵ are as same as [23]. The augmentation exactly follows the previous, in which RandomCrop creates two views. All experiments are conducted with 4 NVIDIA GeForce RTX 3090 GPUs.

9. Discussion about Using CLIP in GCD

An evident fact is that using a more powerful backbone facilitates the transfer of knowledge learned from labeled data

Method	Backbone	NCT-CRC-HE		
		All	Old	New
SimGCD	DINO	77.1	79.9	75.1
SimGCD	CLIP	79.1	93.2	69.2
GET (ours)	CLIP	**83.8**	**94.5**	**76.3**

Table 10. Results on the medical dataset.

to unlabeled data [17, 19, 22, 23]. Due to the strong generalization ability of CLIP, it can encode more discriminative features, and its multi-modal information aids in discovering new categories, making it a natural choice for introducing CLIP. As discussed in the main text, using CLIP in GCD has three significances: methodological significance, forward-looking significance, and practical implications.

We argue that **the key to leveraging CLIP for GCD lies in how to use its text encoder**, given the presence of unlabeled data in GCD tasks. In this section, we provide supplementary analyses to complement the discussions in the main text. To be specific, we validate the effectiveness of our method and substantiate the incorporation of CLIP into GCD by addressing the following questions:

1. Does the performance gain originate from the CLIP (text encoder) being pre-exposed to the new categories?

In other words, we need to verify the effectiveness of our method on categories that are unseen by CLIP. The experimental results on the NEV and TV-100 datasets (Tab. 7 in the main paper) demonstrate the effectiveness of our method in scenarios where CLIP lacks prior information.

The intuition behind our TES can be explained from two perspectives. First, our trained TES can be considered as a special fine-tuned text encoder. This text encoder takes visual images as input and produces corresponding textual features as output. Our align loss ensures modal alignment, while the distill loss facilitates the model's adaptation to the dataset's distribution. Second, TES can be regarded as a caption model [12]. For each input image, TES assigns a corresponding caption, expressing each caption in the form of modal-aligned text features. The text embeddings or captions corresponding to images can serve as valuable supplementary information, assisting the GCD task in a multi-modal manner.

2. In more realistic scenarios where class names (either old or new) cannot be generated or retrieved, does our method remain effective?

To address this concern, we first conducted GCD experiments on a medical dataset: the NCT-CRC-HE dataset. The NCT-CRC-HE [7] dataset comprises histological images of human colorectal cancer, containing nine categories. We selected the first five categories as the old classes. For the medical dataset, generating or retrieving new class names is challenging, its class names need expert knowledge. Our method directly generates text features aligned with visual features, and the experimental results in Tab. 10 demonstrate its effectiveness.

	SoyAgeing-R1			SoyAgeing-R3			SoyAgeing-R4		
Method	All	Old	New	All	Old	New	All	Old	New
SimGCD	37.3	48.4	31.7	32.7	47.2	25.5	35.4	46.4	29.9
GET (ours)	**47.9**	**56.5**	**43.6**	**46.0**	**55.2**	**41.4**	**46.6**	**52.8**	**43.4**

Table 11. Results on ultra-fine-grained datasets using CLIP backbone.

Moreover, in certain scenarios, it is also difficult to obtain the class names of base categories. For example, in ultra-fine-grained datasets [10, 24], different categories represent different types of soybean leaves. In such cases, category discovery using CLIP becomes significantly more challenging. To address this, we remove the distillation loss in TES, allowing the use of CLIP's text encoder even when base class names are unavailable. Tab. 11 presents the experimental results on three ultra-fine-grained datasets, demonstrating the effectiveness of our approach. Meanwhile, the results in Tab. 5 of the main paper further demonstrate the effectiveness of our TES in scenarios where the class names of base categories are unavailable, by removing the distillation loss.

10. Additional Experiments and Analysis

The architecture of TES. In TES, we use a single linear layer to transform the visual embedding to pseudo tokens and set the number of pseudo text tokens to 7 across all datasets. Experiment results on the architecture of TES for the CUB dataset are presented in Fig. 5, proving that a single linear layer can effectively transfer visual features into text tokens while reducing the computational cost.

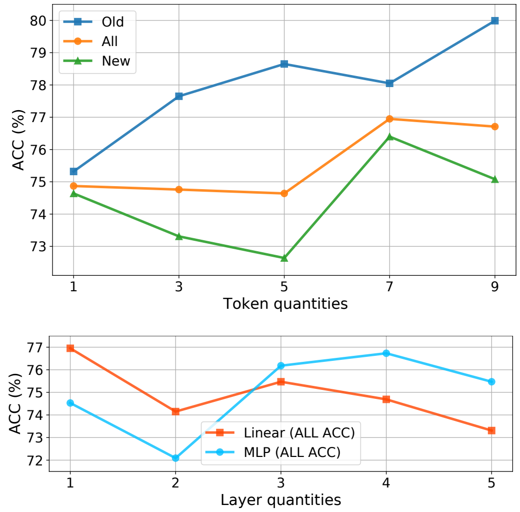

Figure 5. Experiments on the pseudo-tokens and layers in TES.

Effectiveness of text embedding synthesizer. In order

	CUB			Stanford Cars			FGVC-Aircraft			CIFAR10			CIFAR100			ImageNet-100		
Methods	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
PromptCAL [25]	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
PromptCAL-CLIP	65.5	68.7	63.9	74.0	80.8	70.8	54.5	61.8	51.0	88.7	96.5	84.8	80.5	82.4	**76.8**	87.4	93.6	84.3
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 12. Results of PromptCAL-CLIP.

	FGVC-Aircraft			ImageNet-100			ImageNet-R		
Method	All	Old	New	All	Old	New	All	Old	New
GCD-CLIP	45.3	44.4	45.8	75.8	87.3	70.0	44.3	79.0	25.8
+TES	**49.6**	**49.3**	**49.8**	**80.0**	**95.1**	**72.4**	**49.4**	**79.4**	**33.5**

Table 13. Effectiveness of TES in non-parametric GCD.

to prove that our text embedding synthesizer can generate reliable and discriminative representations, we visualize the text embeddings of CIFAR10 with t-SNE. As shown in Fig. 6, the initial text embeddings within the same class exhibit clear clustering, and the learnable embeddings further produce compacter clusters. Moreover, we introduce TES into the non-parametric GCD by straightforwardly concatenating text and image features before semi-supervised k-means classification. As in Tab. 13, with the help of text information, GCD gains about 5% average improvement on 'All' classes over 3 datasets, demonstrating the importance of multi-modal information in GCD task and our TES can be widely used in multiple GCD methods.

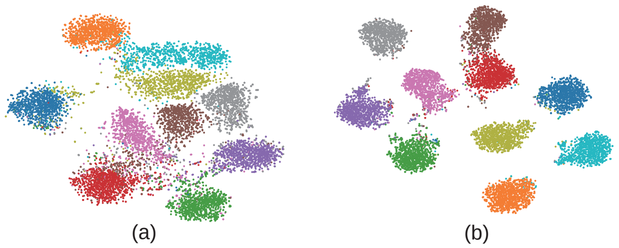

Figure 6. t-SNE visualization of text features for all classes on CIFAR10 test set. (a) shows the distribution of text features generated by TES, while (b) shows the learnable text features.

Different ViT fine-tuning strategies. GCD [17] and SimGCD [23] propose to build the classifier on post-backbone features instead of post-projector. Because the ViT backbone of CLIP contains a lot of knowledge learned from substantial image-text pairs, and the projector plays a role in modal alignment, it's essential to compare the effects of different ViT finetune strategies. As shown in Fig. 7, we conduct multiple evaluations with last-block fine-tuning, projector fine-tuning, and adapter [5] fine-tuning strategies. Though simply fine-tuning the projector can gain a higher accuracy across CUB and Aircraft datasets, it falls behind the last-block fine-tuning method for generic datasets.

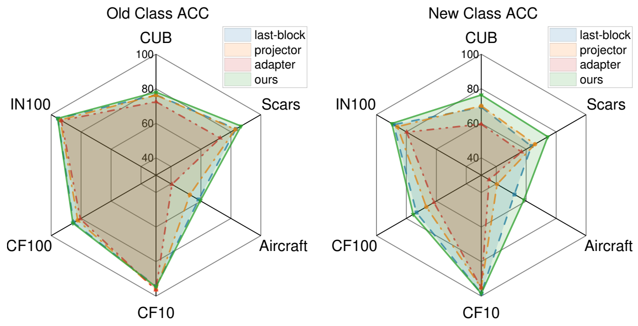

Figure 7. Different ViT finetune strategies.

Overall, our GET perfoms the best among all methods. For a fair comparison, we select the last-backbone fine-tuning strategy for baseline methods and our dual-branch multi-modal learning across all datasets except projector fine-tuning for ImageNet-1K.

Additional baseline results. As shown in Tab. 12. We provide results of PromptCAL-CLIP on three fine-grained datasets and three image classification generic datasets. For three fine-grained datasets, our method outperforms PromptCAL-CLIP on all datasets and classes. In particular, we surpass PromptCAL-CLIP by 11.5%, 4.5%, and 4.4% on 'All' classes of CUB, Stanford Cars, and Aircraft, respectively. As for the generic datasets, our method surpasses PromptCAL-CLIP on all datasets and achieves the best results on CIFAR-100 and ImageNet-100 datasets.

Error bars for main results. The experimental results presented in the paper are the averages of three independent repeated runs. We provide the performance standard deviation of our main results on all evaluation datasets with three runs in Tab. 14.

Dataset	All	Old	New
CIFAR10	97.2±0.1	94.6±0.1	98.5±0.1
CIFAR100	82.1±0.4	85.5±0.5	75.5±0.5
ImageNet-100	91.7±0.3	95.7±0.0	89.7±0.4
CUB	77.0±0.5	78.1±1.6	76.4±1.2
Stanford Cars	78.5±1.3	86.8±1.5	74.5±2.2
FGVC-Aircraft	58.9±1.2	59.6±0.6	58.5±1.8
Herbarium 19	49.7±0.4	64.5±0.8	41.7±0.8
ImageNet-1K	62.4±0.0	74.0±0.2	56.6±0.1
ImageNet-R	58.1±2.4	78.8±0.5	47.0±3.9

Table 14. The standard deviation of our method.

Results of two branches. We report the results of visual and text branches for 'All' classes across six datasets in Tab. 15. For 2 generic datasets (CIFAR10 and ImageNet-100), though the text branch does not achieve state-of-the-art performance, it still exhibits great performance. For 2 fine-grained datasets (CUB and Stanford Cars), both visual and text branches outperform previous methods by a large margin, while the visual branch performs better. For 2 challenging datasets (ImageNet-1K and ImageNet-R), both visual and text branches achieve remarkable results. Due to the challenging datasets comprising a significant number of unknown classes (ImageNet-1k dataset) or diverse visual concepts within the same class (ImageNet-R dataset), the consistency in text information for the same class contributes to the potentially higher discriminative power of the text branch, leading to better performance.

Dataset	Visual Branch	Text Branch
CIFAR10	97.2±0.1	95.1±0.0
ImageNet-100	91.7±0.3	90.1±0.1
CUB	77.0±0.5	73.6±0.8
Stanford Cars	78.5±1.3	73.1±0.6
ImageNet-1K	62.4±0.0	63.5±0.1
ImageNet-R	58.1±2.4	62.6±0.9

Table 15. The results of two branches.

We also provide the performance evolution of two branches throughout the model learning process on the CUB dataset (see in Fig. 8), the mutual promotion and fusion of the two branches resulted in excellent outcomes. In our experiments, we consistently and simply select the results from the visual branch.

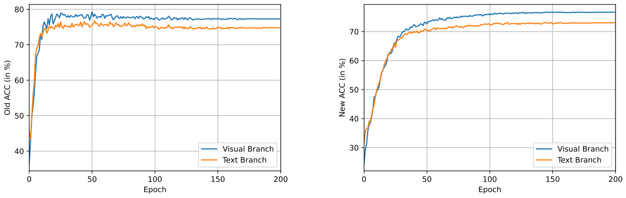

Figure 8. Performance evolution of two branches throughout the model learning process on the CUB dataset.

Results with the estimated number of classes. Vaze *et al.* [17] provides an off-the-shelf method to estimate the class number in unlabelled data. We introduce text embeddings generated by our TES into the off-the-shelf method by simply concatenating text and image features before class number Estimation. As shown in Tab. 16, multi-modal features can estimate a more accurate class, demonstrating our multi-modal method is effective in category number estimation. Following previous works, we assume the number of classes for each dataset is known and provide experimental results in the main paper. Tab. 17 shows the results using the estimated number of classes for CUB and SCars datasets.

Method	CIFAR10	CIFAR100	ImageNet-100	CUB	Stanford Cars
Ground truth	10	100	100	200	196
GCD-CLIP	5 (50%)	94 (6%)	116 (16%)	212 (12%)	234 (19%)
+TES	8 (20%)	97 (3%)	109 (9%)	212 (12%)	220 (12%)

Table 16. Estimation of class number in unlabelled data. The table shows the estimated number and the error.

Method	Known C	CUB			Stanford Cars		
		All	Old	New	All	Old	New
GET	✓	77.0	78.1	76.4	78.5	86.8	74.5
GET	✗ (w/ Est.)	75.6	75.9	75.5	76.8	87.6	71.6

Table 17. Results with the estimated number of classes

Computation complexity analysis. Tab. 18 shows the computation complexity. Our TES uses a frozen visual encoder and stage 2 finetunes the last block in another visual encoder, thus the 2 stages share the same visual encoder for the first 11 blocks but a different last block, resulting in a low computational complexity increase.

Methods	Inference Time (s/per img)	Learnable Params (M)	FLOPs (G)
SimGCD-CLIP	5.2×10^{-3}	13.4	35.2
GET(ours)	5.2×10^{-3}	15.6	38.6

Table 18. Computation complexity analysis.

The anchor prototypes. In the CICO, the anchor prototypes are calculated by averaging the features of labeled anchor samples, making them more dynamic compared to directly using the prototype classifier η. The ablation on CUB is shown in Tab. 19.

Methods	All	Old	New
use classifier η	76.3	77.6	75.6
use anchors (ours)	77.0	78.1	76.4

Table 19. The anchor prototypes.

Experiments on the Clevr-4 dataset Recently, [20] presented a synthetic dataset Clevr-4 to examine whether the GCD method can extrapolate the taxonomy specified by the labeled set. Most attributes of Clevr-4, such as shape, color, and count, are easily clustered (achieving close to 99% accuracy with CLIP). However, texture attributes pose a certain level of challenge. Therefore, we evaluate our method on the texture attributes of Clevr-4. As shown in Tab. 20, our method achieves higher accuracy and lower standard deviation compared to SimGCD-CLIP, proving that the GCD method can cluster data at specified levels based on the constraint of labeled text information, which is worthy of attention and exploration.

Methods	All	Old	New
SimGCD-CLIP	83.1±7.4	99.2±0.3	75.1±10.9
GET(ours)	90.0±1.9	99.2±0.2	85.5±2.8

Table 20. The results on Clevr-4 (Texture) in 5 runs.

	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
WordNet [13]	41.8	35.2	45.1	26.5	21.9	29.0	16.5	13.3	18.2	18.0	18.6	17.8	18.6	18.9	18.8	28.8	39.1	23.6
CC3M [15]	20.8	20.8	20.9	18.7	19.0	18.5	15.4	11.6	17.3	8.1	8.3	8.1	13.0	13.6	11.6	8.9	12.1	7.2

Table 21. Results (%) of retrieval-based approach.

Retrieval baselines. To address the challenge of missing class names, another method might involve utilizing a knowledge base of potential class names (nouns) and then using CLIP to retrieve names from this corpus. Images that share the same retrieved name could be grouped together, and clustering accuracy could then be measured based on these groupings. Therefore, this retrieval-based approach serves as an important baseline. Tab. 21 shows the results of retrieval-based approach, using WordNet [13] and CC3M [15] as corpus.

The impact of hyper-parameter λ_c. In our method, we set λ_c to 1 for all datasets to prevent over-tuning. Tab. 22 shows the ablation of the impact of λ_c.

	CUB			Stanford Cars		
λ_c	All	Old	New	All	Old	New
0.5	76.3	74.7	77.1	79.0	88.7	74.3
1	77.0	78.1	76.4	78.5	86.8	74.5
1.5	75.3	75.9	75.0	79.6	90.7	74.2
2	75.0	77.3	73.8	79.0	86.8	75.3

Table 22. The impact of hyper-parameter λ_c

11. Pseudo-code

The training procedure of the proposed GET is presented in Algorithm 1.

12. Different pre-trained models

In this section, we perform an extensive empirical investigation to explore the impact of different types of pre-trained models on GCD clustering performance, which clearly demonstrates that different types of backbones exhibit varying biases across different datasets, classes, and even paradigms. We choose DINO [1], which is based on teacher-student learning; MoCo v3 [2], based on contrastive learning; iBOT [27], based on contrastive masked image modeling; and CLIP [14], which is based on vision-language contrastive learning.

We first evaluate the results of GCD and SimGCD across different types of pretraining models. As shown in Fig. 9, different types of backbones exhibit varying biases across different datasets, classes, and even paradigms. For example, iBOT outperforms DINO in non-parametric GCD, but DINO excels in parametric GCD. MOCO demonstrates the

Algorithm 1: Pseudocode for GET.

```
Input: Training dataset $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$, a FC layer $l(\cdot|\theta_\mathrm{t})$
       and a MLP layer $g(\cdot|\theta_\mathrm{m})$, fixed CLIP's image
       encoder $E_i$ and text encoder $E_t$, a trainable image
       encoder $f_v(\cdot|\theta_\mathrm{v})$, a prototypical classifier $\eta(\cdot|\theta_\mathrm{c})$
       and a linear projection $p(\cdot|\theta_\mathrm{p})$.
Output: Predicted label $\hat{y}_i$.
/* Stage 1:  TES Training                          */
repeat
    for $(\boldsymbol{x}_i, \boldsymbol{y}_i) \in$ each batch do
        $\boldsymbol{z}_i^v = E_i(\boldsymbol{x}_i)$ // visual embedding
        $\boldsymbol{t}_i = l(\boldsymbol{z}_i^v))$ // pseudo text tokens
        $\hat{\boldsymbol{z}}_i^t = E_t(\boldsymbol{t}_i)$ // pseudo text embedding
        $\mathcal{L}_{align} \leftarrow$ Eq. (5) and Eq. (6)
        $\mathcal{L}_{distill \leftarrow}$ Eq. (7)
        $\mathcal{L}_{TES} = \mathcal{L}_{align} + \mathcal{L}_{distill}$
        Back-propagation and optimize $\theta_\mathrm{t}$.
until reaching max epochs;
/* Stage 2:  Dual-branch training                  */
repeat
    for $(\boldsymbol{x}_i, \boldsymbol{y}_i) \in$ each batch do
        /* Visual-branch                               */
        $\boldsymbol{z}_i^v = f_v(\boldsymbol{x}_i),\ \boldsymbol{h}_i^v = g(\boldsymbol{z}_i^v),\ \boldsymbol{p}_i^v = \eta(\boldsymbol{z}_i^v)$
        Compute $\mathcal{L}_{ucon}^v$ and $\mathcal{L}_{scon}^v$ by replacing $\boldsymbol{h}$ in Eq.
          (1) and Eq. (2) with $\boldsymbol{h}^v$
        $\mathcal{L}_\mathrm{rep}^v \leftarrow$ Eq. (8)
        Compute $\mathcal{L}_\mathrm{cls}^v$ by replacing $\boldsymbol{p}$ in Eq. (3) and Eq.
          (4) with $\boldsymbol{p}^v$
        $\mathcal{L}_\mathrm{db}^v \leftarrow$Eq. (9)
        /* text-branch                                 */
        $\hat{\boldsymbol{z}}_i^t = E_t(l(E_i(\boldsymbol{x}_i)))$
        $\hat{\boldsymbol{z}}_i^{tl} = p(\hat{\boldsymbol{z}}_i^t),\ \boldsymbol{h}_i^t = g(\hat{\boldsymbol{z}}_i^{tl}),\ \boldsymbol{p}_i^t = \eta(\hat{\boldsymbol{z}}_i^{tl})$
        $\mathcal{L}_\mathrm{db}^t = \mathcal{L}_\mathrm{rep}^t + \mathcal{L}_\mathrm{cls}^t$
        Compute the multi-modal mean entropy
          regularization $H_{mm}$
        /* CICO                                        */
        Calculate the visual and text anchors $\mathcal{P}_v, \mathcal{P}_t$
        Compute the instance relationships by Eq. (10)
        $\mathcal{L}_\mathrm{CICO} \leftarrow$ Eq. (11)
        $\mathcal{L}_\mathrm{Dual} \leftarrow$ Eq. (12)
        Back-propagation and optimize $\theta_\mathrm{v},\ \theta_\mathrm{m},\ \theta_\mathrm{c},\ \theta_\mathrm{p}$.
until reaching max epochs;
return $\hat{y}_i = \eta(f_v(\boldsymbol{x}_i))$.
```

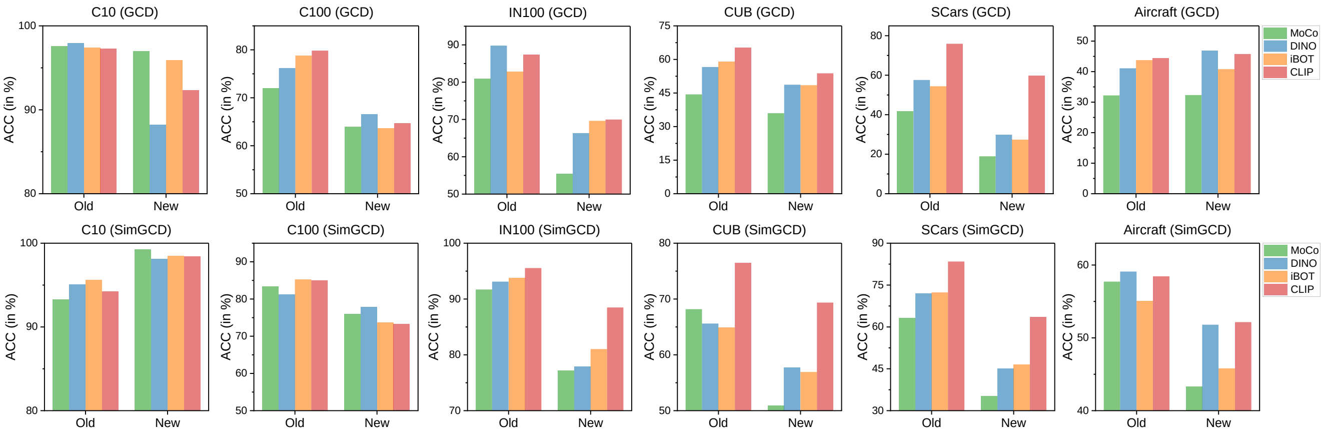

Figure 9. The results of GCD and SimGCD with different backbones across six datasets.

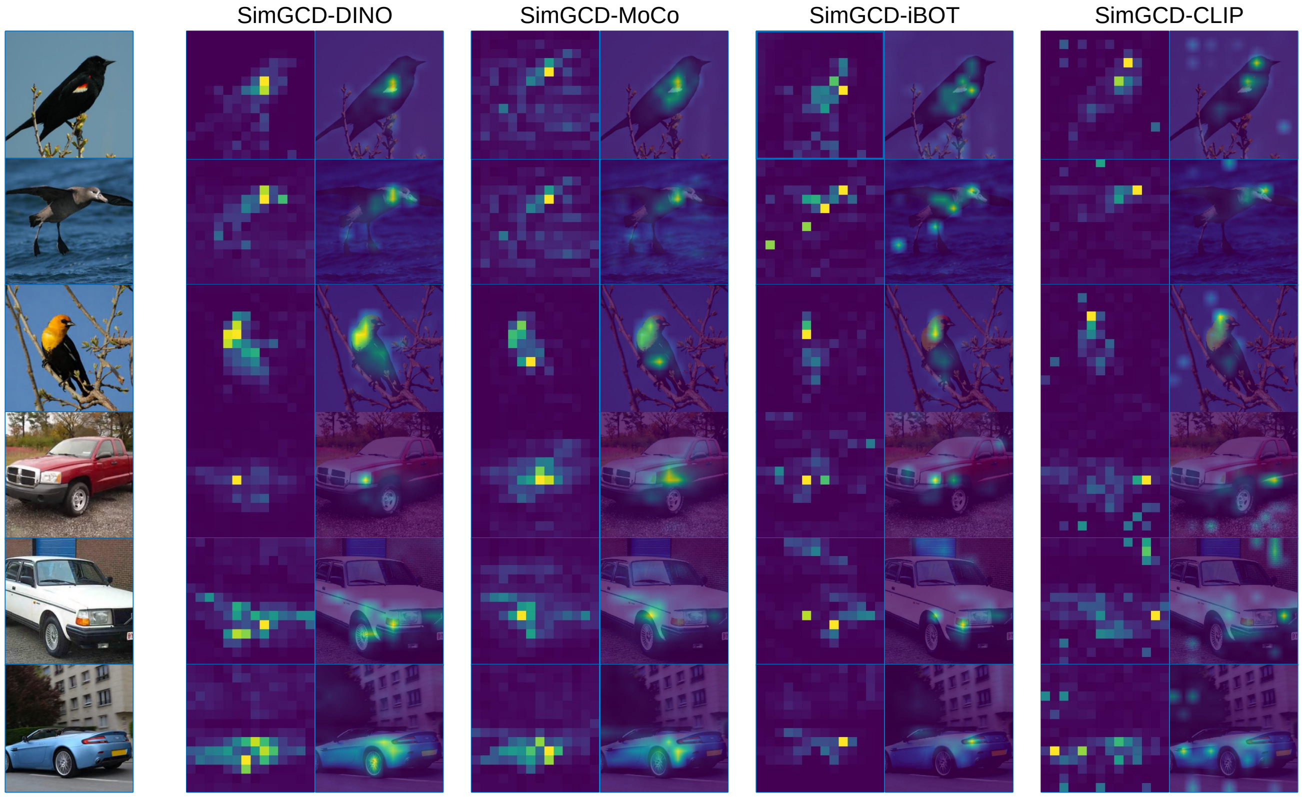

Figure 10. Attention map of class tokens on CUB (first three rows) and StanfordCars (last three rows) datasets. Each row displays the attention areas and attention maps for each image of SimGCD [23] with different backbone models.

strongest category discovery ability for the CIFAR dataset. CLIP performs exceptionally well across all datasets, yet struggles with low-resolution CIFAR data in parametric GCD.

We then visualize and compare the attention map of class tokens of different backbones in Fig. 10. For the CUB dataset, the DINO, iBOT, and MoCO backbones tend to focus more on the feathers of the birds, while CLIP additionally emphasizes the more discriminative head area. For the StanfordCars dataset, the DINO backbone focuses on the car light and wheel; the MoCo backbone focuses on the front fenders of the car, which is less discriminative; the iBOT backbone focuses on the car light and the car window, which is more discriminative than DINO thus leading to better results; the CLIP backbone focus on both the front of the car and global information, showcasing stronger discriminative capabilities.

A key observation is that though promising results have

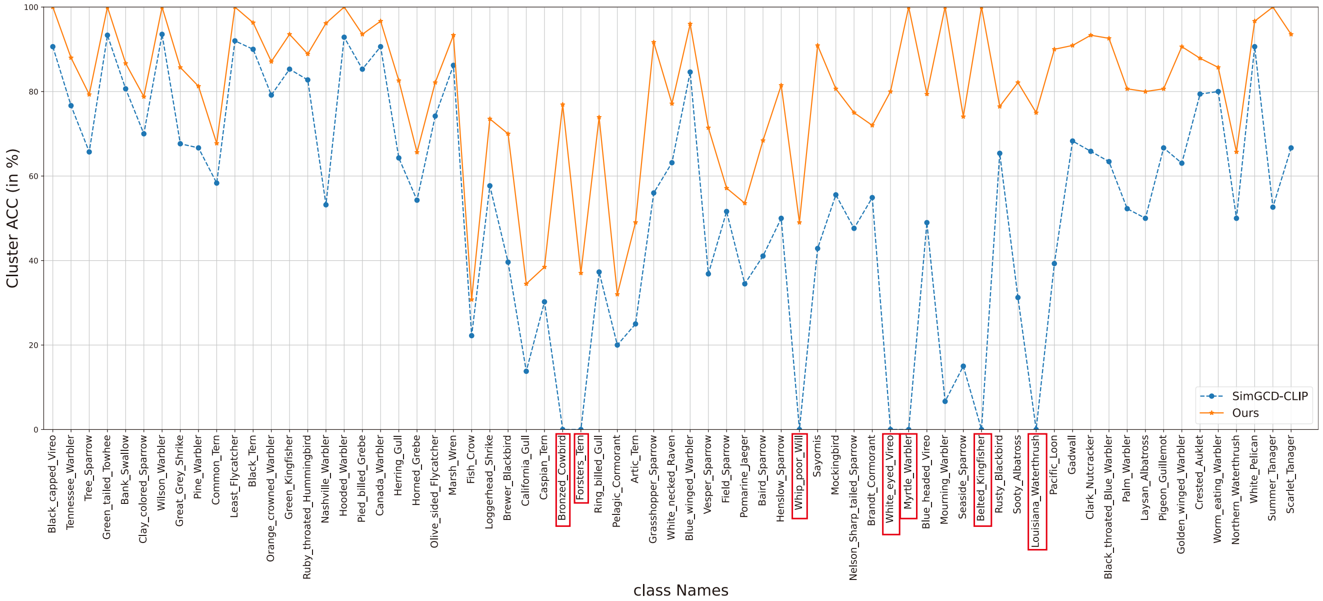

Figure 11. Cluster accuracy of SimGCD-CLIP and our GET on some visually similar classes in CUB datasets. GCD methods relying solely on a single visual modality result in empty clusters(highlighted by red boxes); Our multi-modal approach GET avoids empty clusters and achieves higher classification accuracy.

been achieved, different backbones, even powerful CLIP, still perform inferiorly on distinguishing certain visually similar classes, such as the classes in all fine-grained datasets. We argue that this is due to current methods only utilize a single visual modality of information, another modality may potentially compensate for the lack of discriminative ability. In the meanwhile, the potential of current GCD methods heavily relies on the generalization ability of pre-trained models, prompting us to select a more robust and realistic pre-training model. As a large-scale model, CLIP shows strong generalization ability on downstream tasks and strong multi-modal potential due to its image-text contrastive training, thus we decide to introduce it into the GCD task. This not only unleashes the latent potential performance of existing methods but also serves as a bridge for us to leverage multi-modal information.

13. Cluster results of GET

As shown in Fig. 11, we present the comparative cluster accuracy between our multi-modal approach and previous single-modal methods on some visually similar classes in CUB datasets. It is worth noting that relying solely on visual information, even with a powerful CLIP backbone, the previous method (SimGCD-CLIP) still struggles to differentiate some categories, resulting in empty clusters. However, leveraging the rich and discriminative text information of categories, our GET achieves more accurate classification results on CUB without any empty clusters across all categories, demonstrating the importance of multi-modal information in the GCD task. Furthermore, we showcase the clustering results of SimGCD-CLIP (see in Fig. 12) and our GET (see in Fig. 13) for the 170th class, “Mourning Warbler”, in the CUB dataset. SimGCD-CLIP relies solely on visual information to categorize birds based on shape and posture, the model categorizes many visually similar samples as “Mourning Warbler”. Our approach, by incorporating text information, enhances the model’s discriminative ability and correctly identifies all instances of the “Mourning Warbler” class, achieving 100% classification accuracy for this visually challenging category.

14. Limitations and Broader Impact

Limitations and future works. A limitation of our approach is that we treat visual and text information as equally important. In fact, some samples may have richer and more discriminative visual information than textual information, and vice versa. A more appropriate approach might involve enabling the model to adaptively leverage multimodal information, autonomously assessing which modality’s information is more crucial. We will delve deeper into this aspect in our future work.

Broader impact. Our approach introduces text information into the GCD task through a novel text embedding synthesizer module, extending the GCD to a multi-modal paradigm without extra corpus or models. We believe that the introduction of TES will encourage future research in solving GCD in a multi-modal manner.

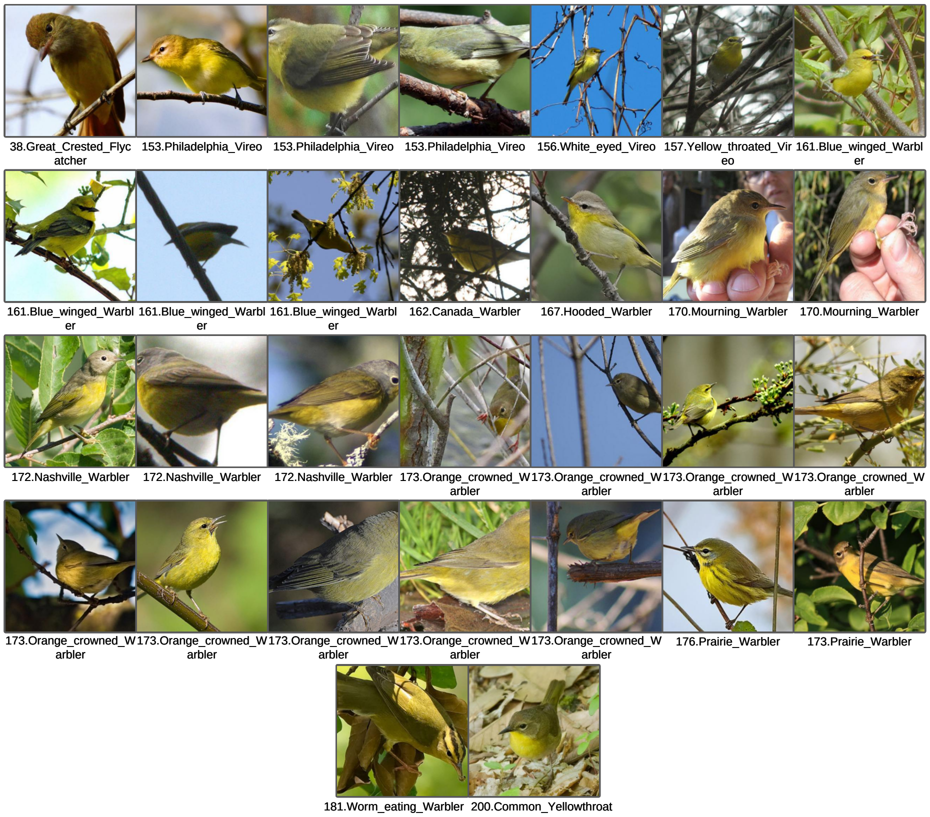

Figure 12. SimGCD-CLIP cluster results visualization for class "Mourning Warbler" in CUB dataset. SimGCD-CLIP categorizes birds based on shape and posture and incorrectly identifies many visually similar categories, resulting in a clustering accuracy of 6.7% for class "Mourning Warbler".

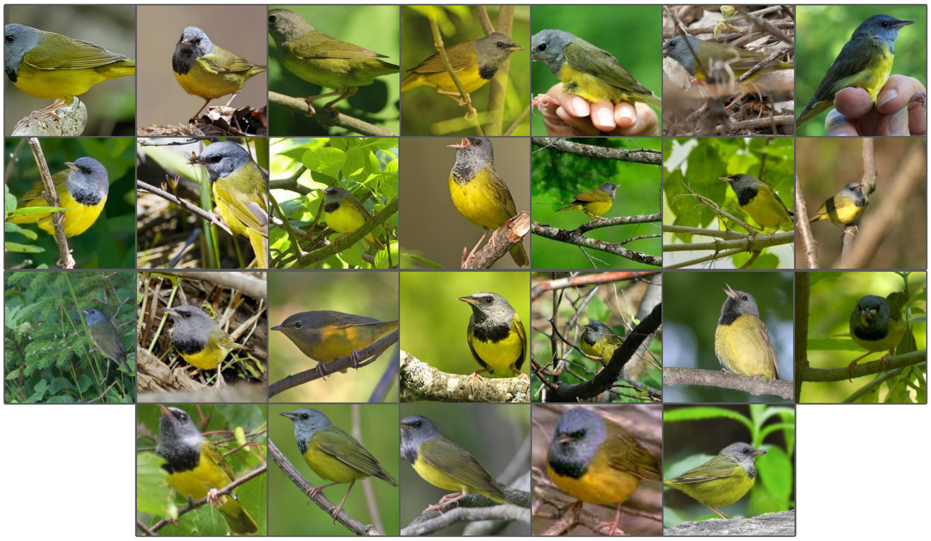

Figure 13. Cluster results visualization for class "Mourning Warbler" in CUB dataset of our GET. Our method uses multi-model information, achieving 100% classification accuracy for this visually challenging category.

References

[1] 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. 5

[2] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers, 2021. 5

[3] 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. 1

[4] 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. 1

[5] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 3

[6] 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. 1

[7] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue, 2018. 2

[8] 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. 1

[9] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical report*, 2009. 1

[10] Yu Liu, Yaqi Cai, Qi Jia, Binglin Qiu, Weimin Wang, and Nan Pu. Novel class discovery for ultra-fine-grained visual categorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17679–17688, 2024. 2

[11] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 1

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

[13] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 1995. 5

[14] 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. 1, 5

[15] 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. 5

[16] 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. 1

[17] 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, 4

[18] 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. 1

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

[20] Sagar Vaze, Andrea Vedaldi, and Andrew Zisserman. No representation rules them all in category discovery. *Advances in Neural Information Processing Systems 37*, 2023. 4

[21] 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. 1

[22] 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

[23] 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, 6

[24] Xiaohan Yu, Yang Zhao, Yongsheng Gao, Xiaohui Yuan, and Shengwu Xiong. Benchmark platform for ultra-fine-grained visual categorization beyond human performance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10285–10295, 2021. 2

[25] 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. 3

[26] 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. 1

[27] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 5