Multimodal Generalized Category Discovery

Supplementary Material

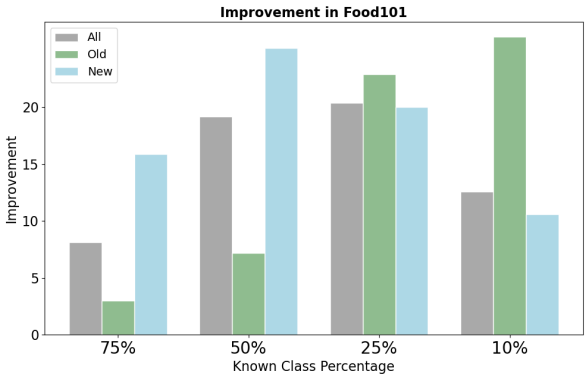

Figure 1. Improvements over baseline under various proportions of known categories on Food101.

1. Robustness to Old-New Category Proportion

In real-world applications, the proportion of new categories is often much larger than that of known categories, posing a significant challenge for traditional models. To address this, we further analyze the robustness and generalization capability of our method under different problem settings. We evaluate different proportions of known categories, i.e., the ratio of labeled and unlabeled categories, including 75%, 50%, 25%, and 10%. We compare our approach against existing baseline [3], using fusion feature for classification. Improvements are shown in Figure 1.

It is evident that our method consistently outperforms the baseline across all proportions. The enhancement is particularly noticeable in the classification of new categories, including the very challenging scenarios with more new categories and fewer labeled data. Our approach significantly mitigates the tendency of the GCD approaches to overfit to old categories.

2. Unknown Category Number Estimation

Here we evaluate the GCD task performance with unknown category numbers. In previous experiments, we adopted the settings from previous work [1–4], which assumed that the number of new classes was known. This assumption can sometimes be difficult to meet in real-world scenarios. Therefore, it is necessary to first estimate the number of categories and then proceed with training based on the estimated number.

We estimate the number of new categories by following the off-the-shelf provided in original GCD [1]. Specifically, we utilize a pretrained model to generate corresponding features of a dataset without finetuning on that dataset. We then performe k-means clustering on these features with different numbers of categories **K**. The clustering result with the highest label accuracy is our estimated number of categories. As we are engaged in a multimodal classification task, we conducted this clustering for both vision and textual features, with the results presented in Table 1a. We found that a simple concatenation of features can significantly enhance the accuracy of initial category estimation. This demonstrates that multimodal information processed by multimodal models like CLIP can provide relatively more accurate clustering under initial conditions.

We then train the model using the number of categories esimated based on the fusion features. The results are shown in Table 1b. Our method consistently shows higher classification accuracy, demonstrating its robustness and generalization ability for real-world scenarios.

3. Comparison of Fusion Features and Soft Voting

We compared the results of using soft voting, where the logits from text and image predictions are summed to select the final label, with those from using fusion features for prediction. We found that fusion features consistently outperformed soft voting. Detailed results are in Table 2. We believe this is because fusion features more effectively integrate and refine semantic information from different modalities, offering richer context. In contrast, soft voting simply combines logits without fully capturing the correlations and complementarities between modalities, leading to reduced robustness and generalization on complex samples.

4. More Visualization of Attention Maps

Figure 2 presents additional examples from the Food101 dataset. In contrast to the baseline models, our model is shown to focus on the key and distinctive features of the food, rather than being influenced by irrelevant background noise. This demonstrates that our model not only makes more accurate predictions, but also provides a more interpretable and reliable decision-making process, which is particularly valuable in specialized fields like medicine and finance.

5. More Visualization Comparison

Table 4 presents additional visualization results from the Food101 dataset. The samples displayed below are cases where SimGCD made incorrect predictions, while our

	Food101	N24News
Image	121	13
Text	165	14
Fusion	**113**	**18**
GT	*101*	*24*

(a) Estimated unknown class number.

Methods	$\|\mathcal{C}\|$	Food101 All	Food101 Old	Food101 New	N24News All	N24News Old	N24News New
SimGCD	*GT*	73.1	85.7	66.8	63.8	**75.1**	57.1
Ours	*GT*	**92.3**	92.9	**92.0**	**68.5**	74.8	**64.5**
SimGCD	*Estimated*	75.6	75.7	75.5	60.9	75.0	52.4
Ours	*Estimated*	91.5	**93.0**	90.7	63.7	73.2	58.0
Δ		**+15.9**	**+17.3**	**+15.2**	**+2.8**	**-1.8**	**+5.6**

(b) Comparison of SimGCD and our approach.

Table 1. GCD task performance with unknown category numbers.

Methods	Food101 All	Food101 Old	Food101 New	N24News All	N24News Old	N24News New
Visual	85.5	92.5	82.0	64.8	69.2	62.6
Text	85.3	92.3	81.9	65.1	68.9	62.8
Voting	87.5	90.5	85.9	67.8	72.1	**65.3**
Fusion	**92.3**	**92.9**	**92.0**	**68.5**	**74.8**	64.5

Table 2. A comparison of different ways for final prediction.

method correctly identified the categories. These examples involve significant noise in both text and images, making accurate classification challenging. The baseline approach struggles in such scenarios due to limited single-modality information processing. In contrast, our multimodal approach better integrates information across modalities, leading to more robust predictions. This highlights the advantages of leveraging complementary signals from both text and images in challenging conditions.

6. Hyperparameter Analysis

τ **(contrastive temperature)**	0.02	0.04	0.06
Accuracy	74.5	85.2	87.5
λ_s **(supervised weight)**	0.1	0.35	0.6
Accuracy	84.8	85.2	73.0
w_i **(image weight)**	0.5	1	1.5
Accuracy	88.1	85.2	79.0
w_t **(text weight)**	0.5	1	1.5
Accuracy	87.1	85.2	77.4

Table 3. Hyperparameter tuning results on Food101 dataset subset

We have conducted hyperparameter tuning on a 10% subset of the Food101 dataset. Table 3 shows our model's strong robustness across a range of settings, with consistently high performance. Given the substantial gains over unaligned methods, we prioritized alignment evaluation over fine-tuning every parameter.

References

[1] 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 (CVPR)*, pages 7492–7501, 2022. 1

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

[3] 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 (ICCV)*, pages 16590–16600, 2023. 1

[4] 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

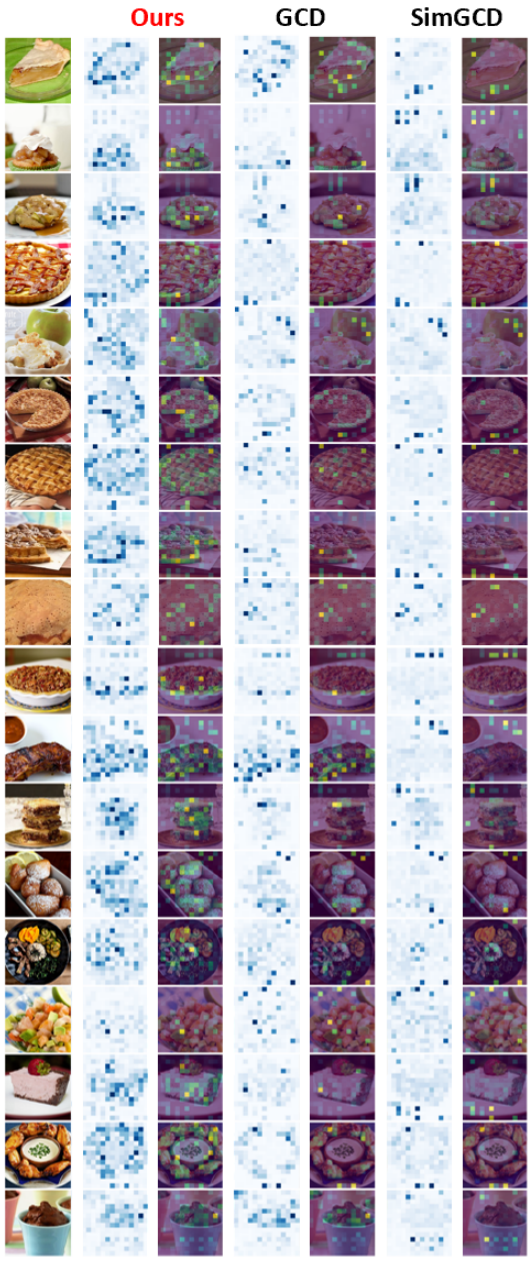

Figure 2. More visualization of attention map

Image	Label	Text
	Apple Pie	top thanksgiving recipes for every family dinner
	Apple Pie	day homemade apple pie recipe imperfectly polished
	Apple Pie	apple pie bourbon sweet tea recipe myrecipes com
	Beef Carpac-cio	beef carpaccio with apple and arugula salad recipe by chef billy parisi ifood tv
	Beef Carpac-cio	kettler cuisine october
	Beef Carpac-cio	lime marinated beef carpaccio with black olive tapenade and sun dried tomato pesto on foodrhythms
	Beef Tartare	the world most recently posted photos of yukhoe flickr hive mind
	Beef Tartare	pairing of the day october november
	Beef Tartare	cruise dining much

Table 4. More visualization results on Food101. SimGCD predicts the category of these samples wrongly while our method predicts it correctly.