

Multimodal Generalized Category Discovery

Yuchang Su¹ Renping Zhou¹ Siyu Huang² Xingjian Li³
 Tianyang Wang⁴ Ziyue Wang³ Min Xu^{3*}

¹Independent Researcher ²Clemson University

³Carnegie Mellon University ⁴University of Alabama at Birmingham

{syccc142857, tehaji007, lixj04}@gmail.com, siyuh@clemson.edu,
 tw2@uab.edu, ziyuew2@andrew.cmu.edu, mxul@cs.cmu.edu

Abstract

Generalized Category Discovery (GCD) aims to classify inputs into both known and novel categories, a task crucial for open-world scientific discoveries. However, current GCD methods are limited to unimodal data, overlooking the inherently multimodal nature of most real-world data. In this work, we extend GCD to a multimodal setting, where inputs from different modalities provide richer and complementary information. Through theoretical analysis and empirical validation, we identify that the key challenge in multimodal GCD lies in effectively aligning heterogeneous information across modalities. To address this, we propose MM-GCD, a novel framework that aligns both the feature and output spaces of different modalities using contrastive learning and distillation techniques. MM-GCD achieves new state-of-the-art performance on the UPMC-Food101 and N24News datasets, surpassing previous methods by 11.5% and 4.7%, respectively.

1. Introduction

Generalized Category Discovery (GCD) [24] aims to classify new inputs into both known and unknown classes, making it particularly valuable for open-world scientific discoveries. For instance, GCD has the potential to analyze genetic data from patients and discover new variants associated with rare diseases, revealing novel disease categories that extend beyond known classifications [3]. This capability is of great interest in rare disease research.

However, existing research [1, 26] has predominantly focused on unimodal data, neglecting the inherently multimodal nature of real-world scenarios. For example, radiology images (e.g., X-rays, CT scans) are frequently paired with clinical reports, each contributing unique information.

Leveraging this rich, multimodal data enhances decision-making accuracy, similar to how radiologists use both images and clinical reports for more accurate diagnoses.

In this work, we extend GCD to the multi-modal setting, where inputs from different modalities are simultaneously used for the classification process (Figure 1). Multimodal GCD presents unique challenges compared to its unimodal counterpart: how can we effectively utilize the rich yet heterogeneous information from different modalities? We find that simply extending existing GCD frameworks to multimodal settings is subject to degraded performance, as the information across modalities is often poorly aligned, making the model biased to spurious correlations and irrelevant information. In §4, we provide theoretical analysis and empirical validation showing that better-aligned multimodal data reduces variance and simplifies the decision boundary, improving classification performance.

Building on these insights, we propose a novel framework, MM-GCD, that directly addresses the alignment challenge in multimodal GCD. We divide the GCD learning process into two stages: the first is generating the feature space, and the second is partitioning it, which corresponds to the results in the output space. Theoretical analysis has already proven that alignment is indispensable in the feature space. We resort to recent advances in multimodal contrastive learning [22] to ensure that features from different modalities are properly aligned. The goal of multimodal contrastive learning is to bring similar concepts from different modalities closer while pushing dissimilar concepts apart. This alignment results in similar concepts from different modalities having similar features, making them easier for models to understand. Additionally, given the classification nature of multimodal GCD, in the output space, we leverage distillation techniques to ensure that classification results are consistent across modalities. Distillation allows predictions from one modality to serve as targets for the other, combined with entropy minimization to ensure

*Corresponding author.

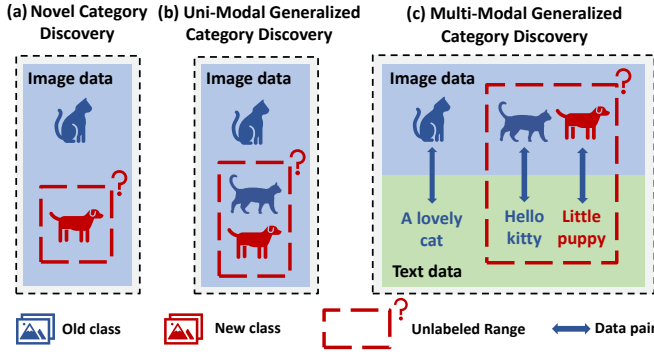


Figure 1. Evolution from NCD to Multimodal GCD. (a) Novel Category Discovery (NCD) dealt with unlabelled images containing only new classes. (b) Generalized Category Discovery (GCD) expanded this by including possible old classes in the unlabelled set but was limited to single modality data. (c) Our multimodal-GCD model addresses these limitations by focusing on multimodal data which is now abundantly present in real life, and leveraging inter-modality interactions to improve learning where labels are missing.

consistency between modalities. By aligning both the feature and output spaces, MM-GCD effectively integrates the heterogeneous information from different modalities, providing an effective solution to the GCD task.

We validate the effectiveness of MM-GCD on two benchmarks and perform ablation studies on its components. MM-GCD sets a new state-of-the-art on the UPMC-Food-101 [27] and N24News [28] datasets, surpassing previous methods by 11.5% and 4.7%, respectively. Leveraging both modalities leads to improvements of 6.8% and 3.4% over single-modality approaches, highlighting that different modalities provide complementary and enriched information for the GCD task. Visualizations of the feature space confirm that our alignment objectives effectively close the modality gap and synchronize the feature spaces. Ablation studies show that, without these alignments, performance drops by 18.8% on Food101, falling below even the single-modality baseline.

In summary, our work makes the following contributions:

- We introduce the multimodal Generalized Category Discovery setting, which closely mirrors real-world scenarios where data naturally exist in multiple modalities.
- We theoretically and empirically demonstrate that modality alignment is the most critical aspect for successfully tackling the multimodal Generalized Category Discovery problem.
- We propose novel and effective alignment methods that solve this problem, achieving new state-of-the-art performance on the Food101 and N24News datasets, with improvements of 11.5% and 4.7%, respectively.

2. Related Works

Multimodal Learning aims to integrate features across multiple modalities, such as audio, text, and images, to enhance task performance [4, 17]. The primary focus in this field is on fusion strategies, categorized into aggregation-based and alignment-based methods [4, 33]. Aggregation methods, like feature concatenation, are popular [2, 18, 28], while alignment-based methods, such as contrastive learning, align and fuse modalities by pre-training on large datasets [22, 30]. These approaches leverage complementary strengths from different modalities, improving classification accuracy and robustness [4].

Novel Class Discovery (NCD) tackles the identification and categorization of new, unseen classes within unlabeled datasets. Early work, such as [9], introduced deep clustering algorithms that significantly improved novel class discovery in image datasets. These methods inspired further research, extending to multimodal settings with cross-modal discrimination and hierarchical frameworks [10, 13]. Despite these advances, NCD methods are not designed for GCD tasks, which require classification of both knowns and unknowns.

Generalized Category Discovery (GCD) requires the classification of both seen and unseen categories within unlabeled data [24]. Methods like SimGCD [29], PromptCAL [31], DPN [1] and SPTNet [26] have advanced GCD by addressing biases towards known classes and improving learning phase interactions. Subsequent studies have further extended the application scenarios of GCD, including its integration into continuous learning [6, 15, 20], federated learning [21], and active learning [16]. However, these approaches are primarily unimodal, focusing on either visual or textual data, and do not leverage the full potential of multimodal information.

Some recent efforts have explored leveraging information from different modalities to enhance classification performance, typically employing strategies designed to generate multimodal content from unimodal inputs [19, 25, 32]. While these methods improve unimodal data interpretation, they fail to fully exploit the rich multimodal data, since their text data is synthesized or selected from certain database. In real-world scenarios, modality correlations are lower, yet the information is richer. Our method effectively harnesses this diverse data while minimizing noise by aligning and integrating multimodal information.

3. Problem Formulation

In this work, we introduce the novel problem of *Multimodal Generalized Category Discovery (Multimodal GCD)*. The goal of Multimodal GCD is to identify both known and unknown categories from multimodal data, where the model has access to labeled data for some categories and must discover new categories in unlabeled data. This problem

$$\mu_{F_k} = \begin{pmatrix} \mu_{X_k} \\ \mu_{Y_k} \end{pmatrix}, \mathcal{S}_F = \begin{pmatrix} \mathcal{S}_{X_k} & \mathcal{S}_{X_k Y_k} \\ (\mathcal{S}_{X_k Y_k})^T & \mathcal{S}_{Y_k} \end{pmatrix} \mathcal{S}_{X_k Y_k} \quad (3)$$

$$\mathcal{S}_{X_k Y_k} = \mathcal{S}_{X_k}^{\frac{1}{2}} R_k \mathcal{S}_{Y_k}^{\frac{1}{2}} \quad (4)$$

$$R_k = \begin{pmatrix} \sigma_{X_k^{(1)}, Y_k^{(1)}} & & \\ & \sigma_{X_k^{(2)}, Y_k^{(2)}} & \\ & & \dots \\ & & & \sigma_{X_k^{(n)}, Y_k^{(n)}} \end{pmatrix} \quad (5)$$

R_k is a diagonal matrix composed of the correlation coefficients, where $\sigma_{X_k^{(i)}, Y_k^{(i)}} \in (0, 1)$ converges to 1 when the distributions of the two modalities are perfectly correlated.

Here we use the determinant of the covariance matrix of a distribution as the objective function $\mathcal{L}$ to measure the distribution of the data:

$$\mathcal{L}_X = \sum_{k \in \mathcal{C}} |\mathcal{S}_{X_k}|, \mathcal{L}_Y = \sum_{k \in \mathcal{C}} |\mathcal{S}_{Y_k}|, \mathcal{L}_F = \sum_{k \in \mathcal{C}} |\mathcal{S}_{F_k}| \quad (6)$$

The determinant can be considered as a measure of volume in high-dimensional space, the smaller this objective function, the more compact the within-class distribution of the data, such that the data is potentially easier to be clustered by clustering algorithms.

We then give the relationship between the objective function of the fused modality and the original distributions of the two modalities, as:

$$\mathcal{L}_F = \sum_{k \in \mathcal{C}} |\mathcal{S}_{X_k}| \cdot |\mathcal{S}_{Y_k} - \mathcal{S}_{X_k}^T \mathcal{S}_{X_k}^{-1} \mathcal{S}_{X_k Y_k}| \quad (7)$$

$$= \sum_{k \in \mathcal{C}} |\mathcal{S}_{X_k}| \cdot |\mathcal{S}_{Y_k} - \left(\mathcal{S}_{X_k}^{\frac{1}{2}} R_k \mathcal{S}_{Y_k}^{\frac{1}{2}} \right)^T \mathcal{S}_{X_k}^{-1} \left(\mathcal{S}_{X_k}^{\frac{1}{2}} R_k \mathcal{S}_{Y_k}^{\frac{1}{2}} \right)| \quad (8)$$

$$= \sum_{k \in \mathcal{C}} |\mathcal{I} - R_k^2| \cdot |\mathcal{S}_{X_k}| \cdot |\mathcal{S}_{Y_k}| \quad (9)$$

From the final result, it can be seen that $\mathcal{L}_F$ is negatively correlated with R_k . This means that as R_k becomes larger, and the two distributions become closer, $\mathcal{L}_F$ decreases, increasing the compactness of the joint distribution. This suggests that even if the individual results of the two distributions do not improve, i.e., when $\mathcal{L}_X$ and $\mathcal{L}_Y$ remain unchanged, we can still make the joint distribution more distinguishable by improving the alignment of the modalities.

We conducted experiments to prove our theory: using the baseline method, we sorted the test data based on the cosine similarity of cross-modality features between the text and visual and subsequently divided into ten groups. As shown in Fig. 3, the accuracy within each group revealed

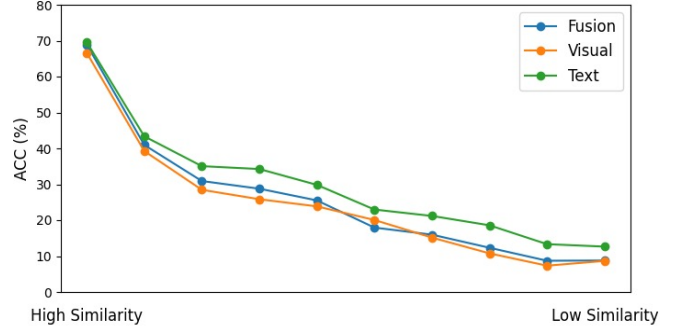


Figure 3. The relationship between accuracy and feature similarity across visual and text modalities. The result shows that groups with higher feature similarity tend to achieve greater accuracy.

that groups with higher cross-modality feature similarity exhibited higher accuracy. This finding suggests that there is a strong correlation between the alignment of features across modalities and the resulting accuracy.

5. Method

Through theoretical derivation and experimental validation, we have identified that the key to the problem is the alignment between modalities. With this objective in mind, we have reconsidered the goal of the loss function in the uni-modal setting and incorporated the goal of inter-modal alignment into it.

5.1. What is required to discover a new category?

The requirements for the Generalized Category Discovery (GCD) problem can be divided into two key aspects: *creating an embedding space that can differentiate each category* and *effectively partitioning this embedding space to include new categories*. These aspects correspond to the goals of representation learning [24] and parametric classification [29] in previous methods. In our approach, we incorporate the objective of modality alignment into the learning processes of these two components. In §5.2, we focus on aligning the features of different modalities within the embedding space, while in §5.3, we address the unified partitioning of this space.

Let us define the basic notations used in this section. Given a sample $\mathbf{x}_i$ from one modality and its augmented view $\mathbf{x}'_i$, along with the corresponding data $\mathbf{y}_i$ from another modality within the same data pair, let $\mathbf{h}_i$ denote the feature representation of $\mathbf{x}_i$ after passing through the backbone encoder f_x , such that $\mathbf{h}_i = f_x(\mathbf{x}_i)$. Similarly, let $\tilde{\mathbf{h}}_i$ represent the feature obtained from $\mathbf{y}_i$ after passing through the backbone encoder f_y , such that $\tilde{\mathbf{h}}_i = f_y(\mathbf{y}_i)$. Additionally, let $\mathbf{z}_i$ represent the output of a projection layer g applied to $\mathbf{h}_i$, followed by normalization, such that $\mathbf{z}_i = \text{norm}(g(\mathbf{h}_i))$.

5.2. Constructing Embedding Space

To build a robust embedding space across different modalities, representation learning with contrastive loss is widely employed. A critical aspect of this process is the selection of positive pairs. In the multimodal setting, we define three types of positive pairs.

The first type involves an unsupervised contrastive loss between two views $\mathbf{x}_i$ and $\mathbf{x}'_i$ of the same sample within a mini-batch B :

$$\mathcal{L}_{\text{rep}}^u = \frac{1}{|B|} \sum_{i \in B} -\log \frac{\exp(\mathbf{z}_i^\top \mathbf{z}'_i / \tau_u)}{\sum_{j \neq i} \exp(\mathbf{z}_i^\top \mathbf{z}'_j / \tau_u)}. \quad (10)$$

The second type introduces a supervised contrastive loss between a sample and other samples with the same label within a labeled mini-batch B^l :

$$\mathcal{L}_{\text{rep}}^s = \frac{1}{|B^l|} \sum_{i \in B^l} \frac{1}{|\mathcal{N}_i|} \sum_{r \in \mathcal{N}_i} -\log \frac{\exp(\mathbf{z}_i^\top \mathbf{z}'_r / \tau_s)}{\sum_{j \neq i} \exp(\mathbf{z}_i^\top \mathbf{z}'_j / \tau_s)}. \quad (11)$$

where $\mathcal{N}_i$ is the sample set that shares the same label as x_i .

The third type applies a cross-modal contrastive loss, aimed at aligning features across modalities. Here, the positive pairs extend across different modalities within the same data pair, bringing their respective embedding spaces closer. Given that the encoders from CLIP are already trained for alignment, we use the features $\mathbf{h}_i$ directly from the encoders for this cross-modal contrastive loss, bypassing the additional projection layer:

$$\mathcal{L}_{\text{rep}}^c = \frac{1}{|B|} \sum_{i \in B} -\log \frac{\exp(\mathbf{h}_i^\top \tilde{\mathbf{h}}_i / \tau_c)}{\sum_{j \neq i} \exp(\mathbf{h}_i^\top \tilde{\mathbf{h}}_j / \tau_c)}. \quad (12)$$

The temperatures τ_u , τ_s , and τ_c control the scaling in each level of contrastive learning. We balance the different representation losses using coefficients λ_u and λ_s . The overall representation learning objective is then defined as:

$$\mathcal{L}_{\text{rep}} = \lambda_u \mathcal{L}_{\text{rep}}^u + \lambda_s \mathcal{L}_{\text{rep}}^s + (1 - \lambda_u - \lambda_s) \mathcal{L}_{\text{rep}}^c. \quad (13)$$

5.3. Partitioning Embedding Space

After constructing the embedding space, it is crucial to effectively partition this space for classification, especially when dealing with both known and unknown categories. Previous work [29] has demonstrated that parametric methods for partitioning can yield superior results. Building on this insight, we develop a trainable prototypical classifier, denoted as $\mathcal{C} = \{\mathbf{c}_1, \dots, \mathbf{c}_K\}$, where K represents the total number of categories, including both old and new classes. For each sample $\mathbf{x}_i$ in a mini-batch, we compute a pseudo-probability p_i over all categories:

$$p_i^{(k)} = \frac{\exp(\mathbf{h}_i^\top \mathbf{c}_k / \tau_p)}{\sum_{k'} \exp(\mathbf{h}_i^\top \mathbf{c}_{k'} / \tau_p)}. \quad (14)$$

Similar to the representation learning section, we also present three levels of classification loss. The first loss is self-distillation of all samples. We produce an additional soft pseudo-label $\mathbf{q}_i$ that is generated from a different view of $\mathbf{x}_i$ using a more acute temperature τ_q . We minimize the cross-entropy loss $\ell(\mathbf{q}, \mathbf{p}) = -\sum_k \mathbf{q}^{(k)} \log \mathbf{p}^{(k)}$ between the predicted probabilities and the pseudo-labels for all samples in a mini-batch B :

$$\mathcal{L}_{\text{cls}}^u = \frac{1}{|B|} \sum_{i \in B} \ell(\mathbf{q}_i, \mathbf{p}_i) \quad (15)$$

The second is towards all the labeled set B_l . Using this existing supervision, we can quickly correct the positions of some old class prototypes, thereby creating a clearer partition of the entire space:

$$\mathcal{L}_{\text{cls}}^s = \frac{1}{|B|} \sum_{i \in B} \ell(\mathbf{y}_i, \mathbf{p}_i) \quad (16)$$

where $\mathbf{y}_i$ is the actual label of $\mathbf{x}_i$.

The third is multimodal prototype distillation that we propose. It effectively leverages the pairwise information among modalities in a self-distillation manner: each modality distills the classification decisions made by models of other modalities. It ensures that samples in a pair occupy analogous positions in the latent image and text spaces, facilitating the clustering/prototyping process that plays a key role in the the problem:

$$\mathcal{L}_{\text{cls}}^c = \frac{1}{|B|} \sum_{i \in B} \ell(\tilde{\mathbf{p}}_i, \mathbf{p}_i) \quad (17)$$

where $\tilde{\mathbf{p}}_i$ is the pseudo probabilities of the corresponding data y_i . By strengthening the decision consistency among modalities, the resulting output become more robust to cross-modal distributional discrepancies, thereby enhancing the clustering/prototyping process in multimodal latent spaces.

Since the classification results for unknown classes from different modalities may be inconsistent (e.g., assigning the same new class to different clusters), we applied the Hungarian algorithm [12] for label mapping to determine the optimal classification correspondence before calculating the entropy loss.

A mean-entropy regularization term $H(\bar{\mathbf{p}})$ is additionally adopted to add balance of the classification result as a learning objective, where $\bar{\mathbf{p}} = \frac{1}{2|B|} \sum_{i \in B} (\mathbf{p}_i + \mathbf{p}'_i)$ signifies the average prediction for a batch, and the entropy $H(\bar{\mathbf{p}}) = -\sum_k \bar{\mathbf{p}}^{(k)} \log \bar{\mathbf{p}}^{(k)}$ measures the uncertainty of this prediction. The overall parametric classification objective is:

$$\mathcal{L}_{\text{cls}} = \lambda_u \mathcal{L}_{\text{cls}}^u + \lambda_s \mathcal{L}_{\text{cls}}^s + (1 - \lambda_u - \lambda_s) \mathcal{L}_{\text{cls}}^c + \epsilon H(\bar{\mathbf{p}}). \quad (18)$$

Dataset	Labelled		Unlabelled	
	#Data	#Class	#Data	#Class
UPMC-Food101 [27]	15K	50	45K	101
N24News [28]	12K	12	36K	24

Table 1. Statistics of the multimodal classification benchmarks used for evaluation.

Dataset	Image	Text	Label
N24News		Ms. Goodman styled Amber Valletta with wings...	Style
N24News		A scene from Meg Stuart's "Until Our Hearts Stop"...	Dance
Food101		Mind Blowingly Awesome Vegan Pizza	Pizza
Food101		Korean Sushi Rolls with Walnut-Edamame Crumble	Sushi

Table 2. Data examples for each dataset.

We also found that compared to using voting methods based on classification results from a single modality, using a fusion modality, where features from two modalities are concatenated and then passed through a linear layer, can better capture the interaction information between modalities, thereby achieving better classification performance. Detailed comparison is shown in Appendix. The loss mentioned below is the same as previously described, except that the features are replaced with the new fusion features:

$$\mathcal{L}_{\text{fusion}} = \lambda_u(\mathcal{L}_{\text{rep}}^u + \mathcal{L}_{\text{cls}}^u) + \lambda_s(\mathcal{L}_{\text{rep}}^s + \mathcal{L}_{\text{cls}}^s). \quad (19)$$

And, the overall objective for multimodal GCD learning is:

$$\mathcal{L} = \mathcal{L}_{\text{rep}} + \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{fusion}}. \quad (20)$$

6. Results

6.1. Experiments Setup

Datasets.

We evaluate our approach on two image-text datasets: UPMC-Food-101 [27] and N24News [28]. UPMC-Food-101 includes recipe descriptions paired with food images across 101 categories. N24News contains news images and four textual components, from which we use the Abstracts for pairing. In both datasets, we selected 50% of the categories as old classes, and from those, we picked out 50% of the samples to form $\mathcal{D}^l$, with the remainder serving as $\mathcal{D}^u$.

Methods	Sup	Modal	Food101			N24News		
			All	Old	New	All	Old	New
<i>MMBT</i> [11]	51	Fusion	92.1	-	-	-	-	-
<i>CMA-CLIP</i> [14]	51	Fusion	93.1	-	-	-	-	-
<i>UniS-MMC</i> [33]	51	Fusion	94.7	-	-	84.7	-	-
GCD [24]	55	Visual	40.0	68.7	25.3	27.8	30.1	27.0
		Text	68.7	79.5	63.2	33.1	43.3	26.9
		Fusion	61.7	75.5	54.8	32.8	44.2	26.0
SimGCD [29]	55	Visual	65.5	75.0	60.8	44.7	54.0	39.2
		Text	80.8	84.3	79.1	62.0	75.2	54.0
		Fusion	73.1	85.7	66.8	63.8	75.1	57.1
SPTNet [26]	55	Visual	64.5	71.4	61.0	44.1	53.5	38.5
DPN [1]	55	Text	75.4	81.3	69.7	59.5	69.7	50.7
Ours	55	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
		Fusion	92.3	92.9	92.0	68.5	74.8	64.5

Table 3. Results on the UPMC-Food101 Dataset [27] and N24News Dataset [28].

The specific divisions are detailed in Table 1.

Evaluation Protocol. We adopted the evaluation system from the original GCD work [24]. For the unlabeled data during training, classification accuracy is calculated by comparing the model-predicted labels with the true labels. The optimal match between predicted and ground truth labels is determined using the Hungarian algorithm [12]. In addition to overall accuracy on the entire $\mathcal{D}^u$ set, we also measured accuracy separately for old and new classes, referred to as **All**, **Old**, and **New** accuracy.

$$ACC = \max_{p \in P(\mathcal{Y}^u)} \frac{1}{M} \sum_{i=1}^M 1(y_i = p(\hat{y}_i)) \quad (21)$$

Implementation Details. To support multimodal representations, we adopted the popular CLIP [22] model with ViT B/16 as the backbone. It is worth noting that, previous GCD papers [24, 29] used ViT B/16 [8] pre-trained with DINO [5] over ImageNet [7] as the backbone. To ensure a fair algorithmic comparison, we replicated those baselines with the CLIP backbone. We fine-tune the last transformer block of image and text encoder as well as the projection head. We employed a learning rate initially set to 0.1, with a cosine schedule, a batch size of 128, and trained for 200 epochs. Additionally, we adopted the practice from previous work of setting the supervised weight to 0.35.

6.2. Main Results

Due to the lack of publicly available multimodal semi-supervised models, we compared the state-of-the-art methods in current unimodal tasks and extended two of them (GCD [24] and SimGCD [29]) to the multimodal setting. Specifically, we used CLIP’s dual encoders to extract image and text features separately, and then concatenated these features to obtain fusion results, as described in Section 2.

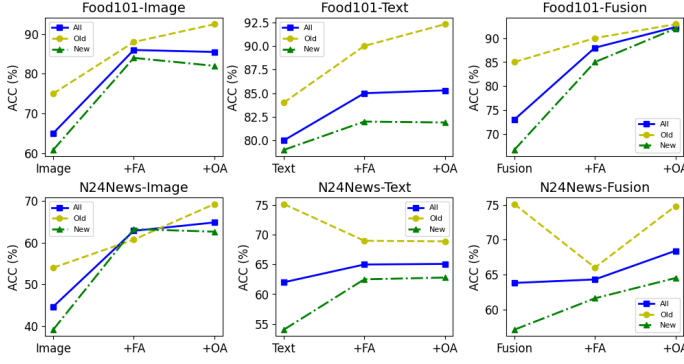


Figure 4. Ablation study for different form of alignment. FA: add feature space alignment using multimodal contrastive learning. OA: add output space alignment using cross-modal distillation

Since we have three different classifiers for image, text and fusion modality, we recorded the accuracy of the predicted labels from each classifier separately, denoted as **Visual**, **Text**, and **Fusion** in Table 3.

Additionally, we included fully supervised models like MMBT [11], CMA-CLIP [14], and UniS-MMC [33] as upper-bound references, which were trained under full supervision without distinguishing between new and old categories.

Our results, summarized in Table 3, demonstrate that our model significantly outperforms the extended GCD and SimGCD models across all configurations. Specifically, our model achieves a 11.5% improvement on Food101 and a 4.7% improvement on N24News compared to SimGCD, emphasizing the impact of our alignment strategies in the multimodal context. Key observations include:

- Our approach consistently improves performance by effectively aligning visual and textual features, unlike SimGCD where modality fusion sometimes reduces accuracy.
- The model shows strong capability in recognizing new categories, with cross-modal feature interaction enhancing robustness and reducing noise.
- On Food101, our fusion-based method outperforms fully supervised models like MMBT, even with only 50% of classes labeled, demonstrating the potential for improving annotation efficiency in multimodal tasks without losing accuracy.

6.3. Ablation Study

In this section, we provide a detailed step-by-step analysis of how our multimodal Generalized Category Discovery approach evolves from a straightforward baseline to the full model. Our approach utilizes three classifiers, each corresponding to **Image**, **Text**, and **Fusion** accuracy, respectively. Therefore, we report the accuracy changes for all three modalities at each step of the process. While we track

the accuracy for image and text separately, our primary focus is on the fusion accuracy, as it reflects the combined effectiveness of the multimodal integration.

We start with the baseline method, where the SimGCD [29] is applied separately to the image, text, and concatenated fusion features. This baseline setup provides the initial classification accuracy for each modality and serves as the foundation for subsequent enhancements.

The first enhancement introduces feature space alignment via cross-modal contrastive learning, as shown in the "FA" stage of Figure 4. This step significantly improves accuracy, especially for "New" categories, by better leveraging complementary information across modalities. Slight drops in "Old" category accuracy on N24News are likely due to previous ways overfitting to well-represented old classes. We mitigate this by using rich information across different modalities. Details under unbalanced conditions are provided in Section 6.5.

The final step involves output space alignment through an entropy-based loss, indicated by the "OA" stage. Even with minimal gains in individual modalities, the fusion results show substantial improvements. This confirms our theoretical insight that aligning predicted distributions across modalities creates a more coherent feature space, leading to the best performance.

Overall, aligning both feature and output spaces consistently boosts performance across categories, ensuring robust multimodal data handling for diverse instances.

6.4. Feature Visualization

To assess the impact of our alignment strategies, we use t-SNE [23] to visualize feature quality. We trained a model on the Food101 dataset and selected 10 random classes to maintain clarity. Figure 5 shows clustering results under three conditions: (a) baseline methods with three modality loss, but no alignment; (b) after adding feature space alignment via cross-modal contrastive learning; and (c) after incorporating output space alignment by matching predicted distributions using entropy minimization. Dots represent text features, and triangles represent image features. Key observations include:

- With only the unimodal loss, image and text feature are poorly aligned, leading to suboptimal classification.
- Feature space alignment reduces intra-class distances and increases inter-class separation, improving clustering. However, some misalignment persists.
- Output space alignment further aligns the feature distributions, with nearly overlapping centroids for images and texts, resulting in more distinct and separable clusters, crucial for effective multimodal GCD

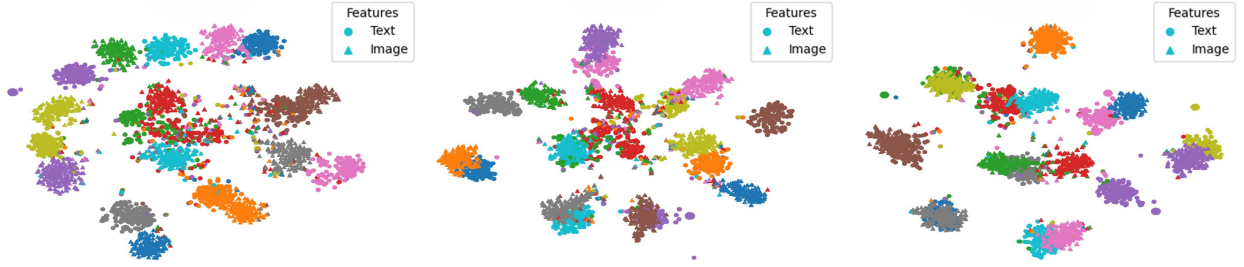


Figure 5. t-SNE result in Food101. From left to right: (a) Using the baseline methods, present unimodal GCD loss with image, text, and fused features; (b) Incorporating multimodal contrastive learning (feature align); (c) Adding multimodal prototype distillation (output align). Dots and triangles represent text and image respectively.

6.5. Analysis of Prediction Bias

An important issue with traditional GCD methods is their tendency to overfit the old classes due to the limited information available for some categories. Therefore, increasing the prediction balance between old and new classes is key to improving performance. We replicated the prediction results of SimGCD[29] on the Food101 dataset[27], following the testing protocol where the ratio of new to old classes is set at 1:2. As shown in Fig.6, a comparison between SimGCD’s predictions and the ground truth reveals a significant disparity. SimGCD tends to over-predict old classes and under-predict new ones, resulting in a pronounced bias towards old classes. The confusion matrix highlights this imbalance, which we believe is the primary factor behind SimGCD’s lower accuracy on multimodal datasets. The model’s strong bias towards old classes limits its ability to generalize to novel categories.

In contrast, our model demonstrates improvements in prediction balance, with predictions closely matching the true class distribution. The confusion matrix shows a more uniform distribution across both old and new classes, indicating reduced bias. We attribute this to our cross-modal supervision strategy, which enhances model robustness by integrating complementary information from different modalities. This effectively mitigates prediction imbalance, leading to more accurate and balanced classification, particularly in real-world scenarios with uneven class distributions.

6.6. Visualization of Attention Maps

We adopt attention weights to generate visualized heatmaps, following common practices [26, 31] of interpreting Vision Transformer models. The visualization results are used to reflect the most essential spatial regions regarding their predictions. Showcases from the Food101 dataset are presented in Figure 7. Our method better utilizes the information from the textual side to help the image side attention focus on the classification target. As shown in Figure 7, both SimGCD and the original GCD models only focus on the apples at the back of the image because they do

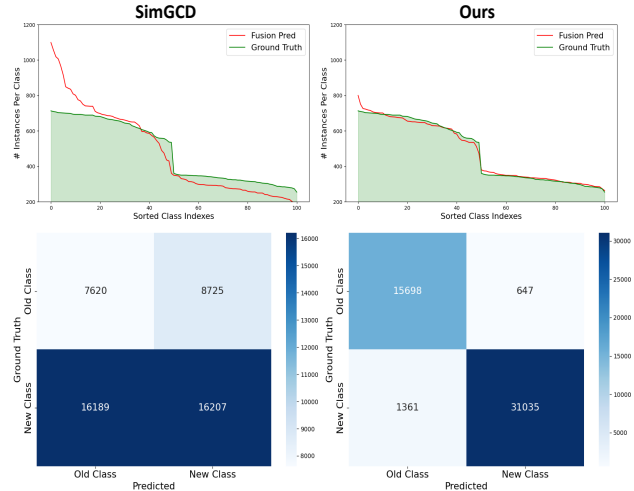


Figure 6. Comparison of prediction results distribution between SimGCD and our methods. The first row shows a comparison between the model’s classification results and the ground truth at the category level. For clarity, all categories are sorted based on the number of ground truth samples, where classes 1-50 are old classes and classes 51-101 are new classes. The second row presents the confusion matrix for the binary classification of new vs. old classes.

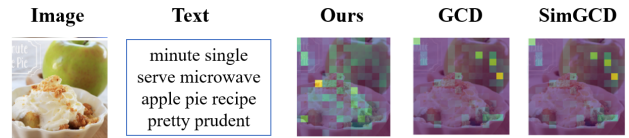


Figure 7. Sample of attention map visualization

not receive textual cues related to the pie. In contrast, our model effectively uses the information provided by the text, focusing more on the apple pie in the bowl. This demonstrates that when the images themselves are somewhat confusing or similar, textual information can effectively help the model focus on relevant content.

7. Conclusion

In conclusion, our research introduces a novel Multi-modal Generalized Category Discovery (MM-GCD) framework, marking a significant advancement in handling mixed labeled and unlabeled datasets. By leveraging multi-modal data, our approach not only addresses the limitations of traditional unimodal GCD but also demonstrates superior performance in identifying new categories across various benchmarks. This study sets a new standard for GCD applications, highlighting the potential of multimodal data in complex real-world classification tasks.

References

- [1] Wenbin An, Feng Tian, Qinghua Zheng, Wei Ding, Qian-Ying Wang, and Ping Chen. Generalized category discovery with decoupled prototypical network, 2023. 1, 2, 6
- [2] Antonios Anastasopoulos, Shankar Kumar, and Hank Liao. Neural language modeling with visual features, 2019. 2
- [3] Elisabet Ars and Roser Torra. Rare diseases, rare presentations: recognizing atypical inherited kidney disease phenotypes in the age of genomics. *Clinical Kidney Journal*, 10(5):586–593, 2017. 1
- [4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019. 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 (ICCV)*, pages 9650–9660, 2021. 6
- [6] Fernando Julio Cendra, Bingchen Zhao, and Kai Han. Promptccd: Learning gaussian mixture prompt pool for continual category discovery, 2024. 2
- [7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6
- [9] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. *arXiv preprint arXiv:2002.05714*, 2020. 2
- [10] Xuhui Jia, Kai Han, Yukun Zhu, and Bradley Green. Joint representation learning and novel category discovery on single- and multi-modal data, 2021. 2
- [11] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text, 2020. 6, 7
- [12] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 5, 6
- [13] Qiang Li, Qiuyang Ma, Weizhi Nie, and Anan Liu. Reinforcement learning based multi-modal feature fusion network for novel class discovery, 2023. 2
- [14] Huidong Liu, Shaoyuan Xu, Jinmiao Fu, Yang Liu, Ning Xie, Chien-Chih Wang, Bryan Wang, and Yi Sun. Cma-clip: Cross-modality attention clip for image-text classification. *arXiv preprint arXiv:2112.03562*, 2021. 6, 7
- [15] Shijie Ma, Fei Zhu, Zhun Zhong, Wenzhuo Liu, Xu-Yao Zhang, and Cheng-Lin Liu. Happy: A debiased learning framework for continual generalized category discovery, 2024. 2
- [16] Shijie Ma, Fei Zhu, Zhun Zhong, Xu-Yao Zhang, and Cheng-Lin Liu. Active generalized category discovery, 2024. 2
- [17] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Ng. Multimodal deep learning. pages 689–696, 2011. 2
- [18] Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. Deep multimodal fusion for persuasiveness prediction. 2016. 2
- [19] Rabah Ouldnooghi, Chia-Wen Kuo, and Zsolt Kira. Clip-gcd: Simple language guided generalized category discovery. *arXiv preprint arXiv:2305.10420*, 2023. 2
- [20] Keon-Hee Park, Hakyung Lee, Kyungwoo Song, and Gyeong-Moon Park. Online continuous generalized category discovery, 2024. 2
- [21] Nan Pu, Zhun Zhong, Xinyuan Ji, and Nicu Sebe. Federated generalized category discovery, 2023. 2
- [22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 6
- [23] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7
- [24] 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, 4, 6
- [25] Enguang Wang, Zhimao Peng, Zhengyuan Xie, Xialei Liu, and Ming-Ming Cheng. Get: Unlocking the multi-modal potential of clip for generalized category discovery. *arXiv preprint arXiv:2403.09974*, 2024. 2
- [26] Hongjun Wang, Sagar Vaze, and Kai Han. Sptnet: An efficient alternative framework for generalized category discovery with spatial prompt tuning, 2024. 1, 2, 6, 8
- [27] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2015. 2, 6, 8

- [28] Zhen Wang, Xu Shan, Xiangxie Zhang, and Jie Yang. N24News: A new dataset for multimodal news classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6768–6775, Marseille, France, 2022. European Language Resources Association. [2](#), [6](#)
- [29] 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. [2](#), [4](#), [5](#), [6](#), [7](#), [8](#)
- [30] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning, 2021. [2](#)
- [31] Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Shahbaz Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3479–3488, 2023. [2](#), [8](#)
- [32] Haiyang Zheng, Nan Pu, Wenjing Li, Nicu Sebe, and Zhun Zhong. Textual knowledge matters: Cross-modality co-teaching for generalized visual class discovery. *arXiv preprint arXiv:2403.07369*, 2024. [2](#)
- [33] Heqing Zou, Meng Shen, Chen Chen, Yuchen Hu, Deepu Rajan, and Eng Siong Chng. Unis-mmcc: Multimodal classification via unimodality-supervised multimodal contrastive learning, 2023. [2](#), [6](#), [7](#)