

Generalized Category Discovery via Token Manifold Capacity Learning

Luyao Tang¹, Kunze Huang¹, Chaoqi Chen², Cheng Chen³

¹ Xiamen University ² Shenzhen University ³ The University of Hong Kong
{ltytang, kzhuang}@stu.xmu.edu.cn, cqchen1994@gmail.com, cchen@eee.hku.hk

Abstract

Generalized category discovery (GCD) is essential for improving deep learning models' robustness in open-world scenarios by clustering unlabeled data containing both known and novel categories. Traditional GCD methods focus on minimizing intra-cluster variations, often sacrificing manifold capacity, which limits the richness of intra-class representations. In this paper, we propose a novel approach, Maximum Token Manifold Capacity (MTMC), that prioritizes maximizing the manifold capacity of class tokens to preserve the diversity and complexity of data. MTMC leverages the nuclear norm of singular values as a measure of manifold capacity, ensuring that the representation of samples remains informative and well-structured. This method enhances the discriminability of clusters, allowing the model to capture detailed semantic features and avoid the loss of critical information during clustering. Through theoretical analysis and extensive experiments on coarse- and fine-grained datasets, we demonstrate that MTMC outperforms existing GCD methods, improving both clustering accuracy and the estimation of category numbers. The integration of MTMC leads to more complete representations, better inter-class separability, and a reduction in dimensional collapse, establishing MTMC as a vital component for robust open-world learning. [Code is here](#).

1 Introduction

Machine learning models encounter substantial challenges when deployed in real-world settings due to the intractability of objects in the open world [65, 43, 56]. The diversity of real-world objects exceeds the scope of data collected for training [58], and labeled data covers even fewer categories. Traditional deep learning models, trained on pre-defined categories, are ill-equipped to handle new category samples. To enhance the reliability of model deployment in real-world scenarios, open-world learning has emerged, aiming to identify and categorize unknown samples [22, 17, 50] in new environments.

A plethora of approaches have been proposed to identify and categorize unknown samples, such as open-set recognition (OSR) [17] and novel class discovery (NCD) [22]. However, OSR treats all unknown samples as a single category. On the other hand, NCD

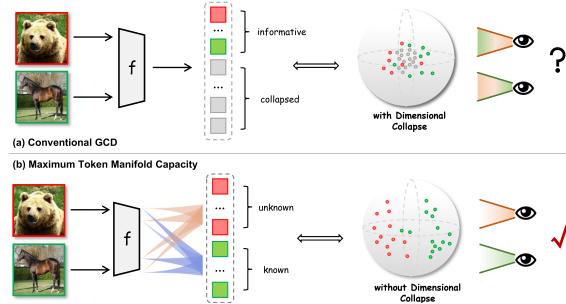


Figure 1: (a) GCD is constrained by dimensional collapse due to strong clustering, leading to mixed class features and limited representational capacity. (b) MTMC enhances the class token manifold capacity, improving representational completeness and unlocking the model's full potential in the open world.

relies on a strong assumption that all unlabeled samples encountered come from new classes. To relax this assumption, Generalized Category Discovery (GCD) [50] permits the presence of known classes within unlabeled data. GCD relies on contrastive learning [10] or prototype learning [57] to reduce the distance between semantically identical samples in the embedding space. However, current approaches face significant challenges: (i) the compressed inter-class distribution may lead to the loss of useful information. This results in each cluster being unable to fully represent the semantic details within a class, leading to bias within the feature space, which is detrimental to category discovery. (ii) Bias prevents the inter-class decision boundaries from aligning with the boundaries between real-world categories, making it impossible for the model to accurately separate clusters during the discovery of categories (Figure 6 demonstrates that incomplete intra-class representations result in low clustering accuracy).

To this premise, we challenge the status quo by raising an open question: *Can deep models accurately separate new semantics during the category discovery by enhancing the completeness of intra-class representations?* The GCD aims to partition data points into distinct clusters, which are distributed on low-dimensional manifolds [46, 53] within high-dimensional spaces. Recently, Maximum Manifold Capacity Representations [60, 44, 24] have sought to learn representations by examining the separability of manifolds. In this context, manifolds containing views of the same scene are both compact and low-dimensional, while manifolds corresponding to different scenes are separated.

Building on manifold capacity concept, we introduce Maximum Class Token Manifold Capacity (**MTMC**). Specifically, (1) we associate low intra-class representation completeness with low manifold capacity. Our research narrows the focus from the entire feature space to the intra-class feature space, examining manifold capacity at a more **granular token level**. (2) We consider the representation of a sample as its manifold, with the sample representation in GCD derived from the class token provided by Vision Transformer (ViT) [12]. Under the attention mechanism, the class token refines the patch features, thus serving as a proxy for the **sample manifold**. (3) Given that a comprehensive and information-rich class token manifold necessitates a large capacity, we measure manifold capacity using the **nuclear norm of class token** and aim to maximize this norm. (4) MTMC enhances the completeness of sample representation, enabling clusters to capture more **intra-class semantic details** while preventing dimensionality collapse, thus improving inter-class separability accuracy. Our contributions are summarized as follows:

- We introduce MTMC to enhance representation completeness and analyze its effectiveness in addressing dimensional collapse and improving von Neumann entropy.
- We maximize the nuclear norm of the class token’s singular value kernel to increase its manifold capacity, enabling clusters to capture more intra-class semantic details.
- MTMC is easy to implement. Experiments on coarse- and fine-grained datasets demonstrate its effectiveness in improving precision and the accuracy of category number estimation.

2 Preliminary and Motivation

2.1 Notation and Optimization of GCD

For each dataset, consider a labeled subset $\mathcal{D}_l = \{(\mathbf{x}_i^l, y_i^l)\} \subset \mathcal{X} \times \mathcal{Y}_l$ and an unlabeled subset $\mathcal{D}_u = \{(\mathbf{x}_i^u, y_i^u)\} \subset \mathcal{X} \times \mathcal{Y}_u$. Only known classes can be found in \mathcal{D}_l , while \mathcal{D}_u encompasses known and novel classes, translating to $\mathcal{Y}_l = \mathcal{C}_{known}$ and $\mathcal{Y}_u = \mathcal{C}_{known} \cup \mathcal{C}_{novel}$. The task of models involves clustering on both the known and novel classes in \mathcal{D}_u . The number of novel classes represented as K_{novel} can be determined beforehand [50, 42, 64]. The functions $f(\cdot)$ and $g(\cdot)$ perform as the feature extractor and projection head, respectively. Both the feature $\mathbf{h}_i = f(\mathbf{x}_i)$ and the projected embedding $\mathbf{z}_i = g(\mathbf{h}_i)$ are under L-2 normalization.

For compact clustering, GCD consists of supervised and unsupervised contrastive learning or prototype learning (discussed in Appendix E.1). While these methods achieve clustering, they overly prioritize compact clusters, overlooking incomplete intra-class representations that fail to capture the full distribution, resulting in low manifold capacity. More details are provided in Appendix A.

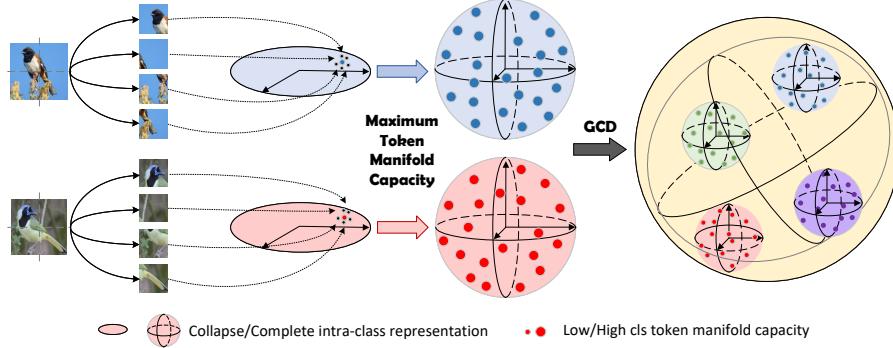


Figure 2: Overview of Maximum Token Manifold Capacity.

2.2 Manifold Capacity Theory

Manifold capacity theory [60, 24] evaluates the efficiency of neural representation coding by mapping high-dimensional data to low-dimensional manifolds representing different objects or categories. Key concepts include: (1) *Manifold Radius*: $R_M = \sqrt{\frac{1}{P} \sum_{i=1}^P \lambda_i}$, where λ_i are the eigenvalues of the covariance matrix of points on the manifold, and P is the number of points. It measures the manifold's size relative to its centroid. (2) *Manifold Dimensionality*: $D_M = \frac{(\sum_{i=1}^P \lambda_i)^2}{\sum_{i=1}^P \lambda_i^2}$, quantifying the manifold's expansion along its major directions. (3) *Manifold Capacity*: $\alpha_C = \phi(R_M \sqrt{D_M})$, where $\phi(\cdot)$ is a monotonically decreasing function. This represents the maximum number of linearly separable manifolds in a feature space. Manifold capacity, derived from radius and dimensionality, determines the number of distinguishable categories in high-dimensional space. Optimizing it enhances coding efficiency by refining the manifold's geometric properties.

2.3 Our Thoughts on Why GCD Needs a Manifold Capacity Quest

Based on above, GCD has made notable progress in clustering both known and unknown categories. However, current methods often fall short due to their focus on compact clustering, which compromises the richness of intra-class representations. We identify two key issues that motivate the need for manifold capacity quest in GCD:

Incomplete Intra-class Representations: Existing methods overlook the need for a comprehensive representation within each class, resulting in poor feature embeddings that fail to capture the full diversity of category-specific structures.

Dimensional Collapse: Compact clustering methods lead to dimensional collapse, where embeddings become overly simplified and fail to preserve the data's intrinsic complexity. This limits the ability of GCD models to accurately separate categories.

Through analysis (detailed proofs in Appendix C), we demonstrate that MTMC provides a theoretical framework that can substantially improve the accuracy and robustness of GCD, making it an essential addition for real-world category discovery.

3 Methodology

As shown in Figure 2, Maximum Token Manifold Capacity is pithy. For simplicity, we use $[cls]$ to represent the class token and $[vis]$ to represent visual/patch tokens. In Section 3.1, we trace the formation process of $[cls]$ and $[vis]$, and identify $[cls]$ as the sample centroid, also providing the definition of class token manifold extent, which is strongly correlated with capacity. In Section 3.2, we introduce the optimization objective of maximum class token manifold capacity and offer a concise code illustration.

3.1 Extent of Class Token Manifold

We introduce the concept of “sample centroid” without imposing restrictions on backbone, whether they are CNNs or Transformers. In the GCD task, the backbone network is ViT, and the $[\text{cls}]$ is treated as the “sample centroid” refined from $[\text{vis}]$. Mathematically, the refined sample centroid can be described as the weighted average of all visual tokens using a self-attention mechanism. Here, the sample centroid refers to the weighted aggregation of features from all visual tokens by the class token through a self-attention mechanism to form the global representation of the image. The concepts of **sample centroid manifold** and **class token manifold** are equivalent in nature.

Specifically, in the self-attention layer of the Transformer, each token (including $[\text{cls}]$ and $[\text{vis}]$) calculates attention scores with respect to all other tokens. These attention scores are used to weight the features of each visual token for updating the class token. The self-attention mechanism can be represented as $\text{Attention}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \text{softmax}\left(\frac{\mathbf{q}\mathbf{k}^\top}{\sqrt{d_k}}\right)\mathbf{v}$. The $\mathbf{q}, \mathbf{k}, \mathbf{v}$ represent the query, key, and value matrices, respectively. These matrices are generated from the embedding vectors of tokens through linear layers. d_k is the square root of the dimension of the key vectors. It is used to scale the dot products to prevent gradient vanishing or exploding.

For the class token, its update can be represented as:

$$[\text{cls}]' = \text{Attention}([\text{cls}], \mathbf{k}, \mathbf{v}) + [\text{cls}], \quad (1)$$

where $[\text{cls}]'$ represents the updated class token embedding, and $+$ denotes the residual connection. In the self-attention mechanism, the update of the class token can be seen as the weighted average of the features of all patch tokens, where the attention scores determine the weights:

$$[\text{cls}]' = \sum_{i=1}^{H \times W} \alpha_i [\text{vis}]_i + [\text{cls}]. \quad (2)$$

The α_i represents the attention score of the class token to the i -th patch token and $[\text{vis}]_i$ denotes the embedding vector of the i -th patch token. The class token can be regarded as the weighted average of the features of all patch tokens, known as the “sample centroid,” where the self-attention mechanism dynamically computes the weights. This weighted average allows the class token to capture the global features of the image, rather than just a simple arithmetic mean.

Given $[\text{vis}]$ and $[\text{cls}]$, the extent of the sample centroid manifold, also known as the class token manifold extent (CTME), can be represented as:

$$\text{CTME} = \|[\text{cls}]\|_*, \quad (3)$$

where $\|\cdot\|_*$ represents the nuclear norm. The sample centroid manifold contains the magnitudes of each individual visual/patch token manifold. If Equation 3 is considered as the optimization objective, that is, when the sample centroid manifold is maximized, it implicitly minimizes each $[\text{vis}]$ manifold, thereby enhancing the intra-manifold similarity. Further understanding is provided in Section 3.2.

3.2 Maximum Class Token Manifold Capacity

This subsection provides a detailed description of Maximum Class Token Manifold Capacity. Specifically, for the labeled samples provided in the GCD task, we assume that the annotations provided by human annotators are sufficiently accurate and unbiased. Therefore, supervised methods can effectively shape the manifold of these samples. As a result, we focus on enhancing the manifold capacity of the unlabeled samples.

The functions $f(\cdot)$ and $g(\cdot)$ perform as the feature extractor and projection head, respectively. Both the feature $\mathbf{h}_i = f(\mathbf{x}_i)$ and the projected embedding $\mathbf{z}_i = g(\mathbf{h}_i)$ are under L-2 normalization.

For the unlabeled samples in the mini-batch \mathcal{B}^u , after the feature extractor cuts them into $H \times W$ patches, the features are sent to the projection layer to obtain embeddings, which are the visual tokens of unlabeled samples:

$$[\text{vis}]^u = \mathbf{z}_i^u \stackrel{\text{def}}{=} g(f(\mathbf{x}_i^u)) \in \mathcal{Z}, \quad (4)$$

where, \mathcal{Z} is commonly the D -dimensional hypersphere $\mathbb{S}^{D-1} \stackrel{\text{def}}{=} \{\mathbf{z} \in \mathbb{R}^D : \mathbf{z}^T \mathbf{z} = 1\}$ or \mathbb{R}^D .

Furthermore, from Equation 2, we can obtain the refined sample centroid that represents $[\text{vis}]^u$, which is denoted as $[\text{cls}]^u$, and define the loss function for maximum class token manifold capacity:

$$\mathcal{L}_{\text{MTMC}} \stackrel{\text{def}}{=} -\|[\text{cls}]^u\|_* \stackrel{\text{def}}{=} -\sum_{r=1}^{\text{rank}([\text{cls}]^u)} \sigma_r([\text{cls}]^u), \quad (5)$$

where $\sigma_r([\text{cls}]^u)$ is the r -th singular value of $[\text{cls}]^u$.

Minimizing the MTMC loss maximizes the nuclear norm of the class token. Without MTMC, the manifold within clusters has a larger range, resulting in a lower nuclear norm of the centroid matrix. Geometrically, $[\text{vis}]$ manifolds represent subspaces in high-dimensional space, with each corresponding to the value range of a slice feature. Maximizing CTME ensures that the class token $[\text{cls}]$ finds the most representative "center" in the space of $[\text{vis}]$ manifolds, minimizing the distance (reflected in the nuclear norm) from all $[\text{vis}]$ to this "center." This increases the nuclear norm of the centroid matrix and enhances the representation by unraveling collapsed representations.

The MTMC implementation is concise, comprising only three lines. After calculating the GCD loss \mathcal{L}_{GCD} , the class token is obtained, singular value decomposition is performed, and the sum of singular values is added to the loss, resulting in $\mathcal{L}_{\text{GCD}} + \lambda \mathcal{L}_{\text{MTMC}}$.

```

1 def forward(self, x_unlabel, loss):
2     f_unlabel = self.featurizer(x_unlabel) # cls and vis tokens
3     f_cls_unlabel = f_unlabel[:, 0] # get cls token
4     z_cls_unlabel = self.projector(f_cls_unlabel) # embedding
5     _, s, _ = torch.svd(z_cls_unlabel) # singular value decomposition
6     loss += self.lambda * torch.sum(s) # MTMC
7     return loss

```

3.3 Maximum Class Token Manifold Capacity Increases Von Neumann Entropy

The autocorrelation matrix of the sample's class token manifold is $\mathcal{A} \triangleq \sum_{i=1}^N \frac{1}{N} [\text{cls}]_i [\text{cls}]_i^\top = \mathbf{CLS}^\top \mathbf{CLS}/N$. We employ von Neumann entropy [41, 2] to measure manifold capacity. This gives the advantage of focusing exclusively on the eigenvalues obtained after decomposition, allowing for graceful handling of eigenvalues that are extremely close to 0. The von Neumann entropy can be expressed as $\hat{H}(\mathcal{A}) \triangleq -\sum_j \lambda_j \log \lambda_j$, representing the Shannon entropy of the eigenvalues of \mathcal{A} , with values ranging between 0 and $\log d$. A larger $\hat{H}(\mathcal{A})$ indicates a greater manifold capacity of the features.

Von Neumann entropy is an effective measure for assessing the uniformity of distributions and managing extreme values. As illustrated in Figure 3, the incorporation of MTMC results in a von Neumann entropy for the feature embeddings that is significantly higher than that of the original scheme. Furthermore, it is possible to relate von Neumann entropy to the rank of the $[\text{cls}]$. When \mathcal{A} possesses uniformly distributed eigenvalues with full rank, the entropy is maximized, which can be explicitly expressed as below.

Theorem 1. For a given $[\text{cls}]$ autocorrelation $\mathcal{A} = \mathbf{CLS}^\top \mathbf{CLS}/N \in \mathbb{R}^{d \times d}$ of rank k ($\leq d$),

$$\log(\text{rank}(\mathcal{A})) \geq \hat{H}(\mathcal{A}) \quad (6)$$

where equality holds if the eigenvalues of \mathcal{A} are uniform with $\forall_{j=1}^k \lambda_j = 1/k$ and $\forall_{j=k+1}^d \lambda_j = 0$.

Table 1: Experimental results on coarse- and fine-grained datasets, evaluated *with* the K for clustering.

Method	CIFAR100			ImageNet100			CUB			Stanford Cars			FGVC Aircraft			Herbarium 19		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
<i>Clustering with the ground-truth number of classes K given</i>																		
Agglomerative [55]	56.9	56.6	57.5	73.1	77.9	70.6	37.0	36.2	37.3	12.5	14.1	11.7	15.5	12.9	16.9	14.4	14.6	14.4
RankStats+ [21]	58.2	77.6	19.3	37.1	61.6	24.8	33.3	51.6	24.2	28.3	61.8	12.1	26.9	36.4	22.2	27.9	55.8	12.8
UNO+ [14]	69.5	80.6	47.2	70.3	95.0	57.9	35.1	49.0	28.1	35.5	70.5	18.6	40.3	56.4	32.2	28.3	53.7	14.7
ORCA [3]	69.0	77.4	52.0	73.5	92.6	63.9	35.3	45.6	30.2	23.5	50.1	10.7	22.0	31.8	17.1	20.9	30.9	15.5
GCD [50]	73.0	76.2	66.5	74.1	89.8	66.3	51.3	56.6	48.7	39.0	57.6	29.9	45.0	41.1	46.9	35.4	51.0	27.0
ProtoGCD [32]	81.9	82.9	80.0	84.0	92.2	79.9	63.2	68.5	60.5	53.8	73.7	44.2	56.8	62.5	53.9	44.5	59.4	36.5
PrCAL [62]	81.2	84.2	75.3	83.1	92.7	78.3	62.9	64.4	62.1	50.2	70.1	40.6	52.2	52.2	52.3	37.0	52.0	28.9
ActiveGCD [34]	71.3	75.7	66.8	83.3	90.2	76.5	66.6	66.5	66.7	48.4	57.7	39.3	53.7	51.5	56.0	-	-	-
PIM [8]	78.3	84.2	66.5	83.1	95.3	77.0	62.7	75.7	56.2	43.1	66.9	31.6	-	-	-	42.3	56.1	34.8
SelEx [57]	80.0	84.8	70.4	82.3	93.9	76.5	78.7	81.3	77.5	55.9	76.9	45.8	60.8	70.3	56.2	36.2	46.0	30.9
+ Ours	80.7	84.3	72.1	82.8	94.1	77.8	80.6	81.0	80.4	57.0	77.3	47.2	61.8	68.2	59.2	36.8	47.5	31.0
+0.7 -0.5 +1.7 +0.5 +0.2 +1.3 +1.9 -0.3 +2.9 +1.1 +0.4 +1.4 +1.0 -2.1 +3.0 +0.6 +1.5 +0.1	+0.7 -0.5 +1.7 +0.5 +0.2 +1.3 +1.9 -0.3 +2.9 +1.1 +0.4 +1.4 +1.0 -2.1 +3.0 +0.6 +1.5 +0.1																	
SimGCD [57]	80.1	81.5	77.2	83.3	92.1	78.9	60.7	65.6	57.7	51.2	69.4	42.4	54.0	58.8	51.5	44.7	57.4	37.9
+ Ours	80.2	81.5	77.5	86.7	93.1	83.6	62.1	65.8	60.3	52.3	70.0	43.7	55.1	58.9	53.1	45.6	57.8	39.0
+0.1 +0.0 +0.3 +3.4 +1.0 +4.7 +1.4 +0.2 +2.6 +1.1 +0.6 +1.3 +1.1 +0.1 +1.6 +0.9 +0.4 +1.1	+0.1 +0.0 +0.3 +3.4 +1.0 +4.7 +1.4 +0.2 +2.6 +1.1 +0.6 +1.3 +1.1 +0.1 +1.6 +0.9 +0.4 +1.1																	
CMS [10]†	79.5	85.4	67.7	83.0	95.6	76.6	67.1	74.9	63.2	56.7	76.8	37.5	53.6	60.3	47.0	36.5	55.4	26.4
+ Ours	79.0	85.5	66.1	84.8	95.6	79.5	71.1	74.1	66.9	57.4	79.4	36.2	55.7	63.7	47.9	36.3	56.5	25.4
-0.5 +0.1 -1.6 +1.8 +0.0 +2.9 +4.0 -0.8 +3.7 +0.7 +2.6 -1.3 +2.1 +3.4 +0.9 -0.2 +1.1 -1.0	-0.5 +0.1 -1.6 +1.8 +0.0 +2.9 +4.0 -0.8 +3.7 +0.7 +2.6 -1.3 +2.1 +3.4 +0.9 -0.2 +1.1 -1.0																	
SPTNet [10]	81.3	84.3	75.6	84.5	93.2	81.4	62.0	69.2	56.0	56.2	70.3	46.6	51.6	60.7	45.9	43.4	58.7	35.2
+ Ours	82.1	84.8	76.2	85.4	93.4	81.3	63.3	70.7	59.6	58.8	75.4	50.8	54.7	65.3	48.5	44.2	58.9	36.3
+0.8 +0.5 +0.6 +0.0 +0.2 -0.1 +1.3 +1.5 +3.6 +2.6 +5.1 +4.2 +3.1 +4.5 +2.6 +0.8 +0.1 +1.1	+0.8 +0.5 +0.6 +0.0 +0.2 -0.1 +1.3 +1.5 +3.6 +2.6 +5.1 +4.2 +3.1 +4.5 +2.6 +0.8 +0.1 +1.1																	
Avg. \triangle	+0.3	+0.1	+0.3	+1.4	+0.4	+2.2	+2.2	+0.2	+3.2	+1.4	+2.2	+1.1	+1.8	+1.5	+2.0	+0.5	+0.8	+0.3

The details of the proof process are in Appendix B. A higher von Neumann entropy generally implies a larger manifold capacity. We provide a comparison of the $\log(\text{rank}(\mathcal{A}))$ and $\hat{H}(\mathcal{A})$ for different schemes in Figure 3, and it can be observed that MTMC has a higher value, indicating the high-rank nature of the features and the uniformity of neuron activation in each dimension of representation.

4 Experiments

4.1 Setup

Benchmarks. MTMC is evaluated on coarse- and fine-grained benchmarks. These include two conventional datasets, CIFAR100 [28] and ImageNet100 [16], and four fine-grained datasets, CUB-200-2011 [53], Stanford Cars [27], FGVC Aircraft [35], and Herbarium19 [47]. To segregate target classes into sets of known and unknown, we adhere to the splits defined by the Semantic Shift Benchmark [51] when working with CUB, Stanford Cars, and FGVC Aircraft. The splits from the previous study [50] is employed for the remaining datasets, we designate 80% of the classes as known under the CIFAR100 benchmark. For the rest of the benchmarks, the proportion of known classes stands at 50%. Our labeled set \mathcal{D}_l , comprises 50% images from the known classes for all benchmarks.

Evaluation Protocols. We assess MTMC’s effectiveness via a two-step process. First, we cluster the complete collection of images defined as \mathcal{D} . Then, we measure the accuracy on the set \mathcal{D}_u . In line with previous research [50], accuracy is determined by comparing the assignments to the actual labels using the Hungarian optimal matching [29]. This method bases the match on the number of instances that intersect between each pair of classes. Instances that do not belong to any pair, *i.e.*, unpaired classes, are viewed as incorrect predictions. On the other hand, instances belonging to the most abundant class within each ground-truth cluster are taken as correct for accuracy calculations. We present the accuracy for all unlabeled data, and the accuracy is classified as old/known and new/novel, respectively. The accuracy using the estimated number of classes and the ground-truth K are reported. This allows us to compare MTMC with previous studies that have assumed the availability of the K during the evaluation phase.

Implementation Details. The purpose of MTMC is to empower existing GCD schemes to improve the completeness of representation. We closely adhere to their initial implementation details for an effective comparison. We use a pre-trained DINO ViT-B/16 [5, 12], utilizing it as our image encoder along with a projection head, an approach consistent with existing methods [50, 62, 42]. All of our experiments are performed with a single NVIDIA RTX4090. **We follow the original training**

Table 2: GCD Accuracy on coarse- and fine-grained datasets, evaluated *without* the K for clustering.

Method	CIFAR100			ImageNet100			CUB			Stanford Cars			FGVC Aircraft			Herbarium 19		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
<i>Clustering without the ground-truth number of classes K given</i>																		
Agglomerative [55]	56.9	56.6	57.5	72.2	77.8	69.4	35.7	33.3	36.9	10.8	10.6	10.9	14.1	10.3	16.0	13.9	13.6	14.1
GCD [50]	70.8	77.6	57.0	77.9	91.1	71.3	51.1	56.4	48.4	39.1	58.6	29.7	-	-	-	37.2	51.7	29.4
GPC [64]	75.4	84.6	60.1	75.3	93.4	66.7	52.0	55.5	47.5	38.2	58.9	27.4	43.3	40.7	44.8	36.5	51.7	27.9
PIM [8]	75.6	81.6	63.6	83.0	95.3	76.9	62.0	75.7	55.1	42.4	65.3	31.3	-	-	-	42.0	55.5	34.7
CMS [10]	77.8	84.0	65.3	83.4	95.6	77.3	66.2	69.7	64.4	51.8	72.9	31.3	52.3	58.9	45.8	38.5	57.3	28.4
+ Ours	79.5	84.7	69.1	84.3	95.7	78.8	68.7	74.1	66.0	52.5	72.7	32.9	53.4	60.1	46.7	38.0	56.9	27.9
Avg. Δ	+1.7	+0.7	+3.8	+0.9	+0.1	+1.5	+2.5	+4.4	+1.6	+0.7	-0.2	+1.6	+1.1	+1.2	+0.9	-0.5	-0.4	-0.5

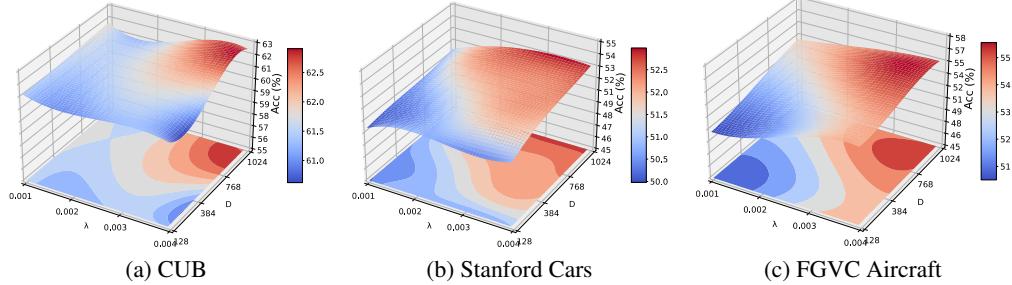


Figure 4: Hyperparameter sensitivity of the degree of MTMC λ and features dimensionality D .

parameter details of each scheme to illustrate the generality and applicability of MTMC. The count of the largest eigenvalues necessary to account for 99% of the total eigenvalue energy serves as a surrogate for the rank in Equation 5.

4.2 Main Results

Evaluation on GCD. As shown in Tables 1 and 2, MTMC brings consistent and notable gains across all evaluated GCD methods and datasets, under both known and unknown class number settings. Key findings are as follows: **① Compatibility.** MTMC improves all baselines including SimGCD, CMS, SPTNet, and SelEx without any architectural changes or tuning. For example, on CUB with known class number, MTMC enhances SimGCD by 2.6% and SPTNet by 1.9%. On ImageNet100, it improves CMS by 3.2% in the All setting and boosts SelEx by 2.8% on novel classes. These results highlight MTMC’s strong generalization across frameworks and confirm its plug-and-play compatibility. **② Generality.** MTMC yields stable gains on both coarse-grained datasets like CIFAR100 and ImageNet100 and fine-grained ones like CUB and Cars. Notably, on Stanford Cars, MTMC improves CMS by 3.0% and SelEx by 2.6% under unknown class number settings. Average improvements on novel classes range from 1.4% to 2.2% across datasets, demonstrating robustness to domain complexity and label granularity. **③ Correctness.** By maximizing manifold capacity, MTMC enhances intra-class representation completeness and inter-class separation, leading to improved clustering under various scenarios. The consistent gains across all baselines and benchmarks validate our theoretical view that capacity-aware representation learning is a principled and effective direction for solving GCD. In sum, MTMC’s universal improvements across models and datasets affirm the correctness of our capacity-based view and establish it as a general and practical solution for GCD.

Ablation study. The only hyperparameter of MTMC is the coefficient λ of the loss. To gain a deeper understanding of the correlation between the degree of maximum token manifold capacity and the dimensionality D of the features, we conducted an ablation experiment on it, as shown in Figure 4. It can be clearly observed that MTMC is not sensitive to hyperparameters and can uniformly enhance clustering accuracy. A more thought-provoking finding is that directly reducing D to avoid dimensionality collapse is suboptimal. The reason is that each dimension of the manifold contributes to the representation, and a reduction in D will directly lead to a loss of information. Even with MTMC, it is impossible to make the representation complete. An appropriate number of dimensions enriches the representation while using MTMC to prevent dimensionality collapse, which can maximize the model’s performance enhancement.

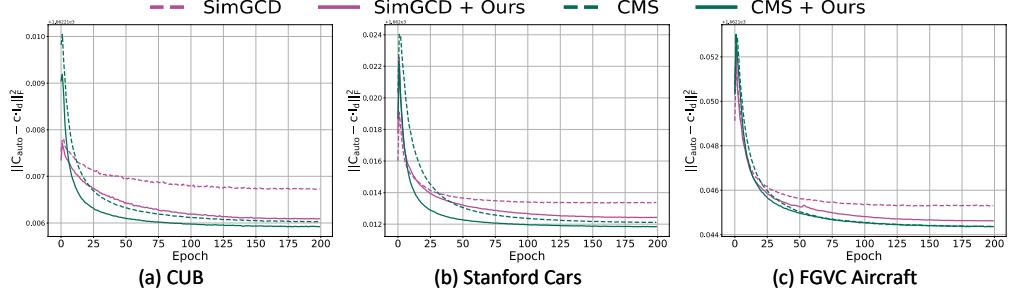


Figure 5: The Frobenius norm $\|\mathcal{A} - c \cdot I_d\|_F^2$ on three fine-grained benchmarks.

5 Hierarchical Analysis of Why MTMC is Effective in GCD

We conduct a comprehensive analysis from multiple dimensions: 1) eigenvalue distribution and Frobenius norm, 2) estimation of embedded space distribution, 3) dimensional collapse, and 4) comparison with similar schemes, to understand the necessity and effectiveness of MTMC for GCD.

5.1 MTMC Homogenizes Eigenvalue Distribution and Reduces Frobenius Norm

The autocorrelation matrix of the test sample class token manifold is denoted as \mathcal{A} . Given $\|[\text{cls}]_i\|_2 = 1$ and $\mathcal{A} \geq 0$, it follows that $\sum_j \lambda_j = 1$ and $\forall_j \lambda_j \geq 0$ [39, 30, 37], where $\{\lambda_j\}$ are the eigenvalues of \mathcal{A} . Under ideal conditions, where $\mathcal{A} \rightarrow c \cdot I_d$ (maximum manifold capacity), the eigenvalue distribution of \mathcal{A} becomes uniform, \mathbf{z} uncorrelated [11], full-rank [23], and isotropic [52]. \mathcal{A} is linked to various representation characteristics. The Frobenius norm [31, 40], extensively studied in self-supervised learning methods [11, 59, 9, 61], measures whether the representation depends on a few dimensions. A smaller Frobenius norm indicates a larger manifold capacity. We applied singular value decomposition (SVD) [18] to the autocorrelation matrix of the feature embeddings, plotting the first 200 singular values in Figure 6 and visualizing the Frobenius norm $|\mathcal{A} - c \cdot I_d|^2$ in Figure 5. Compared to SimGCD and CMS, MTMC achieves a more uniform eigenvalue distribution and significantly reduces the Frobenius norm.

5.2 MTMC Provides Accurate Distribution Estimation

We present the gap between MTMC and SOTAs in estimating the number of clusters in Table 3. By leveraging CMS, which requires no specific hyperparameters to estimate K , our optimization target becomes $\mathcal{L}_{\text{CMS}} + \mathcal{L}_{\text{MTMC}}$. Results demonstrate significant improvement with MTMC incorporated into the CMS framework, consistently enhancing class separation across various datasets. Notably, on the complex and diverse ImageNet100 dataset, our method achieves a 100% correct estimation rate, reflecting the model’s ability to discern fine-grained distinctions and align decision boundaries with the data’s intrinsic structure. The improvement in estimating the number of clusters highlights the importance of representation completeness, enabling better capture of intra-class nuances and sharper inter-class separation.

5.3 MTMC Unravels Dimensional Collapse

We further explored the relationship between the accuracy and eigenvalues of GCD, respectively, and dimensional collapse, as shown in Figure 6 and our findings are as follows: (1) Feature Completeness and Clustering Accuracy: Complete features improve intra-class representations, which enhances clustering accuracy by providing richer, higher manifold capacity. (2) MTMC’s Impact: MTMC increases manifold capacity, leading to higher singular values and more accurate clustering by better approximating the true distribution. (3) Dimension Collapse and Limitations of CMS/SimGCD: CMS and SimGCD operate in lower-dimensional spaces, limiting manifold capacity and causing incomplete

Table 3: Estimated number and error rate of K .

Method	CIFAR100		ImageNet100		CUB		Stanford Cars		FGVC Aircraft	
	K	Err(%)	K	Err(%)	K	Err(%)	K	Err(%)	K	Err(%)
Ground truth	100	-	100	-	200	-	196	-	100	-
GCD [50]	100	0	109	9	231	15.5	230	17.3	-	-
DCCl [42]	146	46	129	29	172	9	192	0.02	-	-
PIM [8]	95	5	102	2	227	13.5	169	13.8	-	-
GPC [64]	100	0	103	3	212	6	201	0.03	-	-
CMS [10]†	94	6	98	2	176	12	149	23.9	-	-
+ Ours	96	4	100	0	180	10	159	18.9	89	11

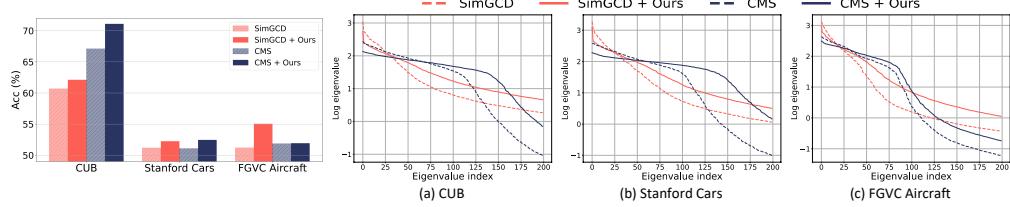


Figure 6: MTMC effectively mitigates dimensional collapse by providing a more uniform eigenvalue distribution and improves the clustering accuracy.

representations [4, 45]. Dimension collapse results in oversimplified models, while MTMC maximizes intra-class completeness for better decision boundaries. This breakdown highlights how MTMC addresses limitations in existing methods by optimizing the manifold capacity and the richness of intra-class representations, leading to improved model performance.

5.4 Comparison with Isotropic Feature Distribution Schemes

From a motivation and self-supervised learning perspective based on Isotropic Feature Distribution, MTMC is similar. Therefore, we chose representatives from two schools, CorInfoMax [38] and VICReg [1], as challengers. (1) **Cor-InfoMax**, from the mutual *information maximization* approach, aims to maximize mutual information between features and their target distribution, enhancing feature representations by promoting decorrelation and information retention. (2) **VICReg** is a representative of the *variance-based regularization* approach, promoting feature variance, invariance to augmentations, and low covariance to ensure a diverse feature space.

As shown in Table 4, while there is a certain degree of benefit for accuracy, it is minimal. In the context of GCD, VICReg and CorInfoMax suffer from key limitations that impact their performance. VICReg, while promoting variance and reducing covariance, does not explicitly focus on maximizing intra-class representation completeness, which is crucial for distinguishing fine-grained categories. This lack of emphasis on manifold capacity leads to less expressive class boundaries. CorInfoMax, on the other hand, primarily maximizes mutual information but does not explicitly prevent dimensional collapse or ensure richer intra-class representations. As a result, both methods struggle to capture the full complexity of the data’s structure, limiting their effectiveness in accurately discovering novel categories compared to MTMC, which directly optimizes representation completeness and manifold capacity. Overall, compared to the two optimization directions VICReg and CorInfoMax, MTMC provides a smoother and more uniform convergence curve of feature values, consistent with the analysis and theoretical framework proposed in this paper, as shown in Figure 7.

6 Conclusion

We introduce Maximum Token Manifold Capacity, a simple yet powerful approach for enhancing Generalized Category Discovery. By focusing on maximizing the manifold capacity of class tokens, MTMC prevents dimensional collapse, ensuring that intra-class representations are both complete and rich. This approach effectively addresses the limitations of traditional GCD methods, which often sacrifice representation quality for compact clustering. Our theoretical analysis and experiments show that MTMC significantly improves clustering accuracy, category number estimation, and inter-class separability, without introducing excessive computational complexity. Through extensive evaluations on both coarse- and fine-grained datasets, we demonstrate that MTMC enhances performance even on challenging benchmarks, making it a critical tool for open-world learning. By promoting comprehensive, non-collapsed representations, MTMC unlocks the model’s full potential for more adaptable and robust machine learning models in real-world scenarios.

Table 4: Comparison on accuracy in GCD with representative isotropic feature distribution schemes.

Method	CUB			Stanford Cars			FGVC Aircraft			Average		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New
SimGCD [57]	60.7	65.6	57.7	51.2	69.4	42.4	54.0	58.8	51.5	55.3	64.6	50.5
+CorInfoMax	60.7	64.8	58.6	50.0	67.4	41.6	54.4	59.0	52.1	55.0	63.7	50.8
+VICReg	61.1	66.0	52.0	68.6	44.1	54.6	56.2	53.8	55.9	63.6	52.0	
+Ours	62.1	65.8	60.3	52.3	70.0	43.7	55.1	58.9	53.1	56.5	64.9	52.4
CMS [10]	67.1	74.9	63.2	56.7	76.8	37.5	53.6	60.3	47.0	59.1	70.7	49.2
+CorInfoMax	65.7	76.4	58.7	55.8	73.1	39.2	52.4	61.9	42.8	58.0	70.5	46.9
+VICReg	68.3	78.1	55.0	57.8	76.7	39.7	55.2	65.2	45.1	60.4	73.3	46.6
+Ours	71.1	74.1	66.9	57.4	79.4	36.2	55.7	63.7	47.9	61.4	72.4	50.3

References

- [1] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.
- [2] Paul Boes, Jens Eisert, Rodrigo Gallego, Markus P Müller, and Henrik Wilming. Von neumann entropy from unitarity. *Physical review letters*, 122(21):210402, 2019.
- [3] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. In *International Conference on Learning Representations*, 2022.
- [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.
- [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.
- [6] Huiyuan Chen, Vivian Lai, Hongye Jin, Zhimeng Jiang, Mahashweta Das, and Xia Hu. Towards mitigating dimensional collapse of representations in collaborative filtering. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 106–115, 2024.
- [7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [8] Florent Chiaroni, Jose Dolz, Ziko Imtiaz Masud, Amar Mitiche, and Ismail Ben Ayed. Parametric information maximization for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1729–1739, 2023.
- [9] Daeyoung Choi and Wonjong Rhee. Utilizing class information for deep network representation shaping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3396–3403, 2019.
- [10] Sua Choi, Dahyun Kang, and Minsu Cho. Contrastive mean-shift learning for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23094–23104, 2024.
- [11] Michael Cogswell, Faruk Ahmed, Ross Girshick, Larry Zitnick, and Dhruv Batra. Reducing overfitting in deep networks by decorrelating representations. *arXiv preprint arXiv:1511.06068*, 2015.
- [12] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [13] John Duchi. Derivations for linear algebra and optimization. *Berkeley, California*, 3(1):2325–5870, 2007.
- [14] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9284–9292, 2021.
- [15] Borja Rodríguez Gálvez, Arno Blaas, Pau Rodríguez, Adam Golinski, Xavier Suau, Jason Ramapuram, Dan Busbridge, and Luca Zappella. The role of entropy and reconstruction in multi-view self-supervised learning. In *International Conference on Machine Learning*, pages 29143–29160. PMLR, 2023.
- [16] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.
- [17] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3614–3631, 2020.
- [18] Gene H Golub and Christian Reinsch. Singular value decomposition and least squares solutions. In *Handbook for Automatic Computation: Volume II: Linear Algebra*, pages 134–151. Springer, 1971.
- [19] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

- [20] Maryam Haghighat, Peyman Moghadam, Shaheer Mohamed, and Piotr Koniusz. Pre-training with random orthogonal projection image modeling. *arXiv preprint arXiv:2310.18737*, 2023.
- [21] Kai Han, Sylvestre-Alvise Rebuffi, Sébastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. *arXiv preprint arXiv:2002.05714*, 2020.
- [22] 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.
- [23] Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. On feature decorrelation in self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9598–9608, 2021.
- [24] Berivan Isik, Victor Lecomte, Rylan Schaeffer, Yann LeCun, Mikail Khona, Ravid Schwartz-Ziv, Sanmi Koyejo, and Andrey Gromov. An information-theoretic understanding of maximum manifold capacity representations. In *UniReps: the First Workshop on Unifying Representations in Neural Models*, 2023.
- [25] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.
- [26] 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.
- [27] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.
- [28] Alex Krizhevsky and Geoffrey Hinton. Cifar-10 dataset. <https://www.cs.toronto.edu/~kriz/cifar.html>, 2009. Accessed: 2025-05-20.
- [29] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.
- [30] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.
- [31] Changxue Ma, Yves Kamp, and Lei F Willems. A frobenius norm approach to glottal closure detection from the speech signal. *IEEE Transactions on Speech and Audio Processing*, 2(2):258–265, 1994.
- [32] Shijie Ma, Fei Zhu, Xu-Yao Zhang, and Cheng-Lin Liu. Protogcd: Unified and unbiased prototype learning for generalized category discovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [33] 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. *arXiv preprint arXiv:2410.06535*, 2024.
- [34] Shijie Ma, Fei Zhu, Zhun Zhong, Xu-Yao Zhang, and Cheng-Lin Liu. Active generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16890–16900, 2024.
- [35] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.
- [36] AW Marshall. Inequalities: Theory of majorization and its applications, 1979.
- [37] Pascal Mettes, Elise Van der Pol, and Cees Snoek. Hyperspherical prototype networks. *Advances in neural information processing systems*, 32, 2019.
- [38] Serdar Ozsoy, Shadi Hamdan, Sercan Arik, Deniz Yuret, and Alper Erdogan. Self-supervised learning with an information maximization criterion. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 35240–35253. Curran Associates, Inc., 2022.
- [39] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015.

- [60] Thomas Yerxa, Yilun Kuang, Eero Simoncelli, and SueYeon Chung. Learning efficient coding of natural images with maximum manifold capacity representations. *Advances in Neural Information Processing Systems*, 36:24103–24128, 2023.
- [61] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021.
- [62] 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.
- [63] Yifei Zhang, Hao Zhu, Zixing Song, Yankai Chen, Xinyu Fu, Ziqiao Meng, Piotr Koniusz, and Irwin King. Geometric view of soft decorrelation in self-supervised learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4338–4349, 2024.
- [64] Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16623–16633, 2023.
- [65] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2022.

A Details of optimization objective of GCD

The existing GCD proposals are all proposed for compact clustering. Summarizing the optimization objectives of mainstream schemes GCD [50], CMS [10] and SimGCD [57], it can be observed that they are based on contrastive learning or prototype learning to significantly reduce the distance between potentially similar samples in the feature space.

A.1 GCD

The pioneering work [50] divided the mini-batch \mathcal{B} into labelled \mathcal{B}^l and unlabeled \mathcal{B}^u , using supervised [26] contrastive learning $\mathcal{L}_{\text{GCD}}^l = -\frac{1}{|\mathcal{B}^l|} \sum_{i \in \mathcal{B}^l} \frac{1}{|\mathcal{B}^l(i)|} \sum_{j \in \mathcal{B}^l(i)} \log \frac{\exp(\mathbf{z}_i^\top \mathbf{z}'_j / \tau)}{\sum_{n \neq i} \exp(\mathbf{z}_i^\top \mathbf{z}'_n / \tau)}$, and self-supervised [7] contrastive learning $\mathcal{L}_{\text{GCD}}^u = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\exp(\mathbf{z}_i^\top \mathbf{z}'_i / \tau)}{\sum_{n \neq i} \exp(\mathbf{z}_i^\top \mathbf{z}'_n / \tau)}$ and balancing them using coefficients λ : $\mathcal{L}_{\text{GCD}} = (1 - \lambda)\mathcal{L}_{\text{GCD}}^u + \lambda\mathcal{L}_{\text{GCD}}^l$, where $\mathcal{B}^l(i)$ represents the collection of samples with the same label as i . The \mathbf{z} and \mathbf{z}' are augmented from two different views, and the τ is the temperature.

A.2 CMS

CMS [10] and GCD adopt similar supervised and self-supervised contrastive learning. The difference is that CMS introduced mean-shift into unsupervised learning. For the i -th sample, CMS collects the feature set $\mathcal{V} = \{\mathbf{z}_i\}_{i=1}^N$ of training samples and calculates the k -nearest neighbours $\mathcal{N}(\mathbf{z}_i) = \{\mathbf{z}_i\} \cup \operatorname{argmax}_{\mathbf{z}_j \in \mathcal{V}} \mathbf{z}_i \cdot \mathbf{z}_j$, where $\operatorname{argmax}_{s \in \mathcal{S}}(\cdot)$ returns a subset of the top- k items. By aggregating neighbor embeddings with weight kernel $\varphi(\cdot)$, it obtains the new embedded representation of samples after mean-shift: $\hat{\mathbf{z}}_i = \frac{\sum_{\mathbf{z}_j \in \mathcal{N}(\mathbf{z}_i)} \varphi(\mathbf{z}_j - \mathbf{z}_i) \mathbf{z}_j}{\left\| \sum_{\mathbf{z}_j \in \mathcal{N}(\mathbf{z}_i)} \varphi(\mathbf{z}_j - \mathbf{z}_i) \mathbf{z}_j \right\|}$. \mathcal{L}_{CMS} and \mathcal{L}_{GCD} are formally approximate.

A.3 SimGCD

SimGCD [57] constructs a prototype classifier $\mathbf{C} = \{\mathbf{c}_1, \dots, \mathbf{c}_{K_{\text{known}}+K_{\text{novel}}}\}$ for both known and unknown classes. It obtains the posterior probability $\mathbf{p}_i^{(k)} = \frac{\exp(\mathbf{h}_i^\top \mathbf{c}_k) / \tau}{\sum_{k'} \exp(\mathbf{h}_i^\top \mathbf{c}'_{k'}) / \tau}$ in a similar way to FixMatch and uses cross-entropy loss $\mathcal{L}_{\text{SimGCD}}^l = \frac{1}{|\mathcal{B}^l|} \sum_{i \in \mathcal{B}^l} \ell(y_i, \mathbf{p}_i)$ on labeled samples. Self-distillation and entropy regularization $\mathcal{L}_{\text{SimGCD}}^u = \frac{1}{|\mathcal{B}|} \ell(\mathbf{p}', \mathbf{p}_i) - \lambda_e H(\frac{1}{2|\mathcal{B}|} \sum_{i \in \mathcal{B}} (\mathbf{p}_i + \mathbf{p}'_i))$ are performed using augmented samples with probability \mathbf{p}'_i .

B Proofs of Theorem

Lemma 1. Given non-negative values p_i such that $\sum_{i=1}^n p_i = 1$, the entropy function $H(p_1, \dots, p_n) = -\sum_{i=1}^n p_i \log p_i$ is strictly concave. Furthermore, it is upper-bounded by $\log n$, as demonstrated by the inequality,

$$\log n = H(1/n, \dots, 1/n) \geq H(p_1, \dots, p_n) \geq 0. \quad (7)$$

Proof B.1. Refer to Section D.1 in [36].

Lemma 2. The Kullback-Leibler (KL) divergence between two zero-mean, d -dimensional multivariate Gaussian distributions can be formulated as follows,

$$\begin{aligned} & D_{\text{KL}}(\mathcal{N}(0, \Sigma_1) \| \mathcal{N}(0, \Sigma_2)) \\ &= \frac{1}{2} \left[\operatorname{tr}(\Sigma_2^{-1} \Sigma_1) - d + \log \frac{|\Sigma_2|}{|\Sigma_1|} \right]. \end{aligned} \quad (8)$$

Proof B.2. Refer to Section 9 in [13].

Theorem 2. For a given [cls] autocorrelation $\mathcal{A} = \mathbf{CLS}^\top \mathbf{CLS} / N \in \mathbb{R}^{d \times d}$ of rank k ($\leq d$),

$$\log(\operatorname{rank}(\mathcal{A})) \geq \hat{H}(\mathcal{A}) \quad (9)$$

where equality holds if the eigenvalues of \mathcal{A} are uniformly distributed with $\forall_{j=1}^k \lambda_j = 1/k$ and $\forall_{j=k+1}^d \lambda_j = 0$.

Proof B.3.

$$\log(\text{rank}(\mathcal{A})) = \log(k) \quad (10)$$

$$\geq H(\lambda_1, \dots, \lambda_k) \text{ (by Lemma 1)} \quad (11)$$

$$= -\sum_{j=1}^k \lambda_j \log \lambda_j \quad (12)$$

$$= -\sum_{j=1}^d \lambda_j \log \lambda_j \quad (13)$$

$$= \hat{H}(\mathcal{A}). \quad (14)$$

According to Lemma 1, the inequality (11) attains equality if and only if $\lambda_j = \frac{1}{k}$ for all $j = 1, 2, \dots, k$. Equation (13) adheres to the convention that $0 \log 0 = 0$, as per the definition in [49].

More details about the definition of $\lambda=1$. Suppose we have a set of n normalized vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$, where the second-order norm (or length) of each vector is 1, that is, $\|\mathbf{v}_i\| = 1$ for all i . The autocorrelation matrix \mathbf{A} of these vectors is defined as:

$$\mathbf{A} = \frac{1}{n} \sum_{i=1}^n \mathbf{v}_i \mathbf{v}_i^T \quad (15)$$

Here, $\mathbf{v}_i \mathbf{v}_i^T$ is the outer product of the vector \mathbf{v}_i with itself, which is a rank-1 matrix. The autocorrelation matrix \mathbf{A} is the average of these outer product matrices.

Next, we need to find the eigenvalues of \mathbf{A} . Since \mathbf{v}_i is normalized, $\mathbf{v}_i^T \mathbf{v}_i = 1$. This means that each \mathbf{v}_i is an eigenvector of \mathbf{A} with the corresponding eigenvalue of $\frac{1}{n}$. This is because:

$$\mathbf{A} \mathbf{v}_i = \frac{1}{n} \left(\sum_{j=1}^n \mathbf{v}_j \mathbf{v}_j^T \right) \mathbf{v}_i = \frac{1}{n} \sum_{j=1}^n \mathbf{v}_j (\mathbf{v}_j^T \mathbf{v}_i) = \frac{1}{n} \mathbf{v}_i (\mathbf{v}_i^T \mathbf{v}_i) = \frac{1}{n} \mathbf{v}_i \cdot 1 = \frac{1}{n} \mathbf{v}_i \quad (16)$$

So, the eigenvalue corresponding to each \mathbf{v}_i is $\frac{1}{n}$. Since \mathbf{A} is a rank- n matrix (assuming the vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ are linearly independent), it has n eigenvalues. We already know that n of the n eigenvalues are $\frac{1}{n}$. Therefore, the sum of all the eigenvalues of \mathbf{A} is:

$$\text{sum of eigenvalues} = \frac{1}{n} + \frac{1}{n} + \dots + \frac{1}{n} = n \cdot \frac{1}{n} = 1 \quad (17)$$

C Theoretically Necessary of GCD with MTMC

GCD is a semi-supervised learning scheme and MTMC is theoretically necessary for GCD. We conduct a comprehensive analysis and derivation from High-Dimensional Probability perspectives (with special consideration given to the more general cases where the number of points P is large and the dimension D is high).

GCD aims to cluster the embeddings of samples from the same category as closely as possible, regardless of whether they are from known or unknown classes. That is each cluster center lies on the hypersphere, and the distribution of the centers on the hypersphere is made as uniform as possible. More formally, this goal can be replaced by two definitions in previous studies [15, 54].

Definition 1 (Perfect Reconstruction). A network f_θ is Perfect Reconstruction if $\forall \mathbf{x} \in \mathcal{X}, \forall t^{(1)}, t^{(2)} \in \mathcal{T}, \mathbf{z}^{(1)} = f_\theta(t^{(1)}(\mathbf{x})) = f_\theta(t^{(2)}(\mathbf{x})) = \mathbf{z}^{(2)}$, where \mathcal{T} is a set of data augmentations such as color jittering, cropping, flipping, etc. The dataset is $\mathbf{x}_{1:P}$ with P samples, and the set after applying $1 : K$ augmentation methods is $t^{(1)}(\mathbf{x}_p), \dots, t^{(K)}(\mathbf{x}_p)$.

Definition 2 (Perfect Uniformity). $p(Z)$ is the distribution over the network representations induced by the data and transformation sampling distributions. If $p(Z)$ is a uniform distribution on the hypersphere, then the network f_θ achieves Perfect Uniformity.

Intuitively, perfect reconstruction means that the network maps all views of the same data to the same embedding, while perfect uniformity means that these embeddings are uniformly distributed on the hypersphere. For brevity, we denote the centroid embedding of the class token representing the p -th sample under different augmentations as \mathbf{z}_p . We prove the following: A network that simultaneously achieves perfect reconstruction and perfect uniformity achieves a lower bound of what MTMC has, that is, it provides the lowest probability of \mathcal{L}_{MTMC} .

Proposition 1. Suppose that, $\forall p \in [P], \mathbf{c}_p^T \mathbf{c}_p \leq 1$. Then, $0 \leq \|C\|_* \leq \sqrt{P \min(P, D)}$.

Proof. Let $\sigma_1, \dots, \sigma_{\min(P,D)}$ denote the singular values of C , so that $\|C\|_* = \sum_{i=1}^{\min(P,D)} \sigma_i$. The lower bound follows by the fact that singular values are nonnegative. For the upper bound, we have

$$\sum_{i=1}^{\min(P,D)} \sigma_i^2 = \text{Tr} [CC^T] = \sum_{n=1}^P \mathbf{c}_p^T \mathbf{c}_p \leq P \quad (18)$$

Then, by Cauchy-Schwarz on the sequences $(1, \dots, 1)$ and $(\sigma_1, \dots, \sigma_{\min(P,D)})$, we get

$$\sum_{i=1}^{\min(P,D)} \sigma_i \leq \sqrt{\left(\sum_{i=1}^{\min(P,D)} 1 \right) \left(\sum_{i=1}^{\min(P,D)} \sigma_i^2 \right)} \leq \sqrt{\min(P, D)P}. \quad (19)$$

Proposition 2. Let f_θ achieve perfect reconstruction. Then, $\|\mathbf{c}_p\|_2 = 1 \forall n$.

Proof. Because f_θ achieves perfect reconstruction, $\forall n, \forall t^{(1)}, t^{(2)}, \mathbf{z}_p^{(1)} = \mathbf{z}_p^{(2)}$. Thus $\mathbf{c}_p = (1/K) \sum_k \mathbf{z}_p^{(k)} = (1/K) \sum_k \mathbf{z}_p^{(1)} = \mathbf{z}_p^{(1)}$, and since $\|\mathbf{z}_p^{(1)}\|_2 = 1$, we have $\|\mathbf{c}_p\|_2 = 1$.

Theorem 1. Let $f_\theta : \mathcal{X} \rightarrow \mathbb{S}^D$ be a network that achieves perfect reconstruction and perfect uniformity. Then f_θ achieves the lower bound of \mathcal{L}_{MTMC} with high probability. Specifically:

$$\|C\|_* = \begin{cases} P(1 - O(P/D)) & \text{if } P \leq D \\ \sqrt{PD}(1 - O(D/P)) & \text{if } P \geq D \end{cases} \quad (20)$$

with high probability in $\min(P, D)$.

This demonstrates that the MTMC loss can be minimized by minimizing the distances of all embeddings corresponding to the same datum and maximizing the distances of all samples' centers.

The above derivations and analyses based on High-Dimensional Probability demonstrate, the **theoretical strong correlation** of MTMC and GCD (as a type of semi-supervised learning).

D More Analysis

D.1 Impact of embedding quality

In Table 1, the accuracy gains on the CIFAR10 and Herbarium19 datasets are insignificant. We use this as a starting point to analyze the conflict between enhancing feature completeness and low embedding quality in GCD. DINO, through self-supervision, already has a good feature representation capability, but due to the distribution of data, its embedding quality still be low. One source of low quality is the data size, and the other is data semantics.

(1) Specifically, when the small-sized CIFAR10 images are interpolated and input into ViT, the high-frequency information is lost. For example, when identifying animal categories, the low-frequency features such as the outline of the animal may be captured relatively well, but the detailed features such as the texture and eyes of the animal (high-frequency features) are difficult to accurately extract. In this case, the model can only cluster through some shortcut information, rather than accurately clustering based on the complete intra-class features. Since the manifold dimension of the low-frequency features is relatively low, it is unable to fully capture the diversity and complexity within the class. Therefore, enhancing the completeness of the intra-class representation on small-sized data is challenging.

(2) Herbarium19 is a large-scale herbal plant recognition dataset, which is not in the model's training data and inherently cannot provide highly discriminative representations. Additionally, the large number of categories makes the decision boundary more chaotic, and existing GCD schemes cannot cluster well. Therefore, enhancing the completeness of intra-class representation on overly low-quality embeddings is not feasible, as the overlap of feature spaces across categories is too large, and samples within a cluster come from multiple categories.

D.2 Analysis of VICReg and CorInfoMax

Compared to the two optimization directions, VICReg and CorInfoMax, MTMC offers a smoother and more uniform convergence of feature values, addressing some key limitations in both methods (Figure 7). VICReg, as a variance-based regularization approach, promotes feature variance and decorrelation but lacks explicit emphasis on intra-class representation completeness. This results in less expressive class boundaries and less

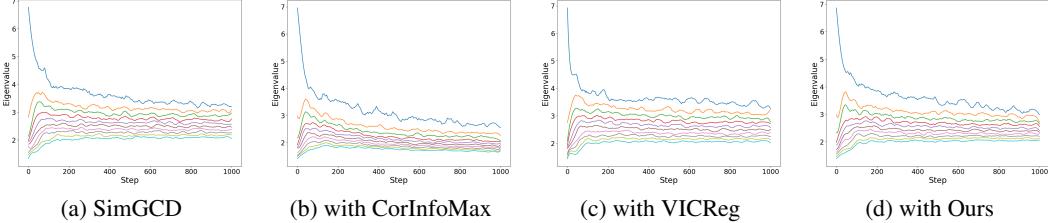


Figure 7: Trends in the top 10 singular values as the number of training steps grows.

effective fine-grained category separation. CorInfoMax, on the other hand, focuses on maximizing mutual information between features and their target distribution but does not sufficiently prevent dimensional collapse or guarantee richer intra-class representations. Both methods, while effective in some contexts, fail to fully capture the complex, high-dimensional structure of the data.

In contrast, MTMC directly targets the manifold capacity of class tokens, ensuring that intra-class representations remain complete and informative. By maximizing the nuclear norm of the class token’s singular values, MTMC ensures that feature values converge uniformly, without the collapse seen in other methods. This leads to more robust and accurate clustering, particularly when discovering novel categories. The smooth convergence of MTMC reflects its ability to optimize representation quality while maintaining high inter-class separability, which is critical for open-world learning tasks.

E Related Works

E.1 Generalized Category Discovery

Generalized category discovery [50, 64, 57, 10] is crucial for identifying and classifying both known and new categories in a dataset, expanding beyond traditional supervised learning to recognize new classes not seen during training. The pioneering work [50] establishes a framework that employs semi-supervised k-means clustering. Following this initial proposition, SimGCD [57] is introduced as a parametric classification approach that utilizes entropy regularization and self-distillation. Expanding on these concepts, CMS [10] is proposed, enhancing representation learning through mean-shift based clustering. Moreover, a deep clustering approach [64] emerges that dynamically adjusts the number of prototypes during inference, facilitating an adaptive discovery of new categories. Most recently, ActiveGCD [34] actively selects samples from unlabeled data to query for labels, with the aim of enhancing the discovery of new categories through an adaptive sampling strategy. Happy [33] explores Continual Generalized Category Discovery (C-GCD), addressing the conflict between discovering new classes and preventing forgetting of old ones through hardness-aware prototype sampling and soft entropy regularization. Each of these contributions addresses the multifaceted challenges of representation learning, category number estimation, and label assignment, redefining the frontiers of open-world learning. Regardless of the flourishing development of GCD, their focus remains on compact clustering, neglecting the integrity of intra-class representation. Our goal is to empower any GCD scheme with concise means to promote the non-collapse representation of each sample, thus shaping more accurate decision boundaries.

E.2 Dimensional Collapse

This Dimensional collapse [19, 4, 45, 25] occurs when the learned embeddings tend to concentrate within a lower-dimensional subspace rather than dispersing throughout the entire embedding space, thereby limiting the representations’ capacity for diversity and expressiveness. DirectCLR [25] presents a direct optimization of the representation space, sidestepping the need for a trainable projector, which inherently mitigates the risk of dimensional collapse by promoting a more even distribution of embeddings across the space. Complementing this, the whitening approach [48] standardizes covariance matrices through whitening techniques, ensuring that each dimension contributes equally to the representation, thus preventing any subset of dimensions from dominating the learning process. Similarly, the non-contrastive learning objective [6] for collaborative filtering avoids data augmentation and negative sampling, focusing on alignment and compactness within the embedding space to prevent dimensional collapse. The Bregman matrix divergence [63] further fortifies the fight against dimensional collapse by minimizing the distance between covariance matrices and the identity matrix, ensuring a uniform distribution of embeddings and directly countering the concentration of information along certain dimensions. Moreover, random orthogonal projection image modeling [20] provides a preventative measure against dimensional collapse by modeling images with random orthogonal projections, which promotes the exploration of a wide range of features and discourages the concentration on a limited subset of dimensions. Rather than directly addressing the issue of dimensional collapse, we focus on maximizing token manifold

capacity to align the radius and dimensions of the manifold with the rich distribution of the real world. This approach also unravels the sample-level dimensional collapse.