LEARNING BIDIRECTIONAL GLOBAL-TO-LOCAL RERANKING FOR GENERALIZED CATEGORY DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalized Category Discovery (GCD) aims to identify both known and novel categories in an unlabeled image collection by exploiting supervision from a subset of known categories. Recent approaches improve representation learning by incorporating neighboring samples from a feature bank. However, these methods typically rely on predefined similarity measures for neighborhood construction, which may retrieve semantically unrelated instances and consequently provide unreliable learning signals. To address this limitation, we propose **G**lobal–**Lo**cal **Re**ranking Network (**GLoRe**), a learnable bidirectional ranking framework for GCD. Rather than treating feature similarity as a fixed indicator of semantic relevance, GLoRe learns to assess the relationship between a query and its retrieved candidates and uses this information to refine their ranking. By learning semantic relationships instead of fixed category predictions, GLoRe can transfer the learned ranking function from labeled examples to unlabeled instances belonging to both known and novel categories. Extensive experiments on fine-grained and coarse-grained benchmarks demonstrate that GLoRe achieves state-of-the-art performance on fine-grained datasets while remaining competitive with state-of-the-art methods on coarse-grained benchmarks. Our code will be made publicly available upon acceptance.

1 INTRODUCTION

Computer vision supports applications in healthcare (Das et al., 2025; Feng et al., 2025; Hsu & Tseng, 2025), autonomous perception (Riz et al., 2023; Wong et al., 2020; Xia et al., 2026), and environmental monitoring (Kareinen et al., 2024; Chen et al., 2026). Despite advances in visual representation learning, supervised recognition relies on annotated datasets whose acquisition is costly and difficult to scale, particularly when expert knowledge is required (Yang et al., 2017). Pseudo-labeling exploits unlabeled data to reduce this burden (Sohn et al., 2020; Pham et al., 2021), but conventional supervised and semi-supervised methods assume a predefined set of categories (Scheirer et al., 2012; Cao et al., 2021). This assumption is restrictive in real-world settings, where unlabeled data may contain previously unseen semantic classes, motivating methods that jointly recognize known categories and discover novel ones from partially labeled data (Vaze et al., 2022).

Novel Class Discovery (NCD) (Han et al., 2019) addresses this problem by transferring knowledge from labeled known categories to discover novel categories in an unlabeled dataset. However, NCD assumes that the labeled and unlabeled sets contain disjoint categories, such that every unlabeled instance belongs to a novel class. Generalized Category Discovery (GCD) (Vaze et al., 2022) relaxes this assumption by considering a mixed unlabeled set containing instances from both known and novel categories. The model must therefore preserve the structure of previously observed categories while simultaneously identifying and organizing instances belonging to previously unseen ones.

A central challenge in GCD is deriving reliable supervision from unlabeled data. Parametric approaches such as SimGCD (Wen et al., 2023) learn classifiers over known and novel categories, while DebGCD (Liu & Han, 2025) and RLCD (Liu et al., 2025) address unreliable self-supervision and prediction bias through debiasing and distribution regularization. A complementary approach exploits relationships among instances to guide representation learning. PromptCAL (Zhang et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2023) refines these relationships through visual prompting and affinity learning, while Contrastive Mean-Shift (CMS) (Choi et al., 2024) incorporates neighborhood aggregation into contrastive learning. Strong pretrained features (Caron et al., 2021) support this process, but the resulting learning signals depend on the semantic relevance of the selected neighbors.

Global similarity can connect images that share an overall appearance but differ in category-defining local attributes. Aggregating such neighbors can blur category boundaries, particularly in fine-grained settings. PartGCD (Wang et al., 2026) demonstrates the value of local evidence by combining explicit part representations with global information for classification. For neighborhood-based learning, the corresponding challenge is to use this evidence to distinguish semantically relevant neighbors from visually similar distractors, despite the absence of labels for novel categories.

We approach neighbor selection through the lens of image retrieval and introduce GLoRe, a bidirectional global-to-local reranking framework for GCD. Given candidates retrieved by global similarity, GLoRe learns a pairwise relevance scorer using supervision from labeled instances.. These bidirectional interactions combine global semantic context with discriminative local details, enabling more informed neighbor selection than global similarity alone. The learned relevance estimates are then used to refine neighborhoods across the labeled and unlabeled data, providing more reliable semantic relationships to guide representation learning. We also introduce several modifications to CMS Choi et al. (2024), including a momentum encoder that mitigates stale memory-bank features and KoLeo Sablayrolles et al. (2018) regularization as a feature-space analogue of SimGCD's Wen et al. (2023) mean-entropy maximization, promoting representation diversity without requiring a classification layer. GLoRe was extensively evaluated on both fine- and coarse-grained GCD benchmarks, demonstrating consistent and substantial improvements over strong existing methods.

Our contributions are as follows:

- We introduce GLoRe, a learnable reranking framework that uses bidirectional global-to-local interactions to refine neighborhoods before mean-shift aggregation.
- We improve feature-bank learning with a dynamic momentum schedule and KoLeo regularization, promoting feature stability and diversity.
- We demonstrate substantial gains on fine-grained GCD benchmarks and competitive performance on coarse-grained benchmarks.

2 Related Work

2.1 Generalized Category Discovery

GCD aims to categorize unlabeled instances from both known and novel classes using a partially labeled dataset (Vaze et al., 2022). Early baselines adapt novel category discovery methods, including RankStats+ (Han et al., 2020) and UNO+ (Fini et al., 2021), while GCD combines contrastive representation learning with semi-supervised k-means. SimGCD (Wen et al., 2023) establishes a strong parametric alternative through self-distillation and entropy regularization. Subsequent approaches address learning biases through distribution-guided debiasing in DebGCD (Liu & Han, 2025) and reciprocal learning with class-wise distribution regularization in RLCD (Liu et al., 2025). Other methods exploit complementary structure: PromptCAL (Zhang et al., 2023) combines visual prompting with affinity learning, while SPTNet (Wang et al., 2024) alternates spatial prompt learning and model adaptation. InfoSieve (Rastegar et al., 2023), SelEx (Rastegar et al., 2024), and SEAL (He et al., 2025) exploit category hierarchies through self-coding, hierarchical pseudo-labeling, and semantic-aware learning, respectively. MOS (Peng et al., 2025) models object–scene associations, whereas PartGCD (Wang et al., 2026) learns discriminative part representations to capture subtle differences between fine-grained categories. Most directly related to our approach, Contrastive Mean-Shift (CMS) (Choi et al., 2024) incorporates neighborhood aggregation into contrastive learning by aligning mean-shifted representations. Its learning signal therefore depends on the semantic relevance of retrieved neighbors, which global similarity alone may not reliably capture in fine-grained settings. Building on CMS, GLoRe uses a learned bidirectional global-to-local scorer to refine these neighborhoods, exploiting local visual evidence to improve semantic neighbor selection for representation learning.

2.2 RANKING IN RETRIEVAL

Visual retrieval commonly uses global descriptors to identify candidates, followed by reranking through more detailed matching (Tan et al., 2021). Contextual methods such as k-reciprocal encoding (Zhong et al., 2017) refine this ordering using reciprocal-neighbor relationships and Jaccard distance. Learning-to-rank approaches instead learn relevance through pointwise candidate scoring, pairwise preferences, as in RankNet (Burges et al., 2005), or listwise objectives, as in ListNet (Cao et al., 2007). GLoRe adopts pointwise scoring with binary same-category supervision for each query–candidate pair, without requiring ground-truth candidate permutations. Relevant architectural precedents include Reranking Transformers (RRTs) (Tan et al., 2021), which jointly process global and local descriptors to learn image-pair matching scores, and R^{2}Former (Zhu et al., 2023), which combines local feature correlations, attention values, and spatial coordinates for visual place recognition. These methods motivate learned interactions between global and local visual evidence. In GLoRe, relevance denotes semantic category agreement, and reranking determines the neighbors used during GCD training. A bidirectional global-to-local scorer matches the query's global representation against the candidate's local tokens and vice versa. Trained on labeled instances from known categories, it refines neighborhoods for unlabeled instances from both known and novel categories.

3 NOTATION AND PRELIMINARIES

Let $\mathcal{C}^k$ and $\mathcal{C}^n$ denote the disjoint sets of known and novel classes, respectively, with $\mathcal{C} = \mathcal{C}^k \cup \mathcal{C}^n$. Training uses a labeled set $\mathcal{D}_L = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_L}$, where $y_i \in \mathcal{C}^k$, and an unlabeled set $\mathcal{D}_U = \{\mathbf{x}_i\}_{i=1}^{N_U}$. Only a subset of known-class samples is labeled; the remaining known-class samples and all novel-class samples form $\mathcal{D}_U = \mathcal{D}_U^k \cup \mathcal{D}_U^n$. During training, neither the labels nor the known/novel membership of unlabeled instances are provided. The objective is to exploit labeled supervision and unlabeled data structure to learn representations that group known and novel instances by semantic category at inference. Following CMS (Choi et al., 2024), we encode two augmented views, x_i and x_i^+, using a DINOv1-pretrained ViT-B/16 backbone f_θ, obtaining final-layer CLS descriptors $\mathbf{c}_i$ and $\mathbf{c}_i^+$ and patch descriptors $\mathbf{P}_i$ and $\mathbf{P}_i^+$. An MLP projection head h_ϕ produces the normalized descriptor $\mathbf{z}_i = h_\phi(\mathbf{c}_i)/\|h_\phi(\mathbf{c}_i)\|_2 \in \mathbb{R}^{768}$, with $\mathbf{z}_i^+$ defined analogously, for contrastive learning and initial retrieval.

4 METHODOLOGY

4.1 OVERVIEW

GLoRe encodes two augmented views with an online network, while an EMA-updated momentum encoder maintains a feature bank. Each image-query retrieves the top-m similarity-based candidates from the feature bank. A bidirectional reranker refines the retrieved candidates through complementary global-to-local interactions. The highest-reranked neighbors are aggregated and combined with the query into a mean-shifted representation, optimized for cross-view consistency and discriminative structure. Figure 1 illustrates the framework.

4.2 MOMENTUM BANK AND CANDIDATE RETRIEVAL

We update a momentum encoder $(f_{\bar{\theta}}, h_{\bar{\phi}})$ solely by EMA (He et al., 2020) after each online update, gradually increasing $\mu \in [0, 1)$ to smooth its reference features. The bank $\mathcal{B}_t = \left\{(\bar{\mathbf{z}}_j, \bar{\mathbf{c}}_j, \bar{\mathbf{P}}_j) : j \in \mathcal{I}_t\right\}$ stores normalized projections together with CLS and patch descriptors for image indices $\mathcal{I}_t$; bars indicate momentum features computed at insertion or the latest refresh. The bank follows a first-in, first-out policy, inserting the momentum features of each processed image and evicting the oldest entry once the bank is full. For each query view, cosine similarity between $\mathbf{z}_i$ and $\bar{\mathbf{z}}_j$ retrieves m candidates $\mathcal{N}_i$. Reranking uses the online query descriptors $(\mathbf{c}_i, \mathbf{P}_i)$ and bank candidate descriptors $(\bar{\mathbf{c}}_j, \bar{\mathbf{P}}_j)$, $j \in \mathcal{N}_i$.

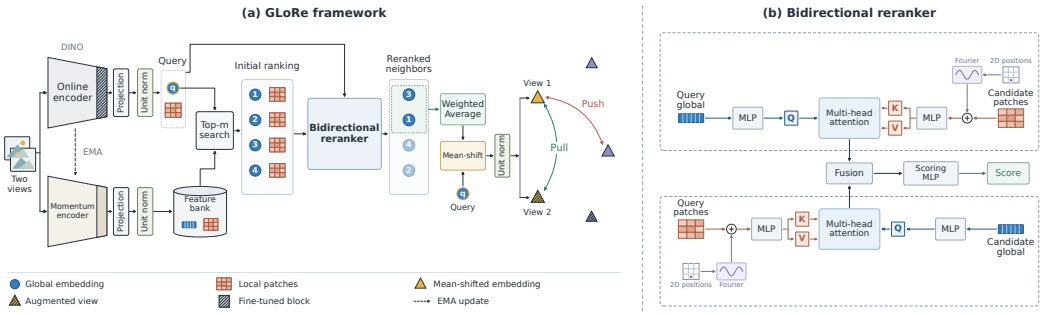


Figure 1: **Overview of GLoRe. (a)** The online encoder extracts query representations, while an EMA-updated momentum encoder populates the feature bank. Initial top-m neighbors are retrieved and reordered by the proposed bidirectional reranker before their aggregation and blending with the query to form the mean-shifted representation. **(b)** The reranker captures complementary global-to-local relations by letting the query global representation attend to candidate patches and, symmetrically, the candidate global representation attend to query patches. The resulting directional features are fused to predict the final relevance score.

4.3 Attention-Guided Retrieval of Local Representations

We select n salient patch descriptors per image using the head-averaged CLS-to-patch attention from the final momentum-backbone block:

$$\mathbf{a}_j = \frac{1}{H}\sum_{h=1}^{H}\left[\mathrm{softmax}\left(\frac{\mathbf{Q}_j^{(h)}(\mathbf{K}_j^{(h)})^\top}{\sqrt{d_\mathrm{k}}}\right)\right]_{0,\,1:N} \in \mathbb{R}^N, \tag{1}$$

where H is the number of attention heads, d_k is the query/key dimension per head, and softmax operates over key tokens. The indices $\mathcal{I}_j = \mathrm{TopK}(\mathbf{a}_j, n)$ select the corresponding rows of the final-layer patch matrix $\bar{\mathbf{P}}_j \in \mathbb{R}^{N\times d}$, with $d = 768$, retaining their original grid coordinates for spatial encoding.

For each selected patch p, we add a Fourier positional encoding (Tancik et al., 2020) to obtain $\widetilde{\mathbf{p}}_{j,p} = \bar{\mathbf{P}}_j[p,:] \oplus \boldsymbol{\gamma}(u_{j,p}, v_{j,p})$, where $(u_{j,p}, v_{j,p})$ are its normalized coordinates in the encoded image's patch grid, $\boldsymbol{\gamma}(u_{j,p}, v_{j,p}) \in \mathbb{R}^d$ maps these coordinates using 192 bands, and $\oplus$ denotes elementwise addition. Stacking the selected descriptors yields $\widetilde{\mathbf{P}}_j \in \mathbb{R}^{n\times d}$. The positional encoding is added before the reranker's key and value projections, and the same construction is applied to the query and every retrieved candidate.

4.4 Bidirectional Global-to-Local Reranking

The reranker evaluates each query-candidate pair (i, j) using two multi-head cross-attention operations. All input features are detached, preventing reranker gradients from reaching the encoder and projection head. The query's global descriptor $\mathbf{c}_i$ attends to the candidate's local descriptors $\widetilde{\mathbf{P}}_j$, while the candidate's global descriptor $\bar{\mathbf{c}}_j$ attends to the query's local descriptors $\widetilde{\mathbf{P}}_i$, for $j \in \mathcal{N}_i$. Linear transformations project global descriptors into queries and local descriptors into keys and values.

$$\mathbf{A}_{i\to j} = \mathrm{softmax}\left(\frac{\mathbf{Q}_i(\mathbf{K}_j)^\top}{\sqrt{d_r}}\right)\mathbf{V}_j, \tag{2}$$

$$\mathbf{A}_{j\to i} = \mathrm{softmax}\left(\frac{\mathbf{Q}_j(\mathbf{K}_i)^\top}{\sqrt{d_r}}\right)\mathbf{V}_i. \tag{3}$$

$\mathbf{A}_{i\to j}$ is the attendance in the tranformer block of the reranker where the global representation of the query embedding is attending to the local (patch-level) representations of the j candidate and $\mathbf{A}_{j\to i}$ is the reversed order where the global representation of the j candidate is attending to the local representations of the query. The bidirectional design of the reranker makes the representation to be pair-dependent and it needs information from both the query and the retrieved candidates.

Each softmax is taken over the n local descriptors of the other image. The final output of the transformer block is a projection of both $\mathbf{A}_{i\to j}$ and $\mathbf{A}_{j\to i}$ through a single layer MLP to corresponding $\mathbf{u}_{i\to j}$ $\mathbf{u}_{j\to i}$. We combine the directional descriptors with their elementwise absolute difference and product:

$$\mathbf{r}_{ij} = \Big[\mathbf{u}_{i\to j}; \mathbf{u}_{j\to i}; \big|\mathbf{u}_{i\to j} - \mathbf{u}_{j\to i}\big|; \big(\mathbf{u}_{i\to j} \odot \mathbf{u}_{j\to i}\big)\Big] \in \mathbb{R}^{4\cdot d_r}. \tag{4}$$

Here, $|\cdot|$ and $\odot$ denote elementwise absolute value and multiplication respectively. The absolute difference exposes discrepancies between the two directional descriptors, while the product captures their multiplicative interactions.

Last step of the reranker is the relevance head which is implemented as a two-layer MLP with GELU activation and dropout, producing a scalar relevance logit for each query–candidate pair; applying a sigmoid yields the relevance score R_{ij}, with the scores of the m retrieved candidates collected in $R_i \in [0,1]^m$. To limit the influence of the reranker on the original retrieval, a coefficient $\beta \in [0,1]$ controls the contribution of the learned relevance scores. Let $S_i^{\text{init}} \in [-1,1]^m$ denote the similarity scores between the projected descriptors of the query and the retrieved candidates $(z_i \cdot \bar{z_j})$. The final scores are computed following equation 5, and the candidates are reranked according to the descending order of S_i'.

$$S_i' = S_i^{\text{init}} + \beta R_i \tag{5}$$

Initial retrieval ranks images using projected global descriptors, which can place visually similar images close together even when they belong to different categories. This is especially problematic in fine-grained discovery, where classes may share an overall appearance but differ in subtle local details. GLoRe exposes spatially indexed patch features to the reranker, allowing it to use distinctions retained in the selected descriptors. Its global-to-local attention aggregates local evidence conditioned on the other image's global context: $A_{i\to j}$ in Eq. (2) aggregates candidate patches $\bar{\mathbf{P}}_j$ conditioned on $\mathbf{c}_i$, while $A_{j\to i}$ in Eq. (3) aggregates query patches $\mathbf{P}_i$ conditioned on $\bar{\mathbf{c}}_j$. These directions provide complementary evidence, since features that support a match in one image may not fully characterize the other. The reranker learns whether two images belong to the same class rather than predicting a known category identity. It can therefore be applied to novel-class pairs; the gains in novel-class accuracy provide empirical evidence that its learned comparisons transfer beyond the labeled categories.

4.5 Learning Objective

For the representation losses, let $\mathcal{J}$ index minibatch views, $\mathcal{J}_{\text{L}} \subseteq \mathcal{J}$ denote labeled views, and i^+ index the other view of the same image.

Supervised contrastive learning. For a labeled anchor i, let $\mathcal{P}(i) = \{p \in \mathcal{J}_{\text{L}} \setminus \{i\} : y_p = y_i\}$ contain the other labeled views with the same class label. Following Khosla et al. (2020), the loss for one anchor is

$$\mathcal{L}_{\text{supcon}}^{(i)} = -\frac{1}{|\mathcal{P}(i)|} \sum_{p\in\mathcal{P}(i)} \log \frac{\exp(z_i \cdot z_p/\tau_{\text{s}})}{\sum_{j\in\mathcal{J}_{\text{L}}\setminus\{i\}} \exp(z_i \cdot z_j/\tau_{\text{s}})}, \tag{6}$$

where $\tau_{\text{s}} > 0$ is the temperature. The denominator includes all other labeled views, and the paired view ensures that $\mathcal{P}(i)$ is nonempty.

Unsupervised contrastive learning. Following CMS (Choi et al., 2024), the retained neighbors k, define normalized weights $w_{ij} = \frac{\exp(\mathbf{z}_i \cdot \bar{\mathbf{z}}_j/\tau_{\text{knn}})}{\sum_{\ell\in\mathcal{N}_i^k} \exp(\mathbf{z}_i \cdot \bar{\mathbf{z}}_\ell/\tau_{\text{knn}})}$ and the final mean-shifted descriptor is constructed based on equation 8 of both labeled and unlabeled samples by adopting the decoupled contrastive formulation of Yeh et al. (2022). We formulate the objective as:

$$\mathcal{L}_{\text{CMS}}^{(i)} = -\log \frac{\exp(ms_i \cdot ms_{i^+}/\tau_{\text{u}})}{\sum_{j\in\mathcal{J}\setminus\{i,i^+\}} \exp(ms_i \cdot ms_j/\tau_{\text{u}})}, \tag{7}$$

$$ms_i = \mathrm{norm}\left(\alpha \mathbf{z}_i + (1-\alpha)\sum_{j\in\mathcal{N}_i^k} w_{ij}\bar{\mathbf{z}}_j\right), \tag{8}$$

where $\mathrm{norm}(\cdot)$ denotes ℓ_2 normalization, $\tau_{\mathrm{u}} > 0$ is the temperature of the contrastive objective, $\alpha \in [0,1]$ balances the contribution of the query and its retrieved neighbors, and $\tau_{\mathrm{knn}} > 0$ scales the neighbor similarities to emphasize more relevant neighbors while suppressing less relevant ones, thereby reducing the influence of potentially noisy retrievals.

KoLeo regularization. KoLeo (Sablayrolles et al., 2018) discourages excessive local concentration of the projected descriptors. Let $\mathcal{J}^{(v)}$ contain the v-th view of each image, with $v \in \{1,2\}$, and $\mathbf{z}_i$ the ℓ_2-normalized projected descriptor. With $\varepsilon > 0$ ensuring numerical stability, For each sample, we compute its nearest-neighbor distance with equation 9 within the corresponding view and apply the KoLeo penalty across both views based on equation 10.

$$\rho_i^{(v)} = \min_{j\in\mathcal{J}^{(v)}\setminus\{i\}} \|\mathbf{z}_i - \mathbf{z}_j\|_2, \tag{9}$$

$$\mathcal{L}_{\mathrm{KoLeo}} = -\frac{1}{2}\sum_{v=1}^{2}\frac{1}{|\mathcal{J}^{(v)}|}\sum_{i\in\mathcal{J}^{(v)}}\log(\rho_i^{(v)} + \varepsilon) \tag{10}$$

Supervised relevance estimation. Let $\mathcal{R}_{\mathrm{L}}$ denote the training query–candidate pairs for which both class labels are known. The reranker predicts $s_{ij} = \sigma(\ell_{ij})$ from its relevance logit ℓ_{ij}, where σ is the sigmoid. We use same-class targets and binary cross-entropy:

$$\mathcal{L}_{\mathrm{reranker}} = -\frac{1}{|\mathcal{R}_{\mathrm{L}}|}\sum_{(i,j)\in\mathcal{R}_{\mathrm{L}}}\left[t_{ij}\log s_{ij} + (1-t_{ij})\log(1-s_{ij})\right]. \tag{11}$$

Here, $t_{ij} = \mathbb{I}[y_i = y_j]$ indicates whether the labeled query and candidate belong to the same class, for each $(i,j) \in \mathcal{R}_{\mathrm{L}}$. The input descriptors are detached, so this loss updates only the reranker.

Overall objective. Averaging the contrastive losses over their respective anchors, the complete objective is

$$\mathcal{L}_{\mathrm{total}} = \frac{\lambda}{|\mathcal{J}|}\sum_{i\in\mathcal{J}}\mathcal{L}_{\mathrm{CMS}}^{(i)} + \frac{1-\lambda}{|\mathcal{J}_{\mathrm{L}}|}\sum_{i\in\mathcal{J}_{\mathrm{L}}}\mathcal{L}_{\mathrm{supcon}}^{(i)} + \lambda_K\,\mathcal{L}_{\mathrm{KoLeo}} + \lambda_R\mathcal{L}_{\mathrm{reranker}}, \tag{12}$$

where $\lambda \in [0,1]$ balances the contrastive terms, $\lambda_K = 0.1$ and $\lambda_R = 1$. The momentum encoder is updated exclusively by exponential moving average.

Inference and clustering. After training, we extract the ℓ_2-normalized, 768-dimensional descriptors $\mathbf{z}_i$ produced by the projection head of the online encoder. Following CMS (Choi et al., 2024), we apply Iterative Mean-Shift refinement (IMS) as a post-processing step and obtain the final cluster assignments through agglomerative clustering.

5 Experiments and Results

5.1 Experimental Setup

Datasets. We evaluate our framework on six standard GCD benchmarks, comprising two coarse-grained datasets, CIFAR100 Krizhevsky et al. (2009) and ImageNet100 Deng et al. (2009), and four fine-grained datasets, CUB-200-2011Wah et al. (2011), FGVC Aircraft Maji et al. (2013), Standford Cars Krause et al. (2013), and Herbarium19 Tan et al. (2019). For the known novel class partition, we adopt the splits introduced by (Vaze et al., 2022), which have become the standard evaluation protocol in GCD.

Evaluation protocol. Following standard practice Vaze et al. (2022), we assess model performance using clustering accuracy (ACC). Let y_i and $\hat{y}_i$ denote the ground-truth label and predicted cluster assignment of sample i, respectively. We compute $\text{ACC} = \frac{1}{|\mathcal{D}_u|} \sum_{i=1}^{|\mathcal{D}_u|} \mathbb{I}(y_i = \mathcal{G}(\hat{y}_i))$, where $\mathcal{G}$ denotes the optimal permutation that aligns the predicted cluster assignments with the ground-truth categories.

Table 1: Comparison of GCD methods on fine-grained benchmarks. Results are reported in ACC across the *All*, *Old*, and *New* categories. † denotes methods that fine-tune two backbone blocks.

	CUB			Stanford-Cars			FGVC-Aircraft			Average		
Method	All	Old	New	All	Old	New	All	Old	New	All	Old	New
k-means (MacQueen, 1967)	34.3	38.9	32.1	12.8	10.6	13.8	16.0	14.4	16.8	21.1	21.3	20.9
RankStats+ (Han et al., 2020)	33.3	51.6	24.2	28.3	61.8	12.1	26.9	36.4	22.2	29.5	49.9	19.5
UNO+ (Fini et al., 2021)	35.1	49.0	28.1	35.5	70.5	18.6	40.3	56.4	32.2	37.0	58.6	26.3
GCD (Vaze et al., 2022)	51.3	56.6	48.7	39.0	57.6	29.9	45.0	41.1	46.9	45.1	51.8	41.8
PromptCAL (Zhang et al., 2023)	62.9	64.4	62.1	50.2	70.1	40.6	52.2	52.2	52.3	55.1	62.2	51.7
SimGCD (Wen et al., 2023)	60.3	65.6	57.7	53.8	71.9	45.0	54.2	59.1	51.8	56.1	65.5	51.5
InfoSieve (Rastegar et al., 2023)	69.4	77.9	65.2	55.7	74.8	46.4	56.3	63.7	52.5	60.5	72.1	54.7
SPTNet (Wang et al., 2024)	65.8	68.8	65.1	59.0	79.2	49.3	59.3	61.8	58.1	61.4	69.9	57.5
CMS (Choi et al., 2024)	68.2	76.5	64.0	56.9	76.1	47.6	56.0	63.4	52.3	60.4	72.0	54.6
SelEx† (Rastegar et al., 2024)	73.6	75.3	72.8	58.5	75.6	50.3	57.1	64.7	53.3	63.1	71.9	58.8
DebGCD† (Liu & Han, 2025)	66.3	71.8	63.5	65.3	81.6	57.4	61.7	63.9	60.6	64.4	72.4	60.5
MOS (Peng et al., 2025)	69.6	72.3	68.2	64.6	80.9	56.7	61.1	66.9	58.2	65.1	73.4	61.0
RLCD (Liu et al., 2025)	70.0	79.1	65.4	64.9	79.3	58.0	60.6	62.2	59.8	65.2	73.5	61.1
SEAL (He et al., 2025)	66.2	72.1	63.2	65.3	79.3	58.5	62.0	65.3	60.4	64.5	72.2	60.7
PartGCD (Wang et al., 2026)	68.6	68.9	68.4	65.6	79.5	58.9	59.4	63.5	57.4	64.5	70.6	61.6
GLoRe (Ours)	**75.6**	**79.6**	**73.6**	**74.4**	**87.9**	**68.0**	**64.4**	**67.9**	**62.6**	**71.5**	**78.5**	**68.1**

Implementation Details. Following Vaze et al. (2022), we fine-tune only the final block of a DINO ViT-B/16 backbone Caron et al. (2021), using the projection-head architecture of dimension 768, batch size of 128, SGD as optimizer, contrastive temperatures $\tau_s = 0.07$, $\tau_u = 0.25$, a learning rate of 0.05 for fine-grained and 0.01 for coarse-grained, and neighborhood size $k = 8$. We retrieve $m = 128$ initial candidates and set $\alpha = 0.5$, $\tau_{\text{knn}} = 0.25$, $\lambda = 0.65$, and $\beta = 0.1$, progressively increasing μ from 0.7 to 0.999 over 200 epochs. The extracted salient patches are set to $n = 32$. The reranker is pretrained for five epochs with frozen DINO features and then jointly trained for 200 epochs, using a learning rate of 3×10^{-4} throughout. It projects global and local features to $d_r = 512$, applies four-head bidirectional cross-attention, and maps their fused 2048-dimensional representation to a scalar relevance score. For fine-grained datasets the bank stores two instances from each unique sample, while for coarse-grained datasets the bank stores a single representation per unique image. Joint training follows a five-epoch reranker warm-up with frozen DINO features. IMS is applied for 20 iterations with $\alpha = 0.1$. All experiments use one NVIDIA RTX A6000 GPU.

5.2 Results

Results on SSB datasets. As shown in Tab. 1, GLoRe establishes a new state of the art across all three fine-grained benchmarks. Remarkably, its superiority extends to every individual evaluation setting, demonstrating that the overall gains do not come at the expense of any particular subset. On average, GLoRe obtains accuracies of 71.5%, 78.5%, and 68.1%, surpassing the strongest prior results by 6.3, 5.0, and 6.5 percentage points, respectively. The improvement is particularly pronounced on Stanford Cars, where GLoRe reaches 74.4% on *All* and 68.0% on *New*, exceeding the corresponding previous best results by 8.8 and 9.1 points. It also improves the best *All* accuracy by 2.0 points on CUB and 2.4 points on FGVC-Aircraft. These consistent gains demonstrate the effectiveness of GLoRe in resolving the subtle local differences that distinguish visually similar fine-grained classes.

Results on coarse-grained benchmarks. CIFAR-100 and ImageNet-100 provide a coarse-grained evaluation setting for assessing whether the learned ranking mechanism generalizes beyond subtle inter-class distinctions. Although the proposed reranker is primarily tailored to the more challenging fine-grained setting, GLoRe remains competitive with state-of-the-art methods on both coarse-grained benchmarks. This is evident on CIFAR-100, where GLoRe achieves 83.6%,

Table 2: Comparison of GCD methods on CIFAR-100 and ImageNet-100. Results are reported in ACC across the *All*, *Old*, and *New* categories. The final three columns report the mean accuracy across the two datasets. † denotes methods that fine-tune two backbone blocks.

	CIFAR-100			ImageNet-100			Average		
Method	All	Old	New	All	Old	New	All	Old	New
k-means (MacQueen, 1967)	52.0	52.2	50.8	72.7	75.5	71.3	62.4	63.9	61.1
RankStats+ (Han et al., 2020)	58.2	77.6	19.3	37.1	61.6	24.8	47.7	69.6	22.1
UNO+ (Fini et al., 2021)	69.5	80.6	47.2	70.3	95.0	57.9	69.9	87.8	52.6
GCD (Vaze et al., 2022)	73.0	76.2	66.5	74.1	89.8	66.3	73.6	83.0	66.4
PromptCAL (Zhang et al., 2023)	81.2	84.2	75.3	83.1	92.7	78.3	82.2	88.5	76.8
SimGCD (Wen et al., 2023)	80.1	81.2	77.8	83.0	93.1	77.9	81.6	87.2	77.9
InfoSieve (Rastegar et al., 2023)	78.3	82.2	70.5	80.5	93.8	73.8	79.4	88.0	72.2
SPTNet (Wang et al., 2024)	81.3	84.3	75.6	85.4	93.2	81.4	83.4	88.8	78.5
CMS (Choi et al., 2024)	82.3	**85.7**	75.5	84.7	**95.6**	79.2	83.5	**90.7**	77.4
SelEx† (Rastegar et al., 2024)	82.3	85.3	76.3	83.1	93.6	77.8	82.7	89.5	77.1
DebGCD† (Liu & Han, 2025)	83.0	84.6	79.9	85.9	94.3	81.6	84.5	89.5	80.8
RLCD (Liu et al., 2025)	83.4	84.2	81.9	**86.9**	94.2	**83.2**	**85.2**	89.2	**82.6**
SEAL (He et al., 2025)	82.1	81.7	**83.0**	84.6	90.9	81.3	83.4	86.3	82.2
PartGCD (Wang et al., 2026)	82.4	84.1	79.0	85.8	94.4	81.5	84.1	89.3	80.3
GLoRe (Ours)	**83.6**	85.2	80.5	85.6	94.0	81.4	84.6	89.6	80.9

and ImageNet-100 reaching 85.6% overall ACC, improving the results of CMS. These findings demonstrate that, despite its fine-grained focus, the proposed framework generalizes effectively to coarse-grained category discovery.

Results on Herbarium19. Herbarium19 combines fine-grained distinctions with severe class imbalance, making it particularly challenging for GCD. As shown in Table 3, leading methods improve over SimGCD (Wen et al., 2023) by less than three percentage points in overall accuracy, considerably less than on the SSB benchmarks. Non-parametric approaches, including GCD (Vaze et al., 2022), PromptCAL (Zhang et al., 2023), SelEx (Rastegar et al., 2024), CMS (Choi et al., 2024), and our original GLoRe formulation, generally trail recent classifier-based methods, suggesting that explicit class-level supervision is beneficial in this setting. To investigate this, we augment GLoRe with the classification head and parametric classification objective of SimGCD (Wen et al., 2023) while retaining all other components. This raises overall accuracy from 40.6% to 47.3%, driven by an 8.5-point gain on novel categories, and places GLoRe above the rest of the approaches. These findings suggest that the absence of a parametric classification objective contributes substantially to the original formulation's performance gap, rather than indicating an inherent limitation of reranking. These results reveal a clear trend in favor of parametric classifiers on Herbarium19 (Tan et al., 2019), with their effect appearing particularly pronounced on this benchmark. While its highly long-tailed distribution provides a plausible explanation, establishing the precise mechanism underlying this advantage requires a more targeted investigation.

Table 3: Comparison of GCD methods on Herbarium19. Results are reported in ACC across the *All*, *Old*, and *New* categories. † denotes methods that fine-tune two backbone blocks.

	Herbarium19		
Method	All	Old	New
k-means (MacQueen, 1967)	13.0	12.2	13.4
RankStats+ (Han et al., 2020)	27.9	55.8	12.8
UNO+ (Fini et al., 2021)	28.3	53.7	14.7
GCD (Vaze et al., 2022)	35.4	51.0	27.0
InfoSieve (Rastegar et al., 2023)	41.0	55.4	33.2
PromptCAL (Zhang et al., 2023)	37.0	52.0	28.9
SimGCD (Wen et al., 2023)	44.0	58.0	36.4
SPTNet (Wang et al., 2024)	43.4	58.7	35.2
CMS (Choi et al., 2024)	36.4	54.9	26.4
SelEx† (Rastegar et al., 2024)	39.6	54.9	31.3
DebGCD† (Liu & Han, 2025)	44.7	59.4	36.8
RLCD (Liu et al., 2025)	46.4	**61.2**	38.4
SEAL (He et al., 2025)	46.9	45.8	**48.2**
PartGCD (Wang et al., 2026)	46.1	59.1	39.0
GloRe (ours)	40.6	57.1	31.8
GloRe (ours) with parametric classifier	**47.3**	60.5	40.3

5.3 Ablation Analysis

Component contributions. Table 4 progressively adds the proposed components, with gains measured relative to the preceding configuration that is based on (Choi et al., 2024). The momentum en-

coder improves *All* accuracy by 2.2/1.4 points on CUB/Stanford Cars, supporting a slowly evolving reference bank. KoLeo adds 1.9/10.4 points on *All* and 6.2/13.7 on *New* by penalizing small nearest-neighbor distances to discourage descriptor concentration. Its motivation parallels SimGCD's entropy regularization (Wen et al., 2023), which balances predictions by maximizing the entropy of batch-averaged class probabilities; however, KoLeo acts on feature geometry without explicitly balancing class assignments. The reranker further improves *All* by 5.1/8.0 points and *New* by 3.5/7.8, supporting learned neighborhood selection even for categories absent from its labeled training pairs. IMS adds only 0.2/0.6 points on *All*, showing that most gains precede this final refinement.

Table 4: Ablation study of the proposed components.

				CUB			Stanford Cars		
Momentum	KoLeo	Reranker	IMS	All	Old	New	All	Old	New
×	×	×	×	66.2	76.2	61.0	54.0	74.1	44.2
✓	×	×	×	68.4	77.7	63.7	55.4	74.7	46.0
✓	✓	×	×	70.3	71.1	69.9	65.8	78.5	59.7
✓	✓	✓	×	75.4	79.5	73.4	73.8	86.8	67.5
✓	✓	✓	✓	**75.6**	**79.6**	**73.6**	**74.4**	**87.9**	**68.0**

Number of retained neighbors. Table 5 examines the number of neighbors retained after reranking, with $k \in \{4, 8, 16\}$. Setting $k = 8$ achieves the highest *All* accuracy on both CUB and Stanford Cars, consistent with the neighborhood-size study in CMS (Choi et al., 2024). Smaller values make the latent space of the neighborhood inexpressive, increasing k broadens the effective neighborhood and increases the likelihood of including less relevant or semantically unrelated samples. This pattern suggests a trade-off between collecting sufficient neighborhood evidence and incorporating less relevant candidates.

Reranking coefficient. Table 6 evaluates the coefficient β controlling the contribution of learned relevance scores to the final ranking. Among the tested values, $\beta = 0.1$ achieves the highest accuracy across *All*, *Old*, and *New* categories on both datasets. Increasing β from 0.05 to 0.1 improves *All* accuracy by 0.9 and 4.8 percentage points on CUB and Stanford Cars, respectively, whereas further increasing it to 0.2 reduces accuracy. These results favor a moderate contribution from the reranker, allowing global-to-local evidence to refine neighbor selection while preserving the influence of the initial similarity scores.

Table 5: Ablation study on the number of neighbors k.

	CUB			Stanford Cars		
k	All	Old	New	All	Old	New
4	75.2	78.2	**73.6**	68.1	81.9	61.4
8	**75.6**	79.6	**73.6**	**74.4**	**87.9**	**68.0**
16	71.0	**80.6**	66.2	71.0	86.8	63.3

Table 6: Ablation study of the reranking coefficient β.

	CUB			Stanford Cars		
β	All	Old	New	All	Old	New
0.05	74.7	77.9	73.1	69.6	83.0	63.2
0.1	**75.6**	**79.6**	**73.6**	**74.4**	**87.9**	**68.0**
0.2	72.7	79.2	69.4	72.2	85.5	65.6

6 Conclusion and Future Work

We introduced GLoRe, a framework for GCD that improves neighborhood-based representation learning through a lightweight, learnable reranker. Its bidirectional global-to-local interactions identify semantically relevant neighbors, complemented by KoLeo regularization for feature diversity and a momentum encoder for stable memory-bank representations. GLoRe outperforms the compared methods across all reported metrics on the SSB datasets, demonstrating the value of refining neighborhoods through learned relevance. Although trained only on labeled pairs, the reranker generalizes to novel categories. Future work will extend relevance learning to unlabeled query–candidate pairs to further improve neighborhood quality and category discovery.

AI USE STATEMENT

Generative AI tools were used solely to aid and polish the writing. All AI-assisted text was reviewed by the authors, who take full responsibility for the paper's content.

REFERENCES

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to Rank using Gradient Descent. In *Proceedings of the International Conference on Machine Learning*, 2005. URL https://www.microsoft.com/en-us/research/publication/learning-to-rank-using-gradient-descent/.

Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. *arXiv preprint arXiv:2102.03526*, 2021.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to Rank: From Pairwise Approach to Listwise Approach. In *Proceedings of the International Conference on Machine Learning*, 2007. URL https://www.microsoft.com/en-us/research/publication/learning-to-rank-from-pairwise-approach-to-listwise-approach/.

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*, 2021. URL https://arxiv.org/abs/2104.14294.

Yuyan Chen, Nico Lang, Brian Schmidt, Aditya Jain, Yves Basset, Sara Beery, Maxim Larrivée, and David Rolnick. Open-insect: benchmarking open-set recognition of novel species in biodiversity monitoring. *Advances in Neural Information Processing Systems*, 38, 2026.

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*, 2024. URL https://arxiv.org/abs/2404.09451.

Ankit Das, Chandan Gautam, Pritee Agrawal, Feng Yang, Yong Liu, and Ramasamy Savitha. Medgcd: Generalized category discovery in medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 427–437. Springer, 2025.

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*, pp. 248–255. Ieee, 2009.

Wei Feng, Sijin Zhou, Yiwen Jiang, Feilong Tang, and Zongyuan Ge. Neighbor-guided unbiased framework for generalized category discovery in medical image classification. *IEEE Journal of Biomedical and Health Informatics*, 29(8):5736–5747, 2025.

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*, pp. 9284–9292, 2021.

Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8400–8408. IEEE, 2019.

Kai Han, Sylvestre-Alvise Rebuffi, Sébastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *International Conference on Learning Representations*, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9726–9735. IEEE, 2020.

Zhenqi He, Yuanpei Liu, and Kai Han. SEAL: Semantic-aware hierarchical learning for generalized category discovery. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Benny Wei-Yun Hsu and Vincent S Tseng. Hitrace: Hierarchical class tracing approach for open-set recognition on skin lesions. *IEEE Journal of Biomedical and Health Informatics*, 29(8):5700–5711, 2025.

Joona Kareinen, Annaliina Skyttä, Tuomas Eerola, Kaisa Kraft, Lasse Lensu, Sanna Suikkanen, Maiju Lehtiniemi, and Heikki Kälviäinen. Open-set plankton recognition. In *European Conference on Computer Vision*, pp. 168–184. Springer, 2024.

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.

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*, pp. 554–561, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Duo Liu, Zhiquan Tan, Linglan Zhao, Zhongqiang Zhang, Xiangzhong Fang, and Weiran Huang. Generalized category discovery via reciprocal learning and class-wise distribution regularization. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Yuanpei Liu and Kai Han. DebGCD: Debiased learning with distribution guidance for generalized category discovery. In *International Conference on Learning Representations*, 2025.

James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 281–297. University of California Press, 1967.

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Zhengyuan Peng, Jinpeng Ma, Zhimin Sun, Ran Yi, Haichuan Song, Xin Tan, and Lizhuang Ma. MOS: Modeling object-scene associations in generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11552–11563. IEEE, 2021.

Sarah Rastegar, Hazel Doughty, and Cees G. M. Snoek. Learn to categorize or categorize to learn? self-coding for generalized category discovery. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://openreview.net/forum?id=m0vfXMrLwF.

Sarah Rastegar, Mohammadreza Salehi, Yuki M. Asano, Hazel Doughty, and Cees G. M. Snoek. SelEx: Self-expertise in fine-grained generalized category discovery. In *European Conference on Computer Vision*. Springer, 2024.

Luigi Riz, Cristiano Saltori, Elisa Ricci, and Fabio Poiesi. Novel class discovery for 3d point cloud semantic segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9393–9402. IEEE, 2023.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Spreading vectors for similarity search. *arXiv preprint arXiv:1806.03198*, 2018.

Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7): 1757–1772, 2012.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608, 2020.

Fuwen Tan, Jiangbo Yuan, and Vicente Ordonez. Instance-level Image Retrieval using Reranking Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://arxiv.org/abs/2103.12236.

Kiat Chuan Tan, Yulong Liu, Barbara Ambrose, Melissa Tulig, and Serge Belongie. The herbarium challenge 2019 dataset. *arXiv preprint arXiv:1906.05372*, 2019.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

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*, 2022. URL https://arxiv.org/abs/2201.02609.

Sagar Vaze, Andrea Vedaldi, and Andrew Zisserman. No representation rules them all in category discovery. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Enguang Wang, Zhimao Peng, Zhengyuan Xie, Haori Lu, Fei Yang, and Xialei Liu. Learning part knowledge to facilitate category understanding for fine-grained generalized category discovery. *IEEE Transactions on Multimedia*, 2026. Early Access.

Hongjun Wang, Sagar Vaze, and Kai Han. SPTNet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. In *International Conference on Learning Representations*, 2024.

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*, 2023. URL https://arxiv.org/abs/2211.11727.

Kelvin Wong, Shenlong Wang, Mengye Ren, Ming Liang, and Raquel Urtasun. Identifying unknown instances for autonomous driving. In *Conference on Robot Learning*, pp. 384–393. PMLR, 2020.

Zhongyu Xia, Jishuo Li, Zhiwei Lin, Xinhao Wang, Yongtao Wang, and Ming-Hsuan Yang. Openad: Open-world autonomous driving benchmark for 3d object detection. *Advances in Neural Information Processing Systems*, 38, 2026.

Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 399–407. Springer, 2017.

Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. In *European conference on computer vision*, pp. 668–684. Springer, 2022.

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*, 2023.

Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking Person Re-identification with k-Reciprocal Encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. URL https://arxiv.org/abs/1701.08398.

Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. R2Former: Unified Retrieval and Reranking Transformer for Place Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2304.03410.

APPENDIX

A Dataset Details 15

B DINOv2 on Fine-Grained Benchmarks 15

C Retrieval Quality 16

D Effect of the Momentum Schedule 16

E Sensitivity to Neighbor-Weighting Temperatures 17

F Category Discovery with an Estimated Number of Classes 18

G Effect of the Reranker Architecture 18

H Effect of total number of Patches 19

I Stability Analysis 19

J Computational overhead 20

K Representation Visualizations and Qualitative Results 20

A Dataset Details

We evaluate GLoRe on six established GCD benchmarks spanning coarse-grained recognition, fine-grained category discovery, and long-tailed fine-grained recognition. The partitions are summarized in Table 7.

Generic datasets. CIFAR-100 (Krizhevsky et al., 2009) contains 50,000 training images of size 32×32, evenly distributed across 100 object categories. ImageNet-100 (Deng et al., 2009) comprises 100 categories selected from ImageNet-1K, featuring natural images with diverse backgrounds, viewpoints, and object appearances.

Fine-grained datasets. We use the three datasets in the Semantic Shift Benchmark (SSB): CUB (Wah et al., 2011), which covers 200 bird species; Stanford Cars (Krause et al., 2013), which comprises 196 car categories distinguished by make, model, and year; and FGVC-Aircraft (Maji et al., 2013), which contains 100 aircraft variants. These benchmarks require recognising subtle visual differences between closely related categories. We additionally evaluate on Herbarium19 (Tan et al., 2019), a collection of herbarium specimen images spanning 683 plant species with a long-tailed class distribution.

Table 7: Statistics of the datasets used in our experiments.

	Labelled		Unlabelled	
Dataset	#Num	#Class	#Num	#Class
CIFAR-100 (Krizhevsky et al., 2009)	20.0K	80	30.0K	100
ImageNet-100 (Deng et al., 2009)	31.9K	50	95.3K	100
Herbarium19 (Tan et al., 2019)	8.9K	341	25.4K	683
CUB (Wah et al., 2011)	1.5K	100	4.5K	200
Stanford Cars (Krause et al., 2013)	2.0K	98	6.1K	196
FGVC-Aircraft (Maji et al., 2013)	1.7K	50	5.0K	100

B DINOv2 on Fine-Grained Benchmarks

To assess the effectiveness of GLoRe with stronger pre-trained representations, we evaluate it using DINOv2 (Oquab et al., 2023) on the three SSB benchmarks. As shown in Table 8, GLoRe achieves the highest overall accuracy on all three datasets, outperforming the strongest competing method by 5.0, 6.3, and 1.0 percentage points on CUB, Stanford-Cars, and FGVC-Aircraft, respectively. It also leads across all evaluation splits on CUB and Stanford-Cars, with particularly substantial gains in novel category discovery. On FGVC-Aircraft, GLoRe achieves competitive performance on both old and new categories while obtaining the best overall accuracy. Averaged across the three benchmarks, GLoRe achieves 81.7% overall and 80.3% novel-category accuracy, exceeding the strongest competing averages by 4.8 and 5.8 percentage points, respectively. These results demonstrate that the benefits of GLoRe extend to stronger DINOv2 (Oquab et al., 2023) representations.

Table 8: Comparison with state-of-the-art GCD methods on the SSB benchmarks using DINOv2 pre-trained weights. Results are reported in ACC across the *All*, *Old*, and *New* categories.

	CUB			Stanford-Cars			FGVC-Aircraft			Average		
Method	All	Old	New	All	Old	New	All	Old	New	All	Old	New
k-means (MacQueen, 1967)	67.6	60.6	71.1	29.4	24.5	31.8	18.9	16.9	19.9	38.6	34.0	40.9
GCD (Vaze et al., 2022)	71.9	71.2	72.3	65.7	67.8	64.7	55.4	47.9	59.2	64.3	62.3	65.4
SimGCD (Wen et al., 2023)	71.5	78.1	68.3	71.5	81.9	66.6	63.9	69.9	60.9	69.0	76.6	65.3
μGCD (Vaze et al., 2023)	74.0	75.9	73.1	76.1	91.0	68.9	66.3	68.7	65.1	72.1	78.5	69.0
DebGCD (Liu & Han, 2025)	77.5	80.8	75.8	75.4	87.7	69.5	71.9	76.0	69.8	74.9	81.5	71.7
RLCD (Liu et al., 2025)	78.7	79.5	78.3	79.5	91.8	73.5	72.6	**77.3**	70.3	76.9	82.9	74.0
SEAL (He et al., 2025)	76.7	78.3	75.9	77.7	88.7	72.4	74.6	73.2	**75.3**	76.3	80.1	74.5
PartGCD (Wang et al., 2026)	77.6	80.6	76.1	78.2	88.7	73.1	71.1	75.6	68.8	75.6	81.6	72.7
GLoRe (ours)	**83.7**	**84.5**	**83.3**	**85.8**	**92.7**	**82.4**	**75.6**	76.5	75.1	**81.7**	**84.6**	**80.3**

C RETRIEVAL QUALITY

To quantify the semantic quality of the retrieved neighborhoods, we perform a post-hoc oracle analysis. During training, the model has access only to the prescribed labeled subset, while the ground-truth labels of unlabeled samples are unavailable. After each epoch, we use the ground-truth labels solely to compute P@8, defined as the fraction of the eight retrieved neighbors that belong to the same class as the query. These labels are not used to train the reranker, select neighbors, or choose model checkpoints; therefore, the curves in Figure 2 are diagnostic measurements of neighborhood purity rather than an additional supervision signal.

The diagnostic shows that the reranker substantially improves neighborhood quality, particularly during the initial ten epochs. Averaged across the three query groups, it improves P@8 by 25.47 percentage points on Stanford Cars and by 11.86 points on CUB during this initial phase. The improvement remains clear at convergence, with corresponding mean gains of 8.41 and 5.73 points, respectively. At convergence, the gains are concentrated on unlabeled queries: the reranker improves P@8 by 13.77 and 11.40 points for unlabeled old and new classes on Stanford Cars, and by 8.18 and 8.99 points on CUB. In contrast, the improvement for labeled queries becomes negligible at convergence, reaching only 0.06 points on Stanford Cars and 0.02 points on CUB. This indicates that, once the projector has learned the labeled categories, the reranker primarily improves neighborhood formation for unlabeled samples from both old and new classes.

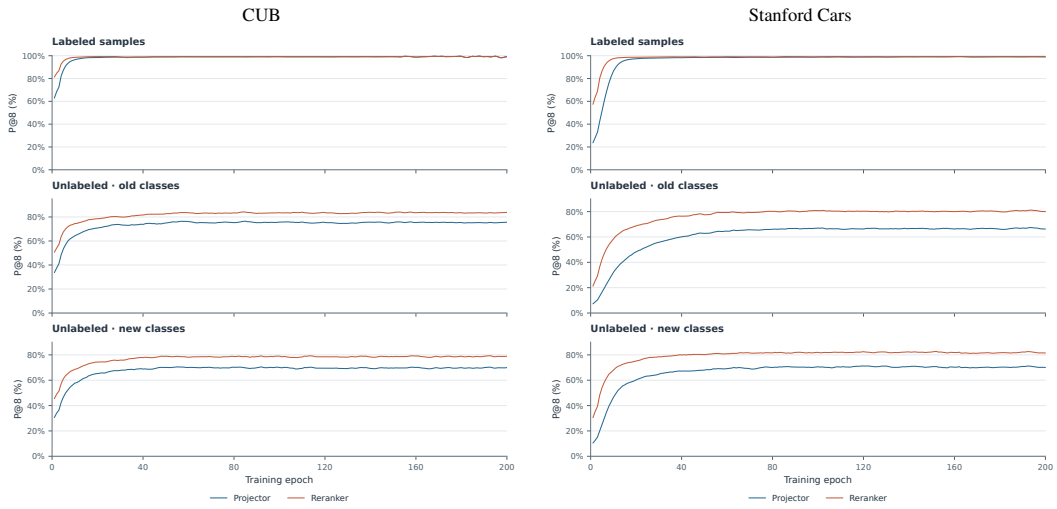


Figure 2: **Post-hoc oracle diagnostic of neighborhood retrieval quality during training.** CUB is shown on the left and Stanford Cars on the right. Within each dataset, the panels show P@8 for labeled queries, unlabeled queries from old classes, and unlabeled queries from new classes, from top to bottom. P@8 is the fraction of the eight retrieved neighbors that share the query's ground-truth class. Ground-truth labels are consulted only after each epoch to compute this diagnostic; labels of unlabeled samples are never provided to the model and do not affect training, retrieval, or model selection. The curves compare retrieval using projected global descriptors (Projector) with the proposed bidirectional reranker (Reranker).

D EFFECT OF THE MOMENTUM SCHEDULE

Table 9 compares a fixed momentum coefficient of 0.999 with schedules increasing from 0.9 or 0.7 to 0.999. Both schedules improve *All* accuracy over fixed momentum, with the $0.7 \rightarrow 0.999$ schedule yielding the strongest results on both benchmarks. It improves *All* accuracy by 1.3 percentage points on CUB and 1.9 points on Stanford Cars, with corresponding gains of 1.7 and 2.9 points on novel categories. Relative to the $0.9 \rightarrow 0.999$ schedule, it further improves *All* accuracy by 0.7 points on each dataset.

The momentum coefficient controls how closely the reference encoder follows the online encoder. Following momentum-based representation learning (He et al., 2020), its backbone parameters are updated as

$$\bar{\theta}_t = \mu_t \bar{\theta}_{t-1} + (1 - \mu_t)\theta_t, \tag{13}$$

with the same update applied to the projection-head parameters $\bar{\phi}_t$. Starting from $\mu_t = 0.7$ assigns a weight of 0.3 to the current online parameters, compared with 0.001 under fixed momentum, allowing the reference encoder to follow early representation changes more closely. As μ_t increases, stronger temporal averaging stabilizes the reference features used for retrieval and relevance estimation. The observed gains, particularly on novel categories, are consistent with the benefit of combining early responsiveness with greater stability later in training.

Table 9: Ablation study of the momentum effect.

	CUB			Stanford Cars		
Momentum	All	Old	New	All	Old	New
0.999	74.3	79.3	71.9	72.5	87.8	65.1
EMA $0.9 \rightarrow 0.999$	74.9	79.1	72.7	73.7	**87.9**	66.9
EMA $0.7 \rightarrow 0.999$	**75.6**	**79.6**	**73.6**	**74.4**	**87.9**	**68.0**

E Sensitivity to Neighbor-Weighting Temperatures

Table 10: Ablation study on the KNN temperature τ_{knn} and the number of retrieved neighbors k.

		CUB			Stanford Cars		
τ_{knn}	k	All	Old	New	All	Old	New
0.1	4	70.8	73.7	69.4	63.7	74.0	58.7
0.1	8	74.4	78.7	72.2	72.3	84.7	66.3
0.1	16	73.3	79.8	70.1	68.4	84.7	60.5
0.25	4	75.2	78.2	**73.6**	68.1	81.9	61.4
0.25	8	**75.6**	79.6	**73.6**	**74.4**	**87.9**	**68.0**
0.25	16	71.0	**80.6**	66.2	71.0	86.8	63.3

Table 10 jointly examines the neighbor-weighting temperature τ_{knn} and the number of retained neighbors k. For a fixed neighborhood $\mathcal{N}_i^k$, the relative weights of two neighbors satisfy

$$\frac{w_{ij}^{knn}}{w_{i\ell}^{knn}} = \exp\left(\frac{\mathbf{z}_i \cdot \bar{\mathbf{z}}_j - \mathbf{z}_i \cdot \bar{\mathbf{z}}_\ell}{\tau_{\text{knn}}}\right), \qquad j, \ell \in \mathcal{N}_i^k. \tag{14}$$

Lower temperatures amplify relative weight differences, concentrating the neighbor contribution to the mean-shifted representation on the most similar descriptors. As $\tau_{\text{knn}} \rightarrow 0^+$, the weights concentrate on the highest-similarity neighbor, or equally among tied maxima. Conversely, larger temperatures produce smoother weights, giving less similar neighbors greater relative influence; as $\tau_{\text{knn}} \rightarrow \infty$, $w_{ij} \rightarrow 1/k$.

With $k = 8$, increasing τ_{knn} from 0.1 to 0.25 improves overall accuracy by 1.2 and 2.1 percentage points on CUB and Stanford Cars, respectively, yielding the highest overall accuracy on both datasets. Increasing k to 16 reduces overall and novel-category accuracy at both temperatures. Moreover, on CUB with $k = 16$, increasing the temperature reduces overall accuracy from 73.3 to 71.0, showing that smoother weighting is not uniformly beneficial. This interaction is consistent with a trade-off between aggregating evidence from multiple neighbors and increasing the influence of potentially less relevant candidates.

F Category Discovery with an Estimated Number of Classes

Table 11 evaluates GLoRe when the number of categories is unknown. Under the shared estimates obtained following Vaze et al. (2022), with $\hat{K} = 231$ for CUB and $\hat{K} = 230$ for Stanford-Cars, GLoRe achieves the highest accuracy across All, Old, and New categories on both datasets. It

surpasses the strongest competing method in overall accuracy by 6.8 and 7.7 percentage points on CUB and Stanford-Cars, respectively, with corresponding novel-category gains of 4.7 and 7.3 points. We additionally evaluate our approach following Choi et al. (2024) approach. The estimates obtained with GLoRe are particularly accurate, recovering the exact number of categories on CUB and closely approximating the true count on Stanford-Cars. In this setting, GLoRe outperforms CMS by 11.2 and 21.9 points in overall accuracy, demonstrating strong performance when ground-truth category counts are unavailable. For the CMS-based estimates, we follow the category-count estimation procedure of Choi et al. (2024), selecting the number of clusters according to clustering accuracy on the labeled validation subset.

Table 11: Comparison on CUB and Stanford-Cars with an estimated number of categories. Results are reported in ACC (%) across the *All*, *Old*, and *New* categories.

	CUB			Stanford-Cars		
	$K = 200$			$K = 196$		
Method	All	Old	New	All	Old	New
(a) Estimation following Vaze et al. (2022) ($\hat{K}_{\text{CUB}} = 231$, $\hat{K}_{\text{Cars}} = 230$)						
GCD (Vaze et al., 2022)	51.1	56.4	48.4	39.1	58.6	29.7
SimGCD (Wen et al., 2023)	61.0	66.0	58.6	49.1	65.1	41.3
SPTNet (Wang et al., 2024)	65.2	71.0	62.3	–	–	–
DebGCD (Liu & Han, 2025)	64.5	68.5	62.5	63.3	78.6	55.8
SEAL (He et al., 2025)	–	–	–	62.4	78.9	54.5
PartGCD (Wang et al., 2026)	66.4	66.5	66.4	64.9	78.7	58.3
GLoRe (ours)	**73.2**	**77.5**	**71.1**	**72.6**	**87.0**	**65.6**
(b) Estimation following Choi et al. (2024)						
CMS (Choi et al., 2024) ($\hat{K}_{\text{CUB}} = 168$, $\hat{K}_{\text{Cars}} = 156$)	64.4	68.2	62.4	51.7	68.9	43.4
GLoRe (ours) ($\hat{K}_{\text{CUB}} = 200$, $\hat{K}_{\text{Cars}} = 190$)	**75.6**	**79.6**	**73.6**	**73.6**	**87.6**	**66.8**

G Effect of the Reranker Architecture

Table 12 compares multi-head self-attention (MHSA), unidirectional cross-attention, and the proposed bidirectional cross-attention. In the MHSA variant, self-attention is applied exclusively to the global embeddings produced by the projection head, without access to patch-level representations. The transformer dimensions, number of attention heads, projection MLPs, and relevance head are kept consistent across variants, with the exception of MHSA, where the transformer block outputs a single 512-dimensional vector. Together with the fixed β, this comparison assesses the attention design under matched architectural hyperparameters.

Table 12: Ablation study of the different architectures of the reranker.

	CUB			Stanford Cars		
Architecture	All	Old	New	All	Old	New
MHSA	73.7	79.4	70.9	61.5	79.2	52.9
Cross-Attention	75.0	**80.1**	72.4	69.4	83.0	62.9
Bidirectional Cross-Attention	**75.6**	79.6	**73.6**	**74.4**	**87.9**	**68.0**

Bidirectional cross-attention achieves the highest *All* and *New* accuracy on both benchmarks, improving novel-category discovery over the unidirectional variant by 1.2 points on CUB and 5.1 points on Stanford Cars. The only exception to its consistent gains across evaluation subsets is CUB *Old*, where unidirectional cross-attention performs 0.5 points better. These results support the effectiveness of incorporating both directional interactions for category discovery.

These differences are consistent with the explicit comparison structure imposed by global-to-local attention. In the MHSA variant, the query, key, and value projections are formed exclusively from the projected global embeddings, such that the interaction is restricted to image-level representations. In contrast, the unidirectional variant uses the query image's CLS descriptor as the query and the candidate's patch descriptors as keys and values. In the unidirectional variant, the query

image's CLS descriptor supplies the query projection, while the candidate's patch descriptors supply the keys and values. This produces a candidate-local representation conditioned on the query's global context. The reverse direction additionally conditions the query's local descriptors on the candidate's CLS descriptor. The relevance head can therefore compare evidence from both images through the directional representations $\mathbf{u}_{i\to j}$ and $\mathbf{u}_{j\to i}$, including their element-wise difference and product. This provides access to complementary local cues that a single directional interaction does not expose.

The effect on neighbor selection can be expressed through the final score margin. For two candidates j and r retrieved for query i,

$$S_{ij} - S_{ir} = \left(S_{ij}^{\text{init}} - S_{ir}^{\text{init}}\right) + \beta\left(R_{ij} - R_{ir}\right). \tag{15}$$

If candidate j initially ranks below r, their order is reversed when $\beta(R_{ij} - R_{ir}) > S_{ir}^{\text{init}} - S_{ij}^{\text{init}}$. Since the sigmoid relevance scores lie in $[0, 1]$, the learned pairwise correction has magnitude at most β. Reranking therefore permits local evidence to resolve sufficiently small global-similarity margins while preserving larger margins. Keeping β fixed compares the attention variants under the same bound on this correction.

H Effect of total number of Patches

Table 13 evaluates the number n of salient patch descriptors retained per image. Selecting $n = 32$ achieves the highest overall and novel-category accuracy on both datasets. Reducing n to 16 lowers overall accuracy, particularly on Stanford Cars, whereas increasing it to 64 substantially degrades performance on CUB while remaining competitive on Stanford Cars. These results suggest that a moderate patch budget balances local evidence coverage with the inclusion of potentially redundant or less informative patches.

Table 13: Ablation study of total number of patches n.

	CUB			Stanford Cars		
Number of Patches	All	Old	New	All	Old	New
16	74.8	80.0	72.2	70.5	85.8	63.1
32	**75.6**	**79.6**	**73.6**	**74.4**	**87.9**	**68.0**
64	63.1	70.1	59.6	73.9	87.6	67.3

I Stability Analysis

Stability Analysis. To assess the robustness of GLoRe to random initialization and stochastic training effects, we repeat each experiment over five independent runs using different random seeds. Table 14 reports the mean performance together with the standard deviation across runs. Overall, GLoRe exhibits stable performance across all evaluated benchmarks, with relatively small variations in *All*, *Old*, and *New* accuracy. The results indicate that the improvements obtained by the proposed framework are consistent across different random seeds and are not dependent on a particular initialization.

Table 14: Stability analysis of GLoRe over five independent runs with different random seeds. Results are reported as mean ± standard deviation.

	GLoRe		
Dataset	All	Old	New
CUB	75.6±0.8	79.6±1.0	73.6±0.7
Stanford Cars	74.4±0.7	87.9±0.9	68.0±0.5
FGVC-Aircraft	64.4±0.6	67.9±0.9	62.6±0.5
CIFAR-100	83.6±0.5	85.2±0.6	80.5±0.4
ImageNet-100	85.6±0.4	94.0±0.7	81.4±0.3
Herbarium19	40.6±0.4	57.1±0.4	31.8±0.3

J COMPUTATIONAL OVERHEAD

Table 15 compares the computational cost of the proposed method with CMS (Choi et al., 2024) on the CUB dataset. Our method introduces approximately 1.97 million additional parameters, increasing the total parameter count by 2.13%. The measured training time per epoch increases from 41.69 to 46.76 seconds, corresponding to an additional 5.07 seconds (12.16%). A single bank-construction takes 29.02 seconds for our method compared with 36.93 seconds for CMS, a reduction of 21.42%.

The frequency of bank construction also differs between the methods. CMS (Choi et al., 2024) reconstructs its feature bank at the beginning of every epoch through a separate feature-extraction pass over the training data. Our method initializes the bank once before training and subsequently updates it with the momentum descriptors of each processed minibatch. This avoids repeated costly full-dataset feature-extraction passes. Nevertheless, for a conservative and fair comparison, we include the additional bank-maintenance overhead in GLoRe's reported per-epoch training cost in the final row of Table 15, despite not reconstructing the full bank at every epoch.

Our method additionally pretrains the reranker for five epochs to adapt its relevance predictions to the pretrained DINO features. This stage requires approximately 18 seconds per epoch, adding a one-time cost of approximately 90 seconds before the main training stage. Comparisons of cumulative training cost should therefore account for both this initial pretraining stage and the different frequencies of bank construction.

Table 15: Computational cost of the proposed method relative to CMS (Choi et al., 2024) on the CUB dataset. Timing values are reported as means $\pm$ standard deviations. Relative changes are computed from the displayed means. Bank-construction time is reported per epoch, and training time per epoch. Reranker pretraining is reported separately in the text.

Metric	CMS	GLoRe	Change (%)
Total parameters (M)	92.37	94.34	+2.13
Bank construction per epoch (s)	36.93 ± 1.28	29.02 ± 1.44	-21.42
Training time per epoch (s)	41.69 ± 0.11	46.76 ± 1.16	$+12.16$
Training epoch + bank construction (s)	78.62 ± 1.28	75.78 ± 1.85	-3.61

K REPRESENTATION VISUALIZATIONS AND QUALITATIVE RESULTS

K.1 QUALITATIVE RESULTS

In Fig. 3 and Fig. 4 we have an illustration of retrieval results on CUB-200-2011 and Standford Cars correspondingly. The figures depict that the projector-based retrievals are relatively sparse in terms of correct neighbors and are predominantly accurate for labeled samples from known categories. In contrast, the proposed reranker consistently promotes semantically related candidates to higher ranks across labeled-known, unlabeled-known, and unlabeled-novel queries, substantially increasing the concentration of correct retrievals within the top-k neighborhood. This behavior indicates that the learned global-to-local interactions provide a more discriminative similarity signal than the projected global descriptor alone.

K.2 T-SNE VISUALIZATIONS

Figure 5 compares t-SNE projections of the representations learned by DINO and GLoRe, with each point representing an image and colors indicating ground-truth categories. DINO exhibits substantial overlap between classes, with many samples distributed across mixed regions. GLoRe produces more compact clusters with clearer separation, although some local inter-class overlap remains. The visualization provides qualitative evidence of greater within-class compactness and reduced inter-class mixing in the projected representation space.

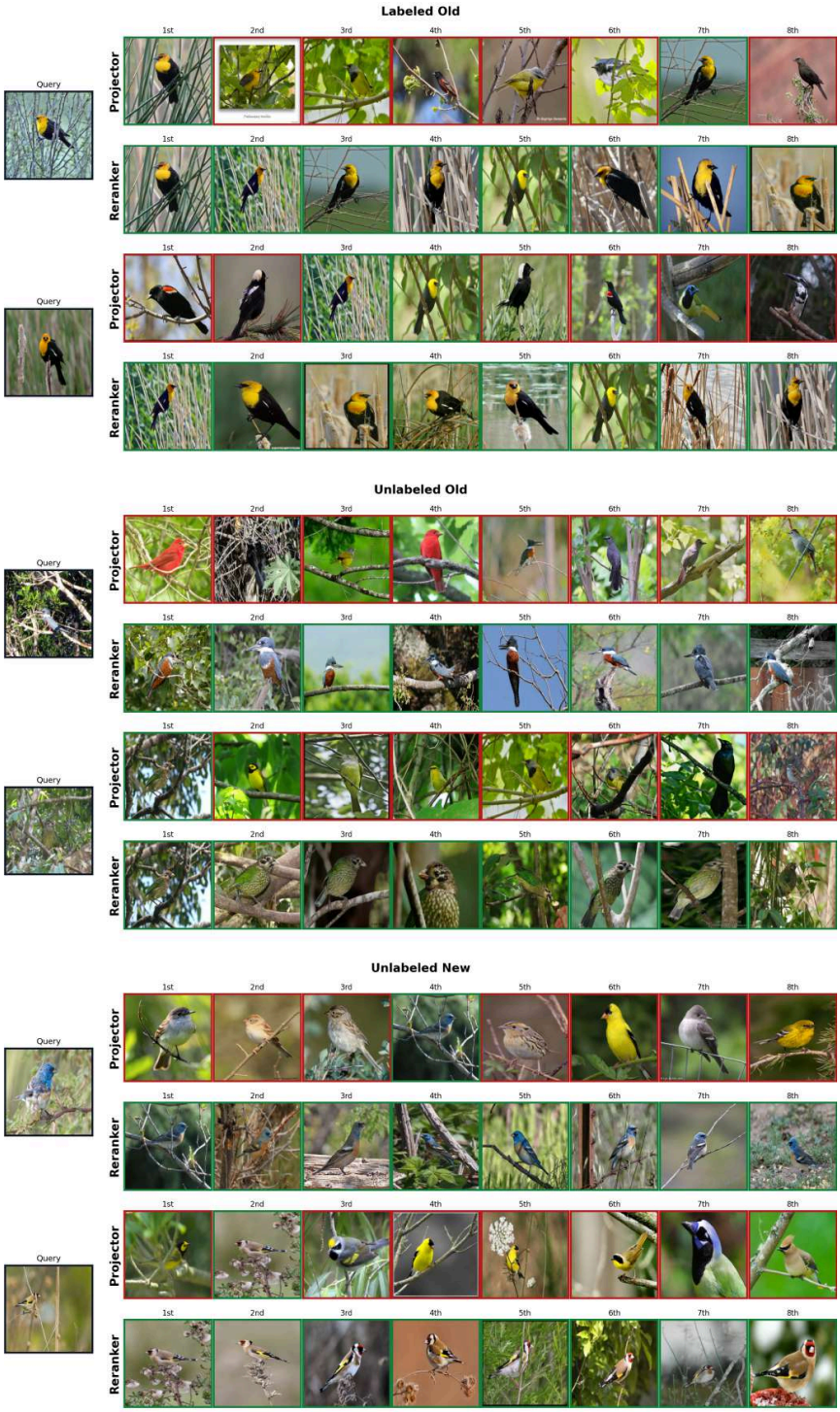


Figure 3: Qualitative reranking results on CUB-200-2011 for labeled-known, unlabeled-known, and unlabeled-novel queries. Green borders indicate correctly retrieved samples from the same class as the query, while red borders denote incorrect retrievals.

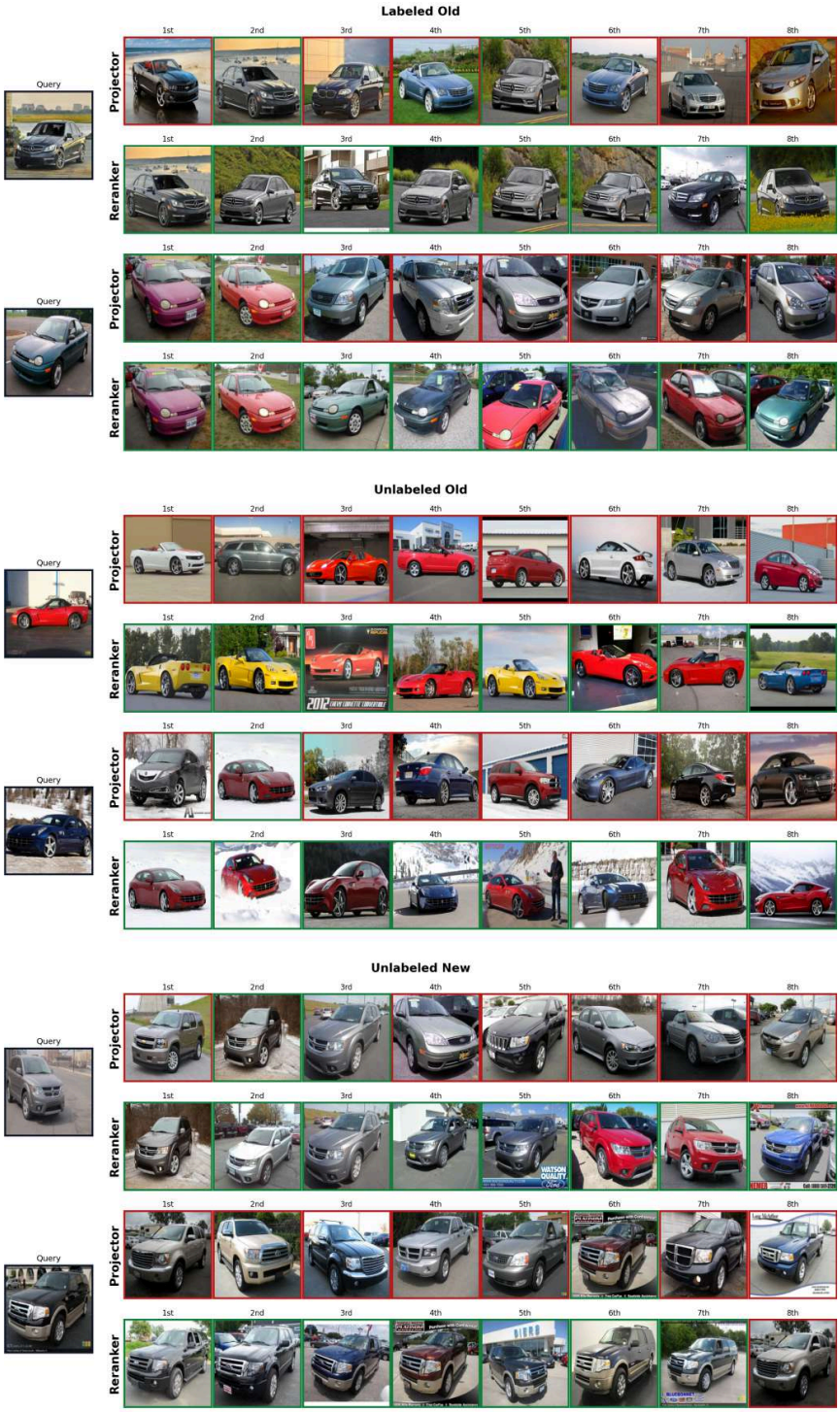


Figure 4: Qualitative reranking results on Stanford Cars for labeled-known, unlabeled-known, and unlabeled-novel queries. Green borders indicate correctly retrieved samples from the same class as the query, while red borders denote incorrect retrievals.

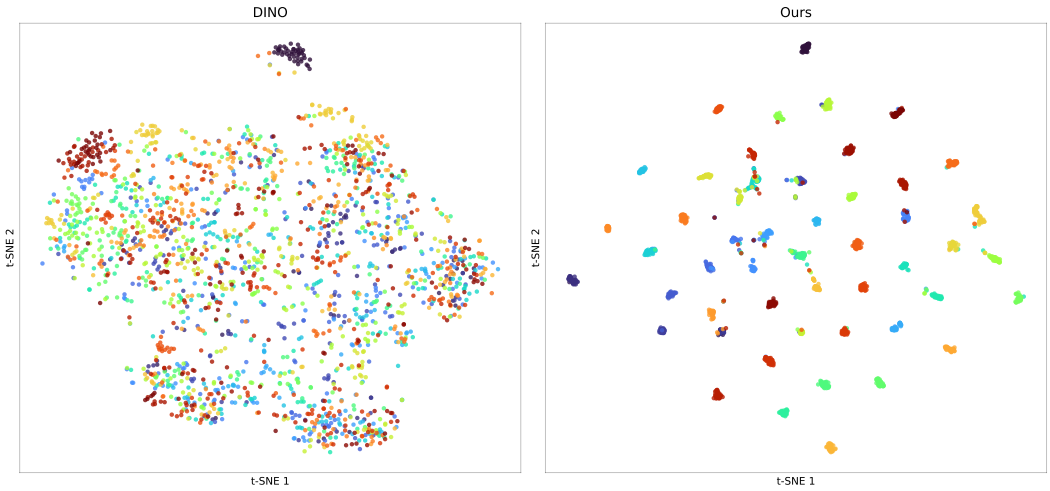


Figure 5: t-SNE visualization of the learned representation space for 50 random classes of the Standford Cars dataset. Each color denotes a different ground-truth category. The visualization compares the structure of the feature space produced by DINO, and the proposed GLoRe framework.