CloSeR: Unified Relational Distillation from Closed-Set Teachers for Category Discovery

Yuanpei Liu, Zhenqi He, Jialu Tang, and Kai Han†

Visual AI Lab, The University of Hong Kong, Hong Kong SAR
ypliu0@connect.hku.hk, kaihanx@hku.hk

Abstract. Generalized Category Discovery (GCD) is an intriguing open-world problem that has garnered increasing attention: given partially labelled data, the goal is to correctly recognize known classes while discovering coherent novel categories from unlabelled samples. Recent GCD methods typically adapt foundation models by jointly optimizing supervised classification and unsupervised discovery objectives on mixed labelled–unlabelled data. While effective, this coupled training can entangle closed-set recognition and open-set discovery, leading to objective conflict and biased predictions, and may disturb the semantic geometry of pretrained representations under limited labels and noisy pseudo-labels. We propose **CloSeR**, a simple plug-and-play framework that injects **Clo**sed-**Se**t **R**elational knowledge into GCD training. CloSeR first builds a domain-adapted closed-set teacher by tuning lightweight block-wise adapters on labelled known-class data while keeping the foundation model backbone frozen, thereby preserving pretrained priors at low training cost. It then transfers the teacher's knowledge to downstream GCD via Unified Relational Distillation (URD), which distills complementary global sample-to-prototype relations to anchor known-class semantics and local sample-to-sample relations to preserve neighborhood structure, using separate feature pathways to reduce optimization interference. CloSeR is head-agnostic and readily integrates with both parametric and non-parametric GCD methods. Extensive experiments with DINO and DINOv2 backbones on six benchmarks (CIFAR-10/100, ImageNet-100, CUB, Stanford-Cars, and FGVC-Aircraft) show consistent gains over GCD baselines, achieving state-of-the-art performance. Project page: https://visual-ai.github.io/closer/

Keywords: Generalized Category Discovery · Self-Supervised Learning · Transfer Learning · Knowledge Distillation

1 Introduction

Once framed as Novel Category Discovery (NCD) [19] and later broadened to Generalized Category Discovery (GCD) [55], *category discovery* [22] has become a focal open-world challenge. In GCD, we are given a partially-labelled dataset:

† Corresponding author.

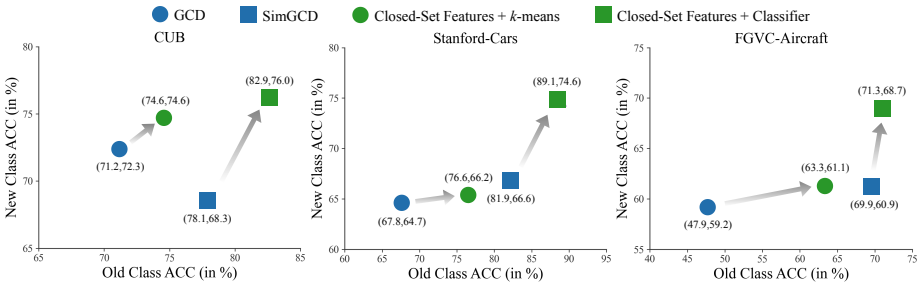

Fig. 1: The *Old*-class *vs. New*-class *ACC* on CUB [58], Stanford-Cars [28], and FGVC-Aircraft [35] using DINOv2 [38] backbone. We compare two representative GCD baselines (the *non-parametric* GCD [55] and the *parametric* SimGCD [63]) with two simple alternatives that first adapt the backbone on labelled known classes (closed-set features) and then perform GCD via either k-means clustering or learning a parametric classifier. *Note:* "GCD" here denotes the method proposed in [55], not the task name.

the unlabelled split contains samples from both known and unknown categories. The goal is to transfer knowledge from the labelled subset so that samples from known classes are correctly recognized while samples from unknown classes are grouped into their underlying novel categories.

Existing GCD methods can be broadly categorized by how they produce category predictions. *Non-parametric* approaches [20, 44, 45, 55] discover categories by clustering features, whereas *parametric* methods [22, 32, 59, 63] learn an explicit classifier (often prototype-based) to assign category indices. With the rise of foundation models [4, 38, 43], recent works from both families typically adapt a pretrained backbone (via full fine-tuning or parameter-efficient tuning) and optimize supervised and unsupervised objectives jointly on the mixed labelled–unlabelled data. While effective, this *one-stage* training entangles known-class classification and novel-class discovery, which may introduce optimization conflict and prediction bias [16,32]. At the same time, prior studies [33,55,59] emphasize that GCD is essentially a *transfer clustering* problem, where learning strong categorical priors from known classes is crucial. Moreover, modern foundation models are pretrained on large-scale data and exhibit strong transferability and semantic structure [4, 38]. In principle, such pretrained priors should be highly beneficial for *transfer clustering* in GCD; however, naïvely adapting these models with mixed supervised–unsupervised objectives can distort the pretrained geometry, especially under limited labelled data and noisy pseudo-labels. This suggests that *how* we adapt foundation models may be as important as *whether* we adapt them. Taken together, these considerations raise a natural question: *Can we explicitly learn to recognize known classes first, and then leverage this closed-set knowledge to guide category discovery?*

To answer this, we conduct a simple pilot study. We *fully fine-tune* a foundation model (*e.g.*, DINOv2 [38]) on labelled known classes using a standard cross-entropy objective with a *small learning rate* to preserve the pretrained prior and

mitigate overfitting, and then evaluate the resulting representations using either vanilla k-means clustering or a lightweight parametric classifier trained on top of the frozen backbone following the standard objective [63]. As shown in Fig. 1, both *non-parametric* and *parametric* solutions improve markedly compared to their standard GCD counterparts. This suggests that *explicitly consolidating categorical priors from known classes* can significantly enhance the representation quality for category discovery. However, this naïve two-step strategy is limited in practice: **(i)** fully fine-tuning foundation models is expensive, especially for larger variants (*e.g.*, ViT-L), which limits the scaling up of GCD models, and **(ii)** simply freezing the adapted backbone and training a separate GCD head is inflexible and does not readily extend to diverse GCD pipelines, particularly non-parametric methods that rely on end-to-end feature learning.

We therefore propose **CloSeR**, a simple yet effective framework that injects **Clo**sed-**Se**t **R**elational knowledge into GCD training in a plug-and-play manner. CloSeR consists of two stages. **First**, we perform Closed-Set Transfer Learning (CST) by inserting *lightweight block-wise adapters* [8] into a *frozen* foundation model and learning closed-set prototypes using *labelled data only*. By updating only a small set of adapter parameters, CST provides an *efficient* way to adapt shallow-to-deep features, yielding a domain-adapted closed-set teacher that preserves the strong pretrained prior while bridging the target-domain gap at low training cost. **Second**, during GCD training, we introduce Unified Relational Distillation (URD), which transfers **(i)** *global* sample-to-prototype relations to anchor semantics to known-class prototypes and **(ii)** *local* sample-to-sample relations to preserve neighborhood geometry. To reduce optimization interference, URD decouples the feature pathways used for global and local relations. Importantly, CloSeR can regularize both *parametric* and *non-parametric* GCD methods: it provides relational supervision to stabilize representation learning, regardless of whether category indices are produced by a classifier head or by clustering.

We evaluate CloSeR with different pretrained backbones (DINO and DINOv2) on six benchmarks, including CIFAR-10 [29], CIFAR-100 [29], ImageNet-100 [11], CUB [58], Stanford-Cars [28], and FGVC-Aircraft [35]. Across GCD baselines spanning both *parametric* and *non-parametric* methods, CloSeR yields consistent improvements. In summary, we make the following contributions: **(i)** We identify a key limitation of the *de facto* GCD training pipeline and show that adapting foundation models to known classes before discovery is crucial for robust category discovery. **(ii)** We propose CloSeR, a staged framework that builds a domain-adapted closed-set teacher via efficient block-wise adapter tuning and transfers its knowledge to downstream GCD training. **(iii)** Within CloSeR, we introduce Unified Relational Distillation (URD), a head-agnostic relational regularizer that distills complementary global and local relations through separate feature pathways. **(iv)** We validate CloSeR through extensive experiments across multiple datasets, backbones, and heterogeneous GCD baselines, showing consistent gains and strong generalization.

2 Related Work

Category Discovery. Category discovery [22], first formulated as Novel Category Discovery (NCD) in a transfer-clustering setting [19], seeks to transfer knowledge from labelled (seen) categories to cluster unlabelled data from novel classes. Generalized Category Discovery (GCD) [55] relaxes the NCD assumption by allowing the unlabelled pool to contain samples from both known and unknown classes, prompting extensive follow-up work that explores diverse strategies for this challenging setting [2, 3, 5, 6, 13, 17, 18, 20, 23, 26, 27, 32, 33, 42, 48, 50, 51, 60, 64, 72, 74]. In the common GCD setup, a self-supervised pretrained image encoder is used as the backbone [4, 38]; based on how category indices are predicted, methods are typically grouped into *parametric* and *non-parametric*. On the *parametric* side, SimGCD [63] first proposes to learn a parametric classifier regularized by mean-entropy minimization; SPTNet [59] extends SimGCD by introducing spatial prompt tuning to better focus on discriminative object parts and strengthen knowledge transfer; DebGCD [32] augments SimGCD [63] with a hard-label debiased auxiliary classifier and a semantic-distribution detector to identify shifts between known and unknown classes; AF [65] mitigates distracted attention in GCD by measuring token importance and adaptively pruning background/non-informative tokens; APL [10] replaces global features with adaptive part-based representations discovered via learnable queries to balance discriminability and generalization; MOS [41] treats scene context as a prior in GCD and models object–scene associations to resolve ambiguity between base/novel classes; and SEAL [23] proposes a semantic-aware hierarchical framework that propagates coarse-to-fine supervision from labelled to unlabelled data via hierarchical contrastive learning and a cross-granularity consistency module. On the *non-parametric* side, SelEx [45] employs hierarchical semi-supervised k-means and achieves strong results on fine-grained datasets, while HypCD [33] operates in hyperbolic space to capture hierarchical class relations and offers a general GCD framework in that geometry. More recently, ConGCD [49] tackles GCD by learning primitive-oriented representations via semantic reconstruction and combining dominant/contextual consensus units with a scheduler to integrate complementary cues for recognizing both known and novel categories.

Knowledge Distillation. Knowledge Distillation (KD) [14] transfers knowledge from a high-capacity teacher to a compact student and has been widely applied across computer vision, including image classification [12, 37, 62], object detection [7, 9, 34, 47, 61, 66], and semantic segmentation [21, 31]. The classical formulation [24] matches softened logits to convey "dark knowledge" beyond hard labels. Extensions exploit intermediate representations: FitNets [46] use intermediate "hints", and attention transfer [68] guides students via spatial attention maps. Similarity- and relation-based KD align embedding structures rather than only logits, including similarity-preserving KD [54], correlation congruence [40], relational KD [39], and contrastive KD [53]. Apart from teacher–student pairs of different networks, self-distillation replaces the external teacher with a model's own multi-branch or temporal outputs [36, 67], while momentum teachers stabilize targets in semi/self-supervised regimes [1, 4, 15, 52, 75]. Teacher–student asym-

metry is central to self-supervised learning, underpinning BYOL [15], DINO [4], iBOT [75], and masked hybrids [1]. Within GCD, self-distillation techniques [1,4] have become a common foundation.

In this work, we propose to inject the categorical prior from the closed-set models into the GCD models by distilling the relations. Unlike previous relational KD [39], we consider both *global* and *local* relations in two *decoupled* feature spaces. Our framework unifies distillation to heterogeneous student models by transferring teacher-induced structure, allowing both *parametric* and *non-parametric* GCD methods with a single training objective.

3 Preliminaries

Problem Setup and Notation. GCD [55] targets a model that jointly handles two goals: classify unlabelled samples belonging to known classes and partition the remaining unlabelled samples into clusters of unknown classes. Let $\mathbf{D}_u = \{(\mathbf{x}_i^u, y_i^u)\} \subset \mathbf{X} \times \mathbf{Y}_u$ be the unlabelled set and $\mathbf{D}_l = \{(\mathbf{x}_i^l, y_i^l)\} \subset \mathbf{X} \times \mathbf{Y}_l$ be the labelled set, with $\mathbf{Y}_u$ and $\mathbf{Y}_l$ denoting their label spaces. By construction, the unlabelled pool spans both known and novel categories, and in particular $\mathbf{Y}_l \subset \mathbf{Y}_u$. We write $M = |\mathbf{Y}_l|$ for the number of labelled classes. Consistent with prior work [18,57,63], we assume the total number of categories $K = |\mathbf{Y}_l \cup \mathbf{Y}_u|$ is given; when K is unknown, it can be estimated using existing techniques [19,55].
Baselines. Vaze *et al.* [55] formalize the task and introduce an early *non-parametric* baseline. The method fine-tunes a self-supervised foundation model [4] to improve representation quality by jointly optimizing a supervised contrastive objective on labelled data and a self-supervised contrastive objective on all data. Concretely, given two random augmentations $\mathbf{x}_i$ and $\mathbf{x}_i'$ of the same image within a mini-batch B, the self-supervised contrastive loss is:

$$\mathcal{L}_{\text{rep}}^u = \frac{1}{|B|} \sum_{i \in B} -\log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_i'/\tau_r)}{\sum_{j \neq i} \exp(\mathbf{z}_i \cdot \mathbf{z}_j'/\tau_r)}, \tag{1}$$

where $\mathbf{z}_i = \text{L2Norm}(\mathcal{M}_{\text{GCD}}(f_\theta(\mathbf{x}_i)))$ denotes the ℓ_2-normalized projected feature, $\mathbf{z}_i'$ is the feature from the other view $\mathbf{x}_i'$, f_θ is the backbone, $\mathcal{M}_{\text{GCD}}$ is the projection head, and τ_r is the temperature. The supervised contrastive loss is defined as:

$$\mathcal{L}_{\text{rep}}^s = \frac{1}{|B_l|} \sum_{i \in B_l} \frac{1}{|N_i|} \sum_{q \in N_i} -\log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_q/\tau_r)}{\sum_{j \neq i} \exp(\mathbf{z}_i \cdot \mathbf{z}_j/\tau_r)}, \tag{2}$$

where N_i indexes the samples in the labelled mini-batch $B_l \subset B$ that share the same label as $\mathbf{x}_i$. The overall representation objective is $\mathcal{L}_{\text{rep}} = (1-\lambda)\mathcal{L}_{\text{rep}}^u + \lambda\mathcal{L}_{\text{rep}}^s$, where λ is the balance factor. Building on this line of work, Rastegar *et al.* [45] propose SelEx, a stronger hierarchical *non-parametric* method for fine-grained GCD, which has become a widely used baseline in the field.

Wen *et al.* [63] propose SimGCD, the first *parametric* GCD baseline that has been widely adopted in subsequent work [57,59]. It learns a parametric classifier via a self-distillation framework [4], initialized with K normalized prototypes

$\mathbf{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$. For an augmented view $\mathbf{x}_i$ with normalized hidden feature $\mathbf{h}_i = f_\theta(\mathbf{x}_i)/||f_\theta(\mathbf{x}_i)||_2$, the predicted probability of class k is:

$$\mathbf{p}_i^{(k)} = \frac{\exp(\mathbf{h}_i \cdot \mathbf{c}_k/\tau_s)}{\sum_{j=1}^{K} \exp(\mathbf{h}_i \cdot \mathbf{c}_j/\tau_s)}, \tag{3}$$

where τ_s is the student temperature. A soft target $\mathbf{q}_i$ is obtained from the teacher prediction computed on another augmented view using a sharper temperature τ_t. The self-distillation objective is the cross-entropy between the two views, $\ell_{\text{ce}}(\mathbf{q}_i', \mathbf{p}_i) = -\sum_{j=1}^{K} \mathbf{q}_i'^{(j)} \log \mathbf{p}_i^{(j)}$. Accordingly, the unsupervised loss over the mini-batch B is:

$$\mathcal{L}_{\text{cls}}^u = \frac{1}{|B|} \sum_{i \in B} \ell_{\text{ce}}(\mathbf{q}_i', \mathbf{p}_i) - \xi \mathcal{H}(\overline{\mathbf{p}}), \tag{4}$$

where $\overline{\mathbf{p}} = \frac{1}{2|B|} \sum_{i \in B} (\mathbf{p}_i + \mathbf{p}_i')$ denotes the batch-mean prediction and $\mathcal{H}(\overline{\mathbf{p}}) = -\sum_{j=1}^{K} \overline{\mathbf{p}}^{(j)} \log \overline{\mathbf{p}}^{(j)}$ is its entropy regularizer, weighted by ξ. For labelled samples, the supervised classification loss is $\mathcal{L}_{\text{cls}}^s = \frac{1}{|B_l|} \sum_{i \in B_l} \ell_{\text{ce}}(\mathbf{p}_i, \mathbf{y}_i)$, where $\mathbf{y}_i$ is the one-hot encoding of y_i. The overall classification objective is $\mathcal{L}_{\text{cls}} = (1-\lambda)\mathcal{L}_{\text{cls}}^u + \lambda \mathcal{L}_{\text{cls}}^s$, with λ being the balance factor. Together with the representation objective $\mathcal{L}_{\text{rep}}$, the full training loss is $\mathcal{L}_{\text{GCD}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{rep}}$.

Limitation of Baselines. All the aforementioned baselines [45, 55, 63] follow a *one-stage* paradigm, directly fine-tuning the backbone on mixed labelled–unlabelled data by jointly optimizing supervised and unsupervised objectives (*e.g.*, $\mathcal{L}_{\text{rep}}^s$ with $\mathcal{L}_{\text{rep}}^u$, and $\mathcal{L}_{\text{cls}}^s$ with $\mathcal{L}_{\text{cls}}^u$). This coupling entangles known-class recognition with novel-class discovery: unsupervised signals computed over all samples (often from noisy pseudo-labels) can conflict with closed-set discrimination and bias representations toward known categories. Moreover, updating pretrained features under mixed objectives can distort the foundation model geometry—especially with limited labelled data—leading to suboptimal *transfer clustering* for unknown classes.

4 Method

4.1 Overview

We propose **CloSeR**, a concise, staged framework with two components (Fig. 2). **First**, we perform Closed-Set Transfer Learning (CST) by inserting a block-wise adapter [8] into a *frozen* foundation model (*e.g.*, DINO/DINOv2) and learning M closed-set prototypes using labelled data only. **Second**, during GCD training we introduce Unified Relational Distillation (URD), which distills **(i)** global sample-to-prototype relations and **(ii)** local sample-to-sample relations from the closed-set space into the GCD student. We train the student with the standard objective (Sec. 3) together with our URD regularization.

4.2 Closed-Set Transfer Learning

Let ϕ denote a pretrained ViT foundation model (*e.g.*, DINO [4]/DINOv2 [38]) trained on large-scale data. Directly fine-tuning all parameters of ϕ on the limited labelled set $\mathbf{D}_l$ is often inefficient and may overfit, while keeping ϕ fully

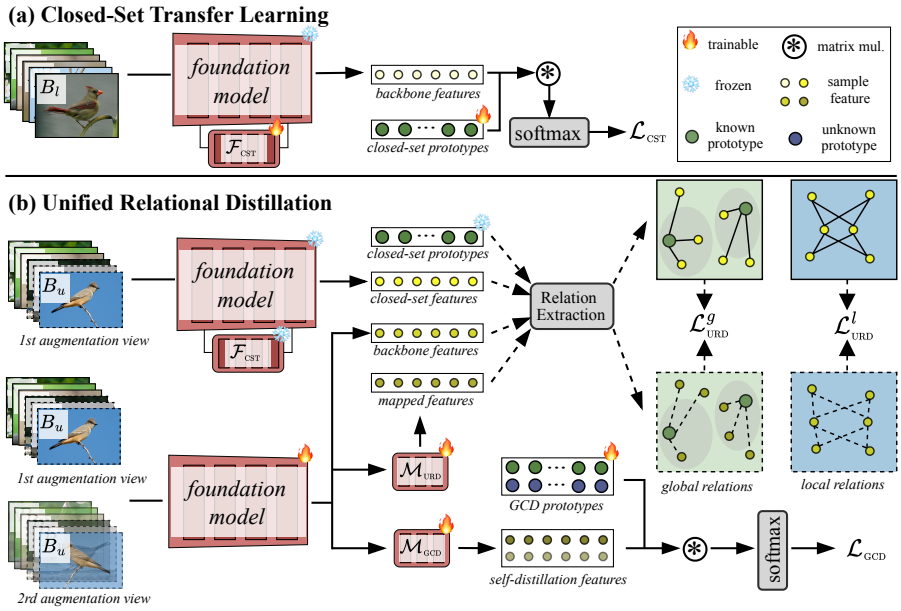

Fig. 2: Overall pipeline of CloSeR (illustrated with SimGCD as an example; it is also applicable to other baselines): **(a)** Closed-Set Transfer Learning; **(b)** Unified Relational Distillation. *Note*: The representation loss in SimGCD is omitted for clarity.

frozen can leave a non-negligible domain gap between the pre-training distribution and the target dataset. To obtain an effective yet stable transfer, we insert a lightweight, *block-wise* adapter [8, 25, 69] into each transformer block. This design enables *feature adaptation throughout the network hierarchy*—from early blocks capturing low-level appearance cues to late blocks encoding high-level semantics—while preserving the strong generalization of the original foundation weights. We denote the collection of all adapter parameters by $\mathcal{F}_{\text{CST}}$ and freeze all original parameters of ϕ.

Given a randomly augmented view $\mathbf{x}_i$ of an image, the closed-set transfer model outputs an ℓ_2-normalized feature:

$$\mathbf{h}_i^{\text{cs}} = \frac{\phi(\mathbf{x}_i; \mathcal{F}_{\text{CST}})}{\|\phi(\mathbf{x}_i; \mathcal{F}_{\text{CST}})\|_2}. \tag{5}$$

We use normalized features for two reasons. First, this matches the cosine-similarity classifier used in both the GCD baseline (Sec. 3) and the foundation model training [4, 38], where logits are computed by inner products between normalized features and normalized prototypes and scaled by a temperature. Second, it stabilizes optimization when transferring from a foundation embedding space to the target domain and makes the learned prototypes comparable across classes.

We maintain M ℓ_2-normalized closed-set prototypes $\mathbf{C}^{\text{cs}} = \{\mathbf{c}_1^{\text{cs}}, \ldots, \mathbf{c}_M^{\text{cs}}\}$, one per known class, acting as a cosine classifier head in the adapted closed-set space. For a labelled sample $(\mathbf{x}_i^l, y_i^l)$, we compute the probability of class k as:

$$\mathbf{p}_i^{\text{cs}(k)} = \frac{\exp(\mathbf{h}_i^{\text{cs}} \cdot \mathbf{c}_k^{\text{cs}} / \tau_{\text{cs}})}{\sum_{j=1}^{M} \exp(\mathbf{h}_i^{\text{cs}} \cdot \mathbf{c}_j^{\text{cs}} / \tau_{\text{cs}})}, \tag{6}$$

where τ_{cs} is consistent with the temperature used in $\mathcal{L}_{\text{cls}}^s$. We train CST with the supervised cross-entropy loss:

$$\mathcal{L}_{\text{CST}} = \frac{1}{|B_l|} \sum_{i \in B_l} \ell_{\text{ce}}(\mathbf{p}_i^{\text{cs}}, \mathbf{y}_i^l), \tag{7}$$

where $\mathbf{y}_i^l$ is the one-hot encoding of y_i^l. The resulting adapted closed-set representation and prototypes are then used to construct the teacher global/local relations distilled by URD (Sec. 4.3).

4.3 Unified Relational Distillation

While the baseline optimizes a K-way prototype classifier with self-distillation techniques (Sec. 3), its discovery of novel categories can be unstable under limited supervision, and the learned embedding may drift away from the strong semantic structure provided by the foundation model. We therefore introduce Unified Relational Distillation (URD), which transfers *relational* knowledge from the closed-set adapted foundation model into the GCD student model. Compared to matching features directly, relation-based distillation [39] is more tolerant to representation mismatch and focuses on preserving the *geometry* of the embedding space that has been proven to be critical for category discovery [23, 33].

At each iteration, we first construct a mixed mini-batch $B = B_l \cup B_u$, where B_l and B_u are sampled from $\mathbf{D}_l$ and $\mathbf{D}_u$, respectively. Following the standard practice, *two* augmented views are used to compute the standard self-distillation losses; however, we empirically find that applying URD to only *one* view per image is sufficient. We denote the selected augmented view by $\mathbf{x}_i$ for $i \in B$ and treat the closed-set adapted model as a teacher. Specifically, we *freeze* the foundation backbone ϕ and the learned adapters $\mathcal{F}_{\text{CST}}$, as well as the closed-set prototypes $\mathbf{C}^{\text{cs}}$, and compute normalized teacher features $\mathbf{h}_i^T = \mathbf{h}_i^{\text{cs}}$. Let f_θ denote the student backbone, initialized from the same foundation model. During GCD training, we fine-tune only the last few transformer blocks (and keep earlier blocks frozen) together with the GCD heads/prototypes. For relational distillation, we extract normalized student features:

$$\mathbf{h}_i^S = \frac{f_\theta(\mathbf{x}_i)}{\|f_\theta(\mathbf{x}_i)\|_2}, \quad i \in B. \tag{8}$$

To compare sample-to-prototype relations in the closed-set space, we further introduce a lightweight mapping network $\mathcal{M}_{\text{URD}}$ and define mapped features:

$$\tilde{\mathbf{h}}_i^S = \frac{\mathcal{M}_{\text{URD}}(\mathbf{h}_i^S)}{\|\mathcal{M}_{\text{URD}}(\mathbf{h}_i^S)\|_2}. \tag{9}$$

This mapping compensates for the representational gap between the GCD student features and the closed-set adapted teacher space, enabling effective relation transfer without constraining the student to mimic teacher features directly.

After obtaining features from the teacher and student models, we define *global relations* as the similarity-induced distribution over the M closed-set prototypes for each sample. Using the fixed closed-set prototypes $\mathbf{C}^{\text{cs}}$, the element in the global relation matrix $\mathbf{R}_g^T$ of the teacher model is:

$$\mathbf{R}_g^T(i,k) = \frac{\exp(\mathbf{h}_i^T \cdot \mathbf{c}_k^{\text{cs}}/\tau_g)}{\sum_{j=1}^{M} \exp(\mathbf{h}_i^T \cdot \mathbf{c}_j^{\text{cs}}/\tau_g)}, \tag{10}$$

and the student counterpart is computed with mapped features:

$$\mathbf{R}_g^S(i,k) = \frac{\exp(\tilde{\mathbf{h}}_i^S \cdot \mathbf{c}_k^{\text{cs}}/\tau_g)}{\sum_{j=1}^{M} \exp(\tilde{\mathbf{h}}_i^S \cdot \mathbf{c}_j^{\text{cs}}/\tau_g)}. \tag{11}$$

After obtaining the relation matrices, we align the student global relations to the teacher via KL divergence with soft targets:

$$\mathcal{L}_{\text{URD}}^g = \frac{1}{|B|} \sum_{i \in B} \text{KL}\big(\mathbf{R}_g^T(i,\cdot) \,\|\, \mathbf{R}_g^S(i,\cdot)\big)\,. \tag{12}$$

Intuitively, this loss encourages the student to preserve the teacher's relative similarity profile to closed-set anchors, which stabilizes the semantic organization of the embedding while still allowing the student to allocate capacity to novel categories.

Global relations anchor each sample to closed-set prototypes; in addition, we distill *local relations* that describe neighborhood structure within the mixed batch. For $i \neq j$, we compute temperature-scaled cosine similarities:

$$s_{ij}^T = \frac{\mathbf{h}_i^T \cdot \mathbf{h}_j^T}{\tau_\ell}, \qquad s_{ij}^S = \frac{\mathbf{h}_i^S \cdot \mathbf{h}_j^S}{\tau_\ell}. \tag{13}$$

We then convert similarities into row-stochastic distributions, and the elements of the local relation matrices $\mathbf{R}_\ell^T / \mathbf{R}_\ell^S$ are defined as:

$$\mathbf{R}_\ell^T(i,j) = \frac{\exp(s_{ij}^T)}{\sum_{t \in B,\, t \neq i} \exp(s_{it}^T)}, \qquad \mathbf{R}_\ell^S(i,j) = \frac{\exp(s_{ij}^S)}{\sum_{t \in B,\, t \neq i} \exp(s_{it}^S)}. \tag{14}$$

These distributions capture, for each anchor sample, which other samples in the batch are most semantically related according to the teacher/student. The neighborhood structures within the local relations of the two models are then matched via KL divergence:

$$\mathcal{L}_{\text{URD}}^l = \frac{1}{|B|} \sum_{i \in B} \text{KL}\big(\mathbf{R}_\ell^T(i,\cdot) \,\|\, \mathbf{R}_\ell^S(i,\cdot)\big)\,. \tag{15}$$

This term encourages the student to preserve fine-grained relational geometry (*e.g.*, intra-class compactness and inter-class separation) induced by the closed-set adapted teacher, which we find beneficial for both known-class recognition and novel-class clustering. Finally, we optimize the standard objective together with URD:

$$\mathcal{L}_{\text{CloSeR}} = \mathcal{L}_{\text{GCD}} + \alpha\, \mathcal{L}_{\text{URD}}^g + \beta\, \mathcal{L}_{\text{URD}}^l, \tag{16}$$

where α and β balance global and local relational transfer.

Table 1: Results on generic datasets (CIFAR-10/100 [29] and ImageNet-100 [11]), reported for *All*/*Old*/*New* categories. Right: average over datasets.

	Method[Venue]	CIFAR-10 [29]			CIFAR-100 [29]			ImageNet-100 [11]			Average		
		All	Old	New	All	Old	New	All	Old	New	All	Old	New
DINO	GCD [55][CVPR 2022]	91.5	97.9	88.2	73.0	76.2	66.5	74.1	89.8	66.3	79.5	88.0	73.7
	PromptCAL [70][CVPR 2023]	**97.9**	96.6	98.5	81.2	84.2	75.3	83.1	92.7	78.3	87.4	91.2	84.0
	GPC [73][ICCV 2023]	90.6	97.6	87.0	75.4	84.6	60.1	75.3	93.4	66.7	80.4	91.9	71.3
	InfoSieve [44][NeurIPS 2023]	94.8	97.7	93.4	78.3	82.2	70.5	80.5	93.8	73.8	84.5	91.2	79.2
	SPTNet [59][ICLR 2024]	97.3	95.0	98.6	81.3	84.3	75.6	85.4	93.2	81.4	88.0	90.8	85.2
	DebGCD [32][ICLR 2025]	97.2	94.8	98.4	83.0	84.6	79.9	85.9	94.3	81.6	88.7	91.2	86.6
	HypCD [33][CVPR 2025]	96.7	97.6	96.3	82.4	85.1	77.0	**86.8**	**94.6**	**82.8**	88.6	92.4	85.4
	AF [65][ICCV 2025]	97.8	95.9	**98.8**	82.2	85.0	76.5	85.4	**94.6**	80.8	88.5	91.8	85.4
	SEAL [23][NeurIPS 2025]	97.2	94.7	98.4	82.1	81.7	**83.0**	84.6	90.9	81.3	88.0	89.1	**87.6**
	SimGCD [63][ICCV 2023]	97.1	95.1	98.1	80.1	81.2	77.8	83.0	93.1	77.9	86.7	89.8	84.6
	+ **CloSeR**	97.6	**99.3**	96.7	**83.8**	85.3	80.8	84.9	92.6	81.0	88.8	92.4	86.2
	SelEx [45][ECCV 2024]	95.9	98.1	94.8	82.3	85.3	76.3	83.1	93.6	77.8	87.1	92.3	83.0
	+ **CloSeR**	97.7	98.3	97.3	83.7	**88.4**	74.3	85.6	**94.6**	81.1	**89.0**	**93.8**	84.2
DINOv2	GCD [55][CVPR 2022]	97.8	**99.0**	97.1	79.6	84.5	69.9	78.5	89.5	73.0	85.3	91.0	80.0
	DebGCD [32][ICLR 2025]	**98.9**	97.5	99.6	90.1	90.9	88.6	**93.2**	**97.0**	**91.2**	94.1	95.1	93.1
	†HypCD [33][CVPR 2025]	98.6	98.1	98.9	88.6	91.5	82.8	92.3	96.4	90.2	93.2	95.3	90.6
	SEAL [23][NeurIPS 2025]	**98.9**	98.1	99.3	89.8	90.4	89.5	91.3	93.3	90.3	93.3	93.9	93.0
	SimGCD [63][ICCV 2023]	98.7	96.7	99.7	88.5	89.2	87.2	89.9	95.5	87.1	92.4	93.8	91.3
	+ **CloSeR**	**98.9**	97.1	**99.8**	**93.2**	**93.0**	**93.5**	92.5	96.6	90.4	**94.9**	95.6	**94.6**
	SelEx [45][ECCV 2024]	98.5*	98.8*	98.5*	87.7*	90.8*	81.5*	90.9*	96.2*	88.3*	92.4	95.3	89.4
	+ **CloSeR**	98.7	98.9	98.6	89.7	92.7	83.7	92.7	95.7	91.1	93.7	**95.8**	91.1

†results based on the best baseline [45]. *results from [33].

5 Experiments

5.1 Implementation Details

We benchmark CloSeR against two representative baselines: the *non-parametric* SelEx [45] and the *parametric* SimGCD [63]. Performance is evaluated using clustering accuracy (*ACC*) [55]. To study the effect of different foundation models, we instantiate all methods with pretrained weights from DINO [4] and DINOv2 [38]. Unless stated otherwise, we use a ViT-B backbone for both teacher and student to ensure fair comparison across methods (ViT-B/14 for DINOv2 and ViT-B/16 for DINO). In the first CST stage, we train the adapter for 100 epochs on fine-grained datasets and 30 epochs on generic datasets with $\tau_{\text{cs}} = 0.1$, using an initial learning rate of 10^{-1} that is cosine-annealed to 10^{-4}. In the subsequent URD stage, for $\mathcal{L}_{\text{GCD}}$ we follow the hyperparameters and augmentation settings of the corresponding baseline. For $\mathcal{L}_{\text{URD}}$, we set the temperatures τ_g and τ_ℓ to 0.3, and the loss weights α and β to 5.0 and 20.0, respectively. Additional details on datasets, optimization, and hyperparameters are provided in the supplementary material.

5.2 Quantitative Comparison

Tab. 1 and Tab. 2 report results on three generic (CIFAR-10/100, ImageNet-100) datasets and three fine-grained (CUB, Stanford-Cars, FGVC-Aircraft) with

Table 2: Results on fine-grained datasets (CUB [58], Stanford-Cars [28], FGVC-Aircraft [35]), reported for *All*/*Old*/*New* categories. Right: average over datasets.

		CUB [58]			Stanford-Cars [28]			FGVC-Aircraft [35]			Average		
	Method[Venue]	All	Old	New	All	Old	New	All	Old	New	All	Old	New
DINO	GCD [55][CVPR 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 [70][CVPR 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
	GPC [73][ICCV 2023]	52.0	55.5	47.5	38.2	58.9	27.4	43.3	40.7	44.8	44.5	51.7	39.9
	μGCD [57][NeurIPS 2023]	65.7	68.0	64.6	56.5	68.1	50.9	53.8	55.4	53.0	58.7	63.8	56.2
	InfoSieve [44][NeurIPS 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 [59][ICLR 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
	FlipClass [30][NeurIPS 2024]	71.3	71.3	71.3	63.1	81.7	53.8	59.3	66.9	55.4	64.6	73.3	60.2
	DebGCD [32][ICLR 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 [41][CVPR 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
	APL [10][CVPR 2025]	68.5	73.1	66.2	62.3	80.7	53.4	60.9	63.5	59.6	63.9	72.4	59.7
	HypCD [33][CVPR 2025]	79.8	75.8	81.8	62.9	80.0	54.7	65.9	67.3	65.1	69.5	74.4	67.2
	AF [65][ICCV 2025]	69.0	74.3	66.3	**67.0**	80.7	**60.4**	59.4	68.1	55.0	65.1	74.4	60.6
	ConGCD [49][ICCV 2025]	81.7	**80.4**	82.4	57.5	77.5	47.9	62.5	**70.2**	58.7	67.2	76.0	63.0
	SEAL [23][NeurIPS 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
	SimGCD [63][ICCV 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
	+ **CloSeR**	68.2	76.9	63.9	63.0	80.2	54.7	56.6	55.5	57.2	62.6	70.9	58.6
	SelEx [45][ECCV 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
	+ **CloSeR**	**82.8**	80.3	**84.1**	66.5	**83.4**	58.3	**66.6**	68.1	**65.8**	**72.0**	**77.3**	**69.4**
DINOv2	GCD [55][CVPR 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
	DebGCD [32][ICLR 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
	APL [10][CVPR 2025]	75.1	79.1	73.2	73.4	87.6	66.7	68.8	74.1	66.6	72.4	80.3	68.8
	†HypCD [33][CVPR 2025]	**90.7**	85.3	**93.4**	**83.8**	93.3	**79.2**	**83.4**	82.0	**84.1**	**86.0**	86.9	**85.6**
	†ConGCD [49][ICCV 2025]	86.3	**87.4**	85.8	79.8	93.1	73.3	81.7	**83.3**	81.0	82.6	**87.9**	80.0
	SEAL [23][NeurIPS 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
	SimGCD [63][ICCV 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
	+ **CloSeR**	78.1	79.5	77.3	78.8	90.7	73.0	71.8	68.7	73.3	76.2	79.6	74.5
	SelEx [45][ECCV 2024]	87.4	85.1	88.5	82.2	93.7	76.7	79.8	82.3	78.6	83.1	87.0	81.3
	+**CloSeR**	89.8	86.5	91.5	83.1	**94.6**	77.6	82.1	81.2	82.6	85.0	87.4	83.9

†results based on the best baseline [45].

DINO and DINOv2 backbones. For brevity, we sometimes abbreviate Stanford-Cars and FGVC-Aircraft as *Cars* and *Aircraft*, respectively.

Generic Datasets. On generic benchmarks, the baselines are already strong, but CloSeR still provides consistent gains. Under DINO, CloSeR improves the SimGCD baseline from 86.7 to 88.8 average *All* accuracy (+2.1 points) and improves SelEx from 87.1 to 89.0 (+1.9 points). Under DINOv2, SimGCD+CloSeR achieves 94.9 average *All* and 94.6 average *New* accuracy, delivering the *best* reported average *New* accuracy. SelEx+CloSeR also improves the average *All* accuracy (92.4→93.7) with clear gains on CIFAR-100 and ImageNet-100.

Fine-Grained Datasets. On fine-grained datasets, our CloSeR is particularly effective. With DINO, CloSeR boosts SimGCD on CUB and Cars by large margins (*e.g.*, Cars *All*: 53.8→63.0, +9.2 points), and improves its average *All* accuracy from 56.1 to 62.6 (+11.6% relative). For SelEx, CloSeR brings even larger improvements: the average *New* accuracy increases from 58.8 to 69.4 (+18.0% relative) and the average *All* increases from 63.1 to 72.0, yielding the *best* DINO-based average performance. With DINOv2, CloSeR continues to substantially improve SimGCD, raising average *New* from 65.3 to 74.5 (+9.2 points; +14.1%

Table 3: Experimental results using different transfer learning methods with DINO [4] and DINOv2 [38] pretrained weights on SimGCD. Results are reported on SSB [56].

			CUB [58]			Stanford-Cars [28]			FGVC-Aircraft [35]		
Adaptation Method	GCD Method	Pretrained	All	Old	New	All	Old	New	All	Old	New
(1) Full Fine-tuning	SimGCD	DINO	68.1	76.3	**64.0**	60.9	77.5	52.9	54.8	53.2	55.6
(2) Fine-tuning Last Block	SimGCD	DINO	58.5	56.2	59.6	46.7	56.8	40.3	45.8	41.8	47.8
(3) Block-wise Adapter	SimGCD	DINO	**68.2**	**76.9**	63.9	**63.0**	**80.2**	**54.7**	**56.6**	**55.5**	**57.2**
(1) Full Fine-tuning	SimGCD	DINOv2	76.9	77.4	76.6	78.4	91.7	72.0	67.9	62.6	70.5
(2) Fine-tuning Last Block	SimGCD	DINOv2	77.3	80.2	75.8	77.9	90.5	71.8	67.7	64.6	69.2
(3) Block-wise Adapter	SimGCD	DINOv2	**78.1**	**79.5**	**77.3**	**78.8**	**90.7**	**73.0**	**71.8**	**68.7**	**73.3**

Table 4: Experimental results using different distillation components with DINO [4] pretrained weights on SelEx. Results are reported on the SSB [56] benchmark.

			CUB [58]			Stanford-Cars [28]			FGVC-Aircraft [35]		
Global Relation	Local Relation	Feature Decoupling	All	Old	New	All	Old	New	All	Old	New
✗	✗	✗	73.6	75.3	72.8	58.5	75.6	50.3	57.1	64.7	53.3
✓	✗	✗	80.5	78.9	81.2	61.9	81.0	52.6	59.0	65.1	56.0
✗	✓	✗	80.2	78.5	81.1	64.3	82.3	55.7	63.0	66.7	61.1
✓	✓	✗	81.0	78.5	82.2	65.8	82.8	57.6	64.9	**68.4**	63.2
✓	✓	✓	**82.8**	**80.3**	**84.1**	**66.5**	**83.4**	**58.3**	**66.6**	68.1	**65.8**

relative). Meanwhile, SelEx+CloSeR reaches 85.0 average *All* accuracy, which is the *second-best* among DINOv2-based methods, and improves *New* category discovery on all three datasets (*e.g.*, CUB *New*: 88.5→91.5).

Discussion. Across all six datasets and both backbones, CloSeR provides a robust performance boost when applied to heterogeneous GCD pipelines, demonstrating strong generality. The consistent improvements on both *Old* and *New* categories suggest that CloSeR mitigates the common training bias of existing methods and better preserves transferable structure from the foundation model.

5.3 Diagnostic Study

Different Closed-Set Transfer Learning Methods. We compare three transfer strategies in the CST stage while keeping the rest of the CloSeR pipeline fixed: (1) full fine-tuning, (2) fine-tuning only the last transformer block (a standard practice in GCD training), and (3) our block-wise adapter adaptation. For (1) and (2), we use an initial learning rate of 10^{-3} to reduce overfitting risk, whereas for (3), we use 10^{-1} given the much smaller set of trainable parameters. As shown in Tab. 3, full fine-tuning yields reasonable accuracy, but fine-tuning only the last block causes a pronounced performance drop, especially with DINO pretraining (*e.g.*, Cars *All* 60.9 → 46.7; Aircraft *All* 54.8 → 45.8). In contrast, block-wise adapters achieve the best overall results, consistently improving *All* accuracy on DINO (CUB 68.1 → 68.2, Cars 60.9 → 63.0, Aircraft 54.8 → 56.6) while substantially strengthening *New*-class discovery (*e.g.*, Cars *New* 52.9 → 54.7). These results support two key observations. First, it is beneficial to preserve the strong pretrained foundation weights rather than fully updating them on limited labelled data. Second, effective transfer requires adapting not only high-

Table 5: Experimental results using different backbone sizes with DINOv2 [38] on SimGCD and SimGCD+CloSeR. Results are reported on SSB [56].

			CUB [58]			Stanford-Cars [28]			FGVC-Aircraft [35]			CIFAR-10 [29]			CIFAR-100 [29]			ImageNet-100 [11]		
Method	Teacher	Student	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
SimGCD	-	ViT-B	71.5	78.1	68.3	71.5	81.9	66.6	63.9	69.9	60.9	98.7	96.7	99.7	88.5	89.2	87.2	89.9	95.5	87.1
+ CloSeR	ViT-B	ViT-B	78.1	79.5	77.3	78.8	90.7	73.0	71.8	68.7	73.3	98.9	97.1	**99.8**	93.2	93.0	**93.5**	92.5	**96.6**	90.4
+ CloSeR	ViT-L	ViT-B	**81.3**	**81.4**	**81.3**	**79.9**	**92.5**	**73.8**	**75.0**	**77.7**	**73.6**	**99.4**	**98.9**	99.7	**93.8**	**94.2**	93.0	**93.4**	95.8	**92.2**
SimGCD	-	ViT-L	71.0	73.7	69.6	71.3	80.8	66.7	70.6	70.1	70.8	99.5	98.7	99.8	91.8	91.8	91.8	92.4	94.0	91.7
+ CloSeR	ViT-L	ViT-L	**81.4**	**83.2**	**80.5**	**83.8**	**92.9**	**79.4**	**76.2**	**82.4**	**73.0**	**99.5**	**98.7**	**99.9**	**94.5**	**94.9**	**93.6**	**93.7**	**97.2**	**91.9**

level semantics but also *shallow* features across the network hierarchy; this is particularly important for weaker pretraining (*e.g.*, DINO), where tuning only the last block is insufficient to close the domain gap.

Ablations on Knowledge Distillation Components. Tab. 4 ablates the three components in URD with SelEx as baseline: global relation distillation (sample-to-prototype), local relation distillation (sample-to-sample), and feature decoupling. Starting from the baseline without distillation, enabling either global or local relations yields large gains across all datasets, confirming the effectiveness of transferring relational structure from the closed-set teacher. Local relations are particularly important on GCD datasets, improving *All* accuracy from 58.5 → 64.3 on Cars and 57.1 → 63.0 on Aircraft, and outperforming using global relations alone on these datasets. Importantly, global and local relations are complementary: combining them further improves performance (*e.g.*, CUB *All* 80.5/80.2 → 81.0, Cars 61.9/64.3 → 65.8, Aircraft 59.0/63.0 → 64.9). Finally, feature decoupling brings an additional and consistent boost on top of global+local distillation. Concretely, we decouple the features used by the two relation pathways: global relations are distilled in the closed-set prototype space using mapped student features, whereas local relations are distilled directly from the student backbone features. This design avoids forcing a single representation to simultaneously satisfy prototype anchoring and neighborhood matching, reducing optimization interference and improving *All* accuracy from 81.0 → 82.8 on CUB, 65.8 → 66.5 on Cars, and 64.9 → 66.6 on Aircraft.

Scaling Up Teacher and Student Backbones. Tab. 5 studies how performance changes with different teacher/student backbone sizes under DINOv2 pretraining. A key finding is that the *default* GCD training strategy in SimGCD does not reliably benefit from scaling up the backbone on fine-grained datasets: replacing a ViT-B student with a larger ViT-L model yields no improvement on CUB (*All* 71.5 → 71.0) and Cars (71.5 → 71.3). This inconsistency suggests that directly optimizing the GCD objective can partially *override* the strong pretrained priors of foundation models, preventing larger backbones from translating additional capacity into better clustering and recognition. In contrast, CloSeR restores a more predictable scaling behavior. With our closed-set teacher and URD supervision, increasing the student from ViT-B to ViT-L (with a ViT-L teacher) improves results across all three fine-grained datasets, especially on Cars (*All* 79.9 → 83.8), while also improving on CUB (81.3 → 81.4) and Aircraft (75.0 → 76.2). Moreover, CloSeR also benefits from a stronger teacher: using

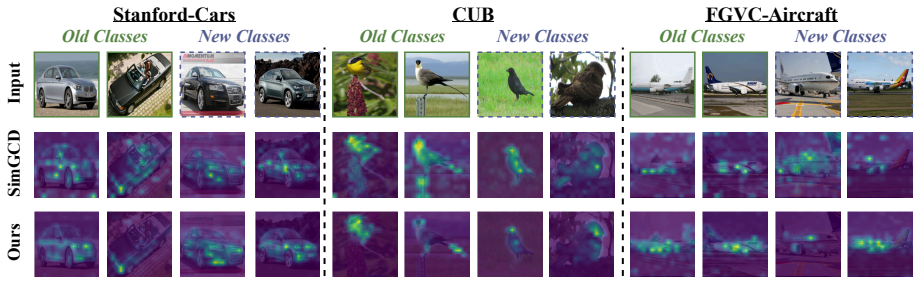

Fig. 3: Comparison of attention maps from the backbone networks of SimGCD [63] and our SimGCD+CloSeR. The results include three datasets: Stanford-Cars [28], CUB [58], and FGVC-Aircraft [35], spanning both *Old* and *New* classes.

a ViT-L teacher improves over a ViT-B teacher with a fixed ViT-B student (*e.g.*, CUB *All* 78.1→81.3). Overall, these results indicate that by constructing a domain-adapted closed-set teacher and distilling both global and local relations, CloSeR better preserves and exploits the foundation model prior, thereby *releasing the power* of larger backbones for category discovery.

5.4 Qualitative Comparison

Attention quality is crucial for fine-grained GCD [55, 59, 71]: models must focus on category-discriminative object parts rather than task-irrelevant background, especially for novel classes where spurious cues can mislead clustering. To examine whether CloSeR improves visual focus, we visualize attention maps of SimGCD [63] and SimGCD+CloSeR in Fig. 3, derived from the last transformer block of the backbone across three fine-grained datasets. Following [4], we average attention over heads and bilinearly upsample the resulting heatmap to the input resolution, then normalize it to $[0, 1]$ for overlay. Qualitatively, our method focuses on semantically diagnostic parts and improves multi-part coverage (*e.g.*, beaks/wing patterns in CUB, grilles/headlights in Stanford-Cars, body/wing markings in FGVC-Aircraft), while the baseline shows diffuse attention with background leakage, particularly on unseen categories. Our method also improves multi-part coverage by highlighting complementary object parts, indicating stronger part-level reasoning and better generalization to novel categories, which supports improved accuracy in the GCD setting.

6 Conclusion

We present CloSeR, a simple yet effective framework to improve generalized category discovery by leveraging foundation models through a *closed-set* transfer learning stage and a unified *relational* distillation stage. First, we build a domain-adapted closed-set teacher via block-wise adapter tuning, which preserves the strong pretrained knowledge while adapting shallow-to-deep features efficiently.

Second, we introduce Unified Relational Distillation (URD) that transfers both prototype-anchored global relations and batch-level local neighborhood structure to the GCD student, with feature decoupling to reduce optimization interference. Extensive experiments on fine-grained and generic benchmarks demonstrate consistent gains over strong baselines, and our ablations show that CloSeR provides a favorable accuracy–efficiency trade-off and better exploits model scaling, effectively unlocking the potential of larger backbones.

Acknowledgements

This work is supported by Hong Kong Research Grants Council – General Research Fund (Grant No. 17211024), Hong Kong Innovation and Technology Commission – Innovation and Technology Fund (Grant No. ITS/488/24FP), and HKU Seed Fund for PI Research.

References

1. Assran, M., Caron, M., Misra, I., Bojanowski, P., Bordes, F., Vincent, P., Joulin, A., Rabbat, M., Ballas, N.: Masked siamese networks for label-efficient learning. In: ECCV (2022)
2. Cao, K., Brbic, M., Leskovec, J.: Open-world semi-supervised learning. In: ICLR (2022)
3. Cao, X., Chen, K., Yang, F., Zheng, X., Tian, Y., Lu, Y.: Allgcd: Leveraging all unlabeled data for generalized category discovery. In: ICCV (2025)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
5. Cendra, F.J., Han, K.: Partco: Part-level correspondence priors enhance category discovery. In: ICML (2026)
6. Cendra, F.J., Zhao, B., Han, K.: Promptccd: Learning gaussian mixture prompt pool for continual category discovery. In: ECCV (2024)
7. Chen, G., Choi, W., Yu, X., Han, T., Chandraker, M.: Learning efficient object detection models with knowledge distillation. In: NeurIPS (2017)
8. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting vision transformers for scalable visual recognition. In: NeurIPS (2022)
9. Cheng, S., Liu, Y., Han, K.: Uadet: A remarkably simple yet effective uncertainty-aware open-set object detection framework. arXiv preprint arXiv:2412.09229 (2024)
10. Dai, Q., Huang, H., Wu, Y., Yang, S.: Adaptive part learning for fine-grained generalized category discovery: A plug-and-play enhancement. In: CVPR (2025)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
12. Duan, S., Sun, Y., Peng, D., Duan, G., Peng, X., Hu, P.: Deep fuzzy multi-view learning for reliable classification. In: ICML (2025)
13. Fini, E., Sangineto, E., Lathuiliere, S., Zhong, Z., Nabi, M., Ricci, E.: A unified objective for novel class discovery. In: ICCV (2021)
14. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. IJCV (2021)

15. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent: a new approach to self-supervised learning. In: NeurIPS (2020)
16. Guo, P., Song, Y., Wang, B.: Towards debiased generalized category discovery. In: IJCAI (2025)
17. Han, K., Rebuffi, S.A., Ehrhardt, S., Vedaldi, A., Zisserman, A.: Automatically discovering and learning new visual categories with ranking statistics. In: ICLR (2020)
18. Han, K., Rebuffi, S.A., Ehrhardt, S., Vedaldi, A., Zisserman, A.: Autonovel: Automatically discovering and learning novel visual categories. IEEE TPAMI (2021)
19. Han, K., Vedaldi, A., Zisserman, A.: Learning to discover novel visual categories via deep transfer clustering. In: ICCV (2019)
20. Hao, S., Han, K., Wong, K.Y.K.: Cipr: An efficient framework with cross-instance positive relations for generalized category discovery. TMLR (2024)
21. He, T., Shen, C., Tian, Z., Gong, D., Sun, C., Yan, Y.: Knowledge adaptation for efficient semantic segmentation. In: CVPR (2019)
22. He, Z., Liu, Y., Han, K.: Category discovery: An open-world perspective. arXiv preprint arXiv:2509.22542 (2025)
23. He, Z., Liu, Y., Han, K.: Seal: Semantic-aware hierarchical learning for generalized category discovery. In: NeurIPS (2025)
24. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NeurIPS 2024 Deep Learning Workshop (2015)
25. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: ECCV (2022)
26. Jia, X., Han, K., Zhu, Y., Green, B.: Joint representation learning and novel category discovery on single-and multi-modal data. In: ICCV (2021)
27. Joseph, K., Paul, S., Aggarwal, G., Biswas, S., Rai, P., Han, K., Balasubramanian, V.N.: Novel class discovery without forgetting. In: ECCV (2022)
28. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: ICCV workshop (2013)
29. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Tech. rep., University of Toronto (2009)
30. Lin, H., An, W., Wang, J., Chen, Y., Tian, F., Wang, M., Wang, Q., Dai, G., Wang, J.: Flipped classroom: Aligning teacher attention with student in generalized category discovery. In: NeurIPS (2024)
31. Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., Wang, J.: Structured knowledge distillation for semantic segmentation. In: CVPR (2019)
32. Liu, Y., Han, K.: Debgcd: Debiased learning with distribution guidance for generalized category discovery. In: ICLR (2025)
33. Liu, Y., He, Z., Han, K.: Hyperbolic category discovery. In: CVPR (2025)
34. Liu, Y., Yin, J., Yu, D., Zhao, S., Shen, J.: Multiple people tracking with articulation detection and stitching strategy. Neurocomputing (2020)
35. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
36. Mobahi, H., Farajtabar, M., Bartlett, P.: Self-distillation amplifies regularization in hilbert space. In: NeurIPS (2020)
37. Müller, R., Kornblith, S., Hinton, G.E.: When does label smoothing help? In: NeurIPS (2019)
38. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

39. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: CVPR (2019)
40. Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., Zhang, Z.: Correlation congruence for knowledge distillation. In: ICCV (2019)
41. Peng, Z., Ma, J., Sun, Z., Yi, R., Song, H., Tan, X., Ma, L.: Mos: Modeling object-scene associations in generalized category discovery. In: CVPR (2025)
42. Pu, N., Zhong, Z., Sebe, N.: Dynamic conceptional contrastive learning for generalized category discovery. In: CVPR (2023)
43. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
44. Rastegar, S., Doughty, H., Snoek, C.: Learn to categorize or categorize to learn? self-coding for generalized category discovery. In: NeurIPS (2023)
45. Rastegar, S., Salehi, M., Asano, Y.M., Doughty, H., Snoek, C.G.M.: Selex: Self-expertise in fine-grained generalized category discovery. In: ECCV (2024)
46. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: ICLR (2015)
47. Shen, J., Liu, Y., Dong, X., Lu, X., Khan, F.S., Hoi, S.: Distilled siamese networks for visual tracking. IEEE TPAMI (2021)
48. Tang, L., Huang, K., Chen, C., Chen, C.: Bures-isotropy alignment: Manifold learning of generalized category discovery. In: ICLR (2026)
49. Tang, L., Huang, K., Chen, C., Yuan, Y., Li, C., Tu, X., Ding, X., Huang, Y.: Dissecting generalized category discovery: Multiplex consensus under self-deconstruction. In: ICCV (2025)
50. Tang, L., Yuan, Y., Chen, C., Zhang, Z., Huang, Y., Zhang, K.: Ocrt: Boosting foundation models in the open world with object-concept-relation triad. In: CVPR (2025)
51. Tang, L., Zheng, J., Huang, K., Chen, C., Huang, Y., Chen, C.: Coge-GCD: Reframing generalized category discovery with compositional generalization. In: ICML (2026)
52. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NeurIPS (2017)
53. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. In: ICLR (2019)
54. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: ICCV (2019)
55. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Generalized category discovery. In: CVPR (2022)
56. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: The semantic shift benchmark. In: ICML workshop (2022)
57. Vaze, S., Vedaldi, A., Zisserman, A.: No representation rules them all in category discovery. In: NeurIPS (2023)
58. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset. Tech. rep., California Institute of Technology (2011)
59. Wang, H., Vaze, S., Han, K.: Sptnet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. In: ICLR (2024)
60. Wang, H., Vaze, S., Han, K.: Hilo: A learning framework for generalized category discovery robust to domain shifts. In: ICLR (2025)
61. Wang, T., Yuan, L., Zhang, X., Feng, J.: Distilling object detectors with fine-grained feature imitation. In: CVPR (2019)

62. Wang, Y., Li, H., Teng, F., Chen, L.: Gorag: Graph-based online retrieval augmented generation for dynamic few-shot social media text classification. In: WWW (2026)
63. Wen, X., Zhao, B., Qi, X.: Parametric classification for generalized category discovery: A baseline study. In: ICCV (2023)
64. Wu, J., Zhang, Y., Fu, R., Liu, Y., Gao, J.: An efficient person clustering algorithm for open checkout-free groceries. In: ECCV (2022)
65. Xu, Q., Hu, Z., Duan, Y., Pei, E., Tai, Y.: A hidden stumbling block in generalized category discovery: Distracted attention. In: ICCV (2025)
66. Yang, X., Yan, J., Feng, Z., He, T.: R3det: Refined single-stage detector with feature refinement for rotating object. In: AAAI (2021)
67. Yun, S., Park, J., Lee, K., Shin, J.: Regularizing class-wise predictions via self-knowledge distillation. In: CVPR (2020)
68. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: ICLR (2017)
69. Zhan, G., Liu, Y., Han, K., Xie, W., Zisserman, A.: Elip: Enhanced visual-language foundation models for image retrieval. In: International Conference on Content-Based Multimedia Indexing (CBMI) (2025)
70. Zhang, S., Khan, S., Shen, Z., Naseer, M., Chen, G., Khan, F.S.: Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In: CVPR (2023)
71. Zhang, W., Zhang, B., Teng, Z., Luo, W., Zou, J., Fan, J.: Less attention is more: Prompt transformer for generalized category discovery. In: CVPR (2025)
72. Zhao, B., Han, K.: Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. In: NeurIPS (2021)
73. Zhao, B., Wen, X., Han, K.: Learning semi-supervised gaussian mixture models for generalized category discovery. In: ICCV (2023)
74. Zhong, Z., Fini, E., Roy, S., Luo, Z., Ricci, E., Sebe, N.: Neighborhood contrastive learning for novel class discovery. In: CVPR (2021)
75. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: Image bert pre-training with online tokenizer. In: ICLR (2020)