GeoGCD: Geometry-Guided Hierarchical Learning for Generalized Category Discovery

Anonymous authors
Paper under double-blind review

Abstract

Generalized Category Discovery (GCD) requires learning a feature space in which both known and previously unseen categories are well organised across multiple semantic granularities. Existing hierarchical approaches rely either on fixed taxonomic priors, which are blind to the actual geometry of the learned features, or on data-driven prototypes, which discard the prior semantic structure; cross-level alignment is typically enforced through KL-based losses that become uninformative or unstable precisely when novel-class predictions are most fragile. We argue that the global geometry of semantics, the shape that taxonomic relations carve in feature and label space, is the natural object to be preserved in this setting, and we introduce a **Geo**metry-Guided Hierarchical Learning framework for **G**eneralized **C**ategory **D**iscovery (**GeoGCD**) that operationalises this principle. GeoGCD models pairwise relations through a hybrid soft-label matrix that fuses taxonomic similarity with a diffusion similarity derived from random walks on the batch graph, and aligns predictions across granularities via a Wasserstein consistency loss for which we establish formal guarantees of finiteness, continuity, and gradient informativeness under disjoint support. On standard fine-grained GCD benchmarks, GeoGCD achieves new state-of-the-art accuracy on CUB-200-2011 and Stanford Cars with **79.93** and **80.10** in overall, respectively, while notably enhancing performance on known classes through better preservation of global semantic geometry with **87.73** and **93.57** in Old accuracy. Code is available at: `https://anonymous.4open.science/r/GeoGCD-7C66`.

1 Introduction

Generalized Category Discovery (GCD) is fundamentally a representation learning problem: given partially labelled data that contains both known and previously unseen classes, the goal is to learn a feature space in which categories at multiple semantic granularities are simultaneously well separated and well organised, even when no supervision is available for the novel ones Vaze et al. (2022a); Wen et al. (2023); He et al. (2025). A representation that supports category discovery in fine-grained domains must therefore satisfy two distinct requirements at once. First, it must respect semantic structure: samples that belong to the same family or order should sit in coherent regions of the feature space, reflecting the prior knowledge encoded in naturally occurring taxonomies He et al. (2025). Second, it must preserve geometric coherence: the global arrangement of clusters, including those formed by novel classes, must reflect the actual density and connectivity of the data manifold, not only the local angular relations between features Nadler et al. (2005); Rastegar et al.; Wang et al. (2023).

Existing hierarchical GCD methods consistently address one side at the expense of the other. A first family relies on fixed taxonomic priors Wang et al. (2023); He et al. (2025) and weights pairwise relations through tree-distance similarity, which gives a clean prior signal but is, by construction, blind to the geometry of the learned features: any cluster structure that the encoder discovers beyond the taxonomy, particularly for novel classes, is invisible to the supervision signal. A second family infers hierarchies directly from data Rastegar et al.; Hao et al. (2023) or organises representations through prototype-level contrast Wen et al. (2023); Wang et al. (2024b), which captures local geometry but discards the explicit semantic prior, leaving the representation free to drift towards arrangements that are inconsistent with the underlying

taxonomy. Cross-granularity consistency, when present at all, is typically enforced through KL-based losses that align coarse predictions with those propagated from finer levels. This choice is comfortable in the standard supervised setting, but it is fragile in category discovery: when novel-class predictions are noisy and the predicted and pseudo-coarse distributions concentrate on disjoint regions of the label simplex, KL is either infinite or carries gradients that ignore the spatial separation of supports, and the cross-level signal silently degrades exactly when it should do the most work. The gap, in our view, is conceptual rather than algorithmic: the global geometry of semantics, the shape that taxonomic relations carve in feature and label space, is rarely treated as a first-class object of learning.

We take this geometry as the guiding principle for the representation. Concretely, we model semantic similarity at every granularity as the fusion of two complementary geometries. A taxonomic similarity, derived from shortest-path distances on the hierarchy tree, encodes prior knowledge that is stable but blind to data. A diffusion similarity, defined through t-step random walks on the batch graph of normalised features and a Gaussian kernel, encodes the data-driven manifold geometry that aggregates information across all paths of length t, robust to noise and to elongated cluster shapes, and intrinsically multi-scale via the spectrum Nadler et al. (2005). The same geometric principle extends naturally to alignment across levels. Since the predicted coarse distribution and the pseudo-coarse distribution induced by the fine-level prediction can have nearly disjoint supports during training, particularly over novel classes, we replace the standard KL term in the cross-granularity consistency by the 1-Wasserstein distance. The distance is finite, continuous, and gradient-rich on the entire label simplex and explicitly encodes the geometric separation between distributions. Fundamentally, KL divergence ignores the underlying metric structure of the label space, whereas the Wasserstein distance explicitly respects it.

We propose GeoGCD (Hierarchical Geometric Generalized Category Discovery), a diffusion-guided semantic contrastive framework that turns the global geometry of semantics into both a representation-learning target and a cross-level alignment criterion. Our contributions are threefold:

- A new formulation of hierarchical category discovery in which the global geometry of semantics, beyond fixed taxonomic priors and local pairwise similarity, drives representation learning.
- A hybrid soft-label matrix combining taxonomic and diffusion-based similarities for hierarchical contrastive learning, together with a 1-Wasserstein cross-granularity consistency loss with formal guarantees of finiteness, continuity, and gradient informativeness under disjoint support.
- Consistent gains on standard fine-grained GCD benchmarks, supported by ablations that isolate the role of each geometric component.

2 Related Work

Generalized Category Discovery. Novel Category Discovery (NCD) Han et al. (2019) introduced the transfer clustering paradigm, in which knowledge from labelled known-class data is used to cluster unlabelled images coming from entirely unseen categories Han et al. (2020; 2021); Jia et al. (2021); Zhong et al. (2021). Generalized Category Discovery (GCD) Vaze et al. (2022a) relaxes the closed-world assumption and allows the unlabelled pool to contain samples from both known and novel classes, which makes the problem closer to a real open-world setting. A number of follow-up works have pushed GCD along different directions Wang et al. (2023); Hao et al. (2023); Pu et al. (2023); Joseph et al. (2022); Cendra et al. (2024); Wang et al. (2024a). SimGCD Wen et al. (2023) proposes a simple end-to-end parametric framework that combines contrastive learning with entropy-regularised classification, so it no longer needs a post-hoc clustering step. SPTNet Wang et al. (2024b) adds spatial prompt tuning to make the backbone focus more on discriminative regions. DebGCD Liu & Han (2025) addresses label bias and semantic shift between known and novel classes through an OOD-based debiased classifier together with a certainty-driven curriculum. PromptCAL Zhang et al. (2023) calibrates contrastive affinities with the help of auxiliary prompts. Despite this progress, most of these methods still operate at a single semantic granularity: each image is assigned to exactly one class, and the multi-level taxonomic structure that naturally exists in many fine-grained domains is largely ignored.

Hierarchical Learning for Category Discovery. A few works Wei et al. (2021); Chang et al. (2021); Du et al. (2021); Qu et al. (2016) have started to bring hierarchical structure into GCD in order to go beyond the single-granularity setting. InfoSieve Rastegar et al. and CiPR Hao et al. (2023) infer abstract hierarchies directly from the data via iterative semi-supervised clustering, so they build a coarse-to-fine tree without using any external taxonomy. TIDA Wang et al. (2023) extends the open-world semi-supervised setting by defining prototypes at several manually chosen granularity levels and forcing the predictions across levels to stay consistent. More recently, SEAL He et al. (2025) relies on naturally occurring biological taxonomies, order, family, and species, to guide a multi-task contrastive framework across granularities. SEAL has two main components: a Hierarchical Semantic-Guided Soft Contrastive loss, which weights negative pairs by their taxonomic distance instead of treating all negatives the same way, and a Cross-Granularity Consistency (CGC) module, which minimises the KL divergence between a fine-level prediction and a pseudo-coarse distribution obtained by passing the fine-level softmax through a learnable species-to-family mapping matrix. GeoGCD builds on this hierarchical formulation but departs by replacing fixed similarity and probability-based alignment with geometry-aware similarity and metric-based consistency.

Diffusion Maps and Manifold-Based Similarity. Diffusion maps Nadler et al. (2005) give a principled way to study the intrinsic geometry of high-dimensional data through random walks on a weighted graph and are applied in machine learning for representation learning Farghly et al. (2025); Wang & Pehlevan (2025); Argov & Wagner (2025). The diffusion distance is defined as the ℓ_2 distance between rows of the t-step Markov transition matrix, so it accumulates information across all paths of length t between two points. This makes the resulting metric robust to small perturbations, naturally multi-scale, and more sensitive to global cluster topology than to single pairwise distances. GeoGCD uses diffusion-based geometry as a supervisory signal that directly guides hierarchical contrastive learning, allowing the representation to reflect emerging density modes of novel classes.

Optimal Transport and Wasserstein Distance. The Wasserstein distance Arjovsky et al. (2017) measures the minimum cost of transporting probability mass from one distribution to another under a ground metric on the sample space. Compared to KL divergence, Wasserstein distance stays finite and informative even when the two distributions have completely disjoint support, and the geometry it induces respects the underlying metric structure. This property has made it popular in generative modelling Arjovsky et al. (2017), domain adaptation, and knowledge distillation. GeoGCD leverages this property for cross-granularity alignment, where predicted and pseudo-coarse distributions frequently have disjoint support during novel-class discovery.

3 Preliminary

3.1 Generalized Category Discovery

We follow the standard GCD setting introduced in Vaze et al. (2022a). Let $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{N_l}$ denote the labelled set, where each image $x_i \in \mathcal{X}$ has a class label $y_i \in \mathcal{Y}_l$, and let $\mathcal{D}_u = \{x_j\}_{j=1}^{N_u}$ denote the unlabelled set, whose underlying labels lie in $\mathcal{Y}_u \supseteq \mathcal{Y}_l$. The novel-class set is $\mathcal{Y}_n = \mathcal{Y}_u \setminus \mathcal{Y}_l$. The total number of classes $|\mathcal{Y}_u| = K$ is either given or estimated by an off-the-shelf method Vaze et al. (2022a).

The goal is to learn an encoder $f_\theta : \mathcal{X} \to \mathbb{R}^d$ together with a classification head $g_\phi : \mathbb{R}^d \to \Delta^{K-1}$ that produces correct predictions on both known and novel classes:

$$\min_{\theta,\phi} \;\; \mathbb{E}_{(x,y)\sim\mathcal{D}_l}[\mathcal{L}_{\text{sup}}(g_\phi(f_\theta(x)), y)] \;+\; \lambda\, \mathbb{E}_{x\sim\mathcal{D}_l\cup\mathcal{D}_u}[\mathcal{L}_{\text{unsup}}(f_\theta(x))]\,, \tag{1}$$

where $\mathcal{L}_{\text{sup}}$ is a supervised loss on the labelled set, $\mathcal{L}_{\text{unsup}}$ is an unsupervised representation loss applied on all images, and $\lambda > 0$ balances the two terms. At evaluation time, accuracy is reported on three splits of the unlabelled set: All, Old (samples whose label is in $\mathcal{Y}_l$), and New (samples whose label is in $\mathcal{Y}_n$).

Hierarchical extension. When a semantic taxonomy is available, each class y at the finest granularity is also associated with labels at $H-1$ coarser levels. We write $y^{(h)} \in \mathcal{Y}^{(h)}$ for the label at level h, with

$h = H$ being the finest level and $h = 1$ the coarsest, and we denote the full hierarchical label as $y^{(1:H)} = (y^{(1)}, \ldots, y^{(H)})$. Hierarchical GCD aims to predict $y^{(1:H)}$ for every image, and to keep predictions across levels consistent, e.g. a sample predicted as a given species should also be predicted as the corresponding family.

3.2 Revisiting the Baseline: Soft Contrastive Learning and Cross-Granularity Alignment

We briefly revisit two design conventions, used by SEAL He et al. (2025) among others, that our method takes as a starting point.

Soft contrastive learning. Standard contrastive losses use a hard $\{0, 1\}$ assignment between positive and negative pairs. A common refinement is to soften this assignment through a similarity matrix $\mathbf{Y}^{(h)} \in [0, 1]^{B \times B}$ derived from the taxonomy, where the shortest-path distance between two labels is converted into a graded similarity. This matrix replaces the binary mask in an InfoNCE-style loss and serves as the soft supervision signal at granularity h.

Cross-granularity alignment. To keep predictions consistent across levels, a learnable mapping matrix $\mathbf{M}^{(h)} \in [0, 1]^{K_H \times K_h}$ from fine to coarse classes (rows summing to one) maps the fine-level prediction $p_i^{(H)}$ to a pseudo-coarse distribution $\tilde{p}_i^{(h)} = (\mathbf{M}^{(h)})^\top p_i^{(H)}$, which is then aligned with the directly predicted $p_i^{(h)}$ by minimising the KL divergence $\mathrm{KL}(p_i^{(h)} \,\|\, \tilde{p}_i^{(h)})$.

Both conventions are taken as a baseline; our contributions lie in how we construct $\mathbf{Y}^{(h)}$ (Section 4.2) and the divergence we use to align $p_i^{(h)}$ with $\tilde{p}_i^{(h)}$ (Section 4.3).

3.3 Diffusion Maps

Diffusion maps Nadler et al. (2005) are a non-linear dimensionality reduction and similarity-analysis tool based on the spectrum of a Markov chain defined on the data graph.

Construction. Let $\mathbf{Z} = [z_1, \ldots, z_N]^\top \in \mathbb{R}^{N \times d}$ be a batch of ℓ_2-normalised features. We first build a Gaussian affinity matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$,

$$W_{ij} \;=\; \exp\!\left(-\frac{\|z_i - z_j\|_2^2}{\sigma^2}\right), \tag{2}$$

where $\sigma > 0$ is a bandwidth (a common choice is the median pairwise distance). We then row-normalise $\mathbf{W}$ to obtain a Markov transition matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$, where $\mathbf{D} = \mathrm{diag}\left(\sum_j W_{ij}\right)$, so that P_{ij} is the one-step probability of jumping from i to j.

Diffusion distance. For an integer scale $t \geq 1$, the t-step transition matrix $\mathbf{P}^t$ has entries $[\mathbf{P}^t]_{ij}$ equal to the probability of going from i to j in exactly t steps. The *diffusion distance* between two points i and j at scale t is the weighted ℓ_2 distance between the corresponding rows of $\mathbf{P}^t$:

$$D_t^2(i, j) \;=\; \sum_{k=1}^{N} \frac{\left([\mathbf{P}^t]_{ik} - [\mathbf{P}^t]_{jk}\right)^2}{\pi_k}, \tag{3}$$

where $\boldsymbol{\pi}$ is the stationary distribution of $\mathbf{P}$. The diffusion distance integrates information over all paths of length t joining i and j, which makes it less sensitive to noise and to elongated cluster shapes than the raw cosine distance. Two points are close under D_t when many short walks of length t connect them, which captures membership in the same density mode rather than mere geometric proximity.

3.4 1-Wasserstein Distance

We finally recall the definition of the Wasserstein distance and the closed form that will be used in the rest of the paper.

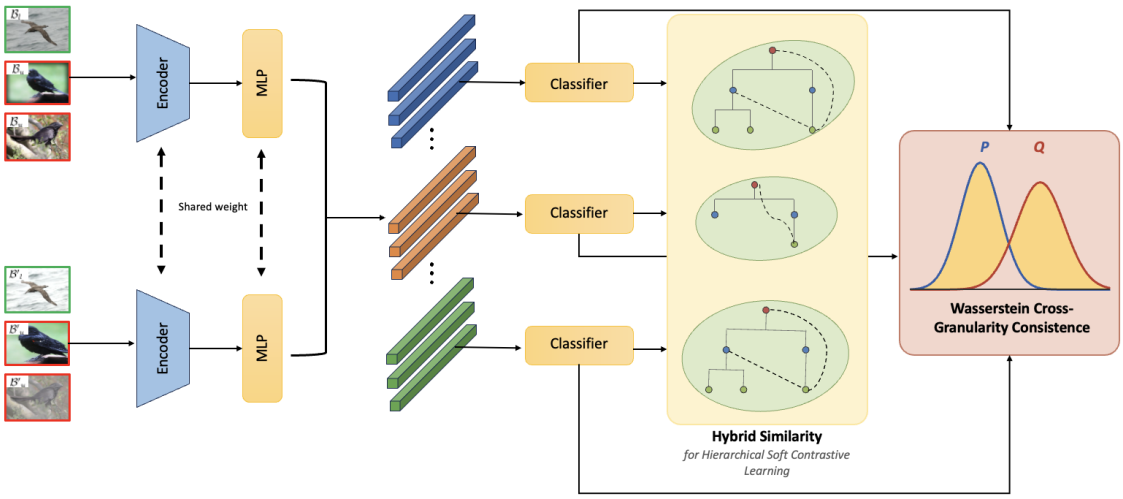

Figure 1: **Overview of the proposed GeoGCD framework.** Two augmented views of each batch are encoded by a shared encoder–MLP backbone and split into per-level feature stacks (one per granularity). Each level feeds a classifier and the **Hybrid Similarity** module, which fuses taxonomic and diffusion-based signals into a soft label matrix for hierarchical contrastive learning. The classifier predictions are then aligned by the **Wasserstein Cross-Granularity Consistency** module.

General definition Villani (2008); Peyré & Cuturi (2019) Given two probability measures p and q on a metric space $(\mathcal{Z}, c)$ with ground cost $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$, the 1-Wasserstein distance is

$$W_1(p, q) \;=\; \inf_{\gamma \in \Pi(p,q)} \int_{\mathcal{Z}\times\mathcal{Z}} c(z, z')\, \mathrm{d}\gamma(z, z'), \tag{4}$$

where $\Pi(p, q)$ is the set of joint distributions (couplings) with marginals p and q.

Closed form on ordinal label spaces Bonneel et al. (2014); Vallender (1974). When the support is one-dimensional and ordinal, e.g. $\mathcal{Z} = \{1, 2, \ldots, K\}$ with the natural metric $c(k, k') = |k - k'|$, W_1 admits a simple closed form in terms of the cumulative distribution functions:

$$W_1(p, q) \;=\; \sum_{k=1}^{K-1} \big|\, F_p(k) - F_q(k) \,\big|, \qquad F_p(k) = \sum_{j \leq k} p(j). \tag{5}$$

This expression is fully differentiable in both p and q and can be evaluated in $\mathcal{O}(K)$ time, which makes it convenient as a training loss.

4 Method

4.1 GeoGCD: Geometry-Guided Hierarchical Generalized Category Discovery

Building on the advantages of semantic hierarchies, we propose the **Hi**erarchical **Geo**metric framework for **GCD** (**GeoGCD**). The overall framework is outlined in Fig. 1. Unlike prior GCD methods that rely on single-granularity supervision Vaze et al. (2022a); Wen et al. (2023); Wang et al. (2024b); Liu & Han (2025), infer abstract hierarchies from data Rastegar et al.; Wang et al. (2023), or guide learning with a fixed taxonomy tree under KL-based alignment He et al. (2025), GeoGCD grounds hierarchical learning in the geometry of two complementary spaces, the feature manifold and the label simplex, through three components: (1) a multi-task backbone producing level-specific features and classifiers; (2) a hybrid soft contrastive loss that fuses taxonomic similarity with a manifold-aware diffusion similarity Nadler et al. (2005),

capturing both prior structure and the evolving cluster geometry of novel classes; and (3) a Wasserstein cross-granularity consistency objective that remains finite and gradient-rich even when the predicted and inferred coarse distributions have nearly disjoint support, the typical regime for novel classes early in training. The remainder of this section details the two new modules in turn: the hybrid contrastive supervision (Section 4.2) and the Wasserstein cross-granularity consistency loss (Section 4.3).

4.2 Hybrid Similarity for Hierarchical Soft Contrastive Learning

For each input x_i, the encoder f_θ produces a backbone feature $\mathbf{e}_i$, which is routed through H projection heads to give level-specific features $\mathbf{f}_i^{(h)} = \pi^{(h)}(\mathbf{e}_i)$, $h = 1, \ldots, H$, all ℓ_2-normalised. At each granularity, we want a soft label matrix that drives contrastive learning by reflecting how related two samples are. A purely taxonomic signal carries strong prior structure but ignores the current geometry of the embedding; a purely manifold-based signal does the opposite. We build this module to leverage the utilities of them.

Taxonomic similarity. Inside a mini-batch of size B, we build a per-level taxonomic similarity $\mathbf{Y}_{\text{hier}}^{(h)} \in \mathbb{R}^{B\times B}$ as a weighted combination of cosine similarities at every granularity, giving more weight to the target level and progressively less to coarser ones. Concretely, with $\mathbf{S}_{ij}^{(h)} = (\mathbf{f}_i^{(h)})^\top \mathbf{f}_j^{(h)}$, we set

$$\mathbf{Y}_{\text{hier}}^{(H)} \;=\; \alpha\, \mathbf{S}^{(H)} + \beta\, \mathbf{S}^{(H-1)} + \gamma\, \mathbf{S}^{(1)}, \tag{6}$$

and analogous mixtures for the coarser levels, with $\alpha \geq \beta \geq \gamma \geq 0$. Each $\mathbf{Y}_{\text{hier}}^{(h)}$ is then min-max normalised and made row-stochastic.

Diffusion similarity. The taxonomic matrix is mostly a function of the prior structure and is blind to the actual cluster geometry of the learned features. We complement it with a diffusion-based similarity that aggregates information along random walks on the batch graph. At each level h, we form a Gaussian affinity

$$\mathbf{W}_{ij}^{(h)} \;=\; \exp\big(-\|\mathbf{f}_i^{(h)} - \mathbf{f}_j^{(h)}\|_2^2/\sigma_W^2\big), \tag{7}$$

row-normalise it into a Markov transition matrix $\mathbf{P}^{(h)} = (\mathbf{D}^{(h)})^{-1}\mathbf{W}^{(h)}$, and apply t diffusion steps. The diffusion distance between samples i and j at scale t is the ℓ_2 distance between the corresponding rows of $(\mathbf{P}^{(h)})^t$, and converting it to a similarity through a Gaussian kernel and row-normalising yields

$$[\mathbf{Y}_{\text{diff}}^{(h)}]_{ij} \;\propto\; \exp\Big(-\big\|[(\mathbf{P}^{(h)})^t]_{i,:} - [(\mathbf{P}^{(h)})^t]_{j,:}\big\|_2 \,/\, \sigma_D\Big). \tag{8}$$

Two samples receive high diffusion similarity when many short walks of length t on the batch graph connect them, that is, when they belong to the same density mode. This captures the structure of *novel* classes as soon as they form coherent clusters, even before their taxonomic position is known.

Hybrid similarity and contrastive loss. We fuse the two through a convex combination with blend coefficient $\lambda \in [0, 1]$ and re-normalise:

$$\mathbf{Y}_{\text{hybrid}}^{(h)} \;=\; \text{rownorm}\Big(\lambda\, \mathbf{Y}_{\text{hier}}^{(h)} + (1-\lambda)\, \mathbf{Y}_{\text{diff}}^{(h)}\Big). \tag{9}$$

The two limits recover known designs: $\lambda = 1$ gives a purely taxonomic signal, and $\lambda = 0$ relies entirely on the data manifold. The hybrid matrix replaces the hard $\{0, 1\}$ assignment in a standard InfoNCE-style loss; with temperature τ, the soft contrastive loss at level h is

$$\mathcal{L}_{\text{con}}^{(h)} \;=\; -\frac{1}{B}\sum_{i=1}^{B}\sum_{j\neq i}[\mathbf{Y}_{\text{hybrid}}^{(h)}]_{ij}\, \log\frac{\exp\big((\mathbf{f}_i^{(h)})^\top \mathbf{f}_j^{(h)}/\tau\big)}{\sum_{k\neq i}\exp\big((\mathbf{f}_i^{(h)})^\top \mathbf{f}_k^{(h)}/\tau\big)}. \tag{10}$$

Pairs with high hybrid similarity contribute strongly to the gradient, while pairs with low similarity behave as hard negatives, giving a graded supervision signal that adapts both to the taxonomy and to the evolving feature geometry. The diffusion matrix is detached from the computation graph; gradients flow only through the features $\mathbf{f}_i^{(h)}$ inside the contrastive term. Small values of t ($t \in \{2, 3\}$) are sufficient in practice: larger t collapses the random walk towards the stationary distribution and erases the cluster structure we want to preserve.

4.3 Wasserstein Cross-Granularity Consistency

The second module ties together predictions made at different granularities. We adopt the mapping-matrix design reviewed in Section 3.2: at each coarse level h, the model produces a directly predicted coarse distribution $p_i^{(h)} = g_\phi^{(h)}(\mathbf{e}_i) \in \Delta^{K_h-1}$ and a pseudo-coarse distribution $\tilde{p}_i^{(h)} = (\mathbf{M}^{(h)})^\top p_i^{(H)}$ obtained by routing the fine-level prediction through the learnable mapping matrix $\mathbf{M}^{(h)}$. Cross-granularity consistency requires the two distributions to agree, since they describe the same underlying coarse class of x_i. Where our design departs from prior work is in the choice of divergence used to enforce this agreement.

Wasserstein objective. Building on the properties reviewed in Section 3.4, we replace the standard KL term with the 1-Wasserstein distance, which on the ordinal coarse-class space with ground metric $c(k,k')=|k-k'|$ admits the closed form

$$\mathcal{L}_{\mathrm{W-CGC}}^{(h)} \;=\; W_1\big(p_i^{(h)},\, \tilde{p}_i^{(h)}\big) \;=\; \sum_{k=1}^{K_h-1} \big|F_{p_i^{(h)}}(k) - F_{\tilde{p}_i^{(h)}}(k)\big|, \tag{11}$$

with $F_p(k) = \sum_{j\le k} p(j)$. This expression is differentiable in both arguments, runs in $\mathcal{O}(K_h)$ time, and inherits the coarse-class ordering from the taxonomy itself. The qualitative claim that W_1 stays bounded, continuous and gradient-rich under disjoint support is formalised in Lemma 1 and Propositions 1 and 2 in the appendix.

4.4 Final Training Objective

The full training objective combines a per-level supervised cross-entropy on the labelled set, the hybrid soft contrastive loss in equation 10, and the Wasserstein cross-granularity consistency loss in equation 11:

$$\mathcal{L} \;=\; \sum_{h=1}^{H}\Big(\mathcal{L}_{\mathrm{sup}}^{(h)} + \mu\,\mathcal{L}_{\mathrm{con}}^{(h)}\Big) \;+\; \nu \sum_{h=1}^{H-1} \mathcal{L}_{\mathrm{W-CGC}}^{(h)}, \tag{12}$$

with scalar weights $\mu, \nu > 0$.

4.5 Formal Analysis of the Wasserstein Cross-Granularity Objective

This appendix collects the formal statements and proofs supporting the discussion in Section 4.3. Throughout, we work on a finite ordinal label space $\mathcal{Y}=\{1,\dots,K\}$ with ground metric $c(k,k')=|k-k'|$, and write $F_p(k)=\sum_{j\le k} p(j)$ for the cumulative distribution function of $p\in\Delta^{K-1}$.

Lemma 1 (Disjoint support is the typical regime at initialisation)**.** *Consider a coarse level h with K_h classes. Assume that at initialisation $p_i^{(h)}$ is approximately concentrated on a single coarse index drawn uniformly from $\{1,\dots,K_h\}$, and that the pseudo-coarse distribution $\tilde{p}_i^{(h)}$, obtained by routing the fine-level prediction through a randomly initialised mapping matrix $\mathbf{M}^{(h)}$, is approximately concentrated on a coarse index drawn independently and uniformly from the same set. Then*

$$\Pr\big[\,\mathrm{supp}(p_i^{(h)}) \cap \mathrm{supp}(\tilde{p}_i^{(h)}) = \emptyset\,\big] \;\ge\; 1 - \frac{1}{K_h}. \tag{13}$$

More generally, if both distributions concentrate on s coarse indices each, chosen uniformly at random and independently, the probability of disjoint supports is at least $1 - s^2/K_h$.

Proof. For the single-mode case, the two predicted indices coincide with probability $1/K_h$ (uniform independent draws), so they are disjoint with probability $1 - 1/K_h$. For the s-mode case, the expected size of the intersection of two uniformly chosen size-s subsets of $\{1,\dots,K_h\}$ is s^2/K_h, so by Markov's inequality the probability that the intersection is non-empty is at most s^2/K_h, which gives the stated bound. □

Proposition 1 (Boundedness and continuity under disjoint support)**.** *Let $p, q \in \Delta^{K-1}$.*

1. *$W_1(p,q) \le K-1$, and W_1 is continuous on $\Delta^{K-1} \times \Delta^{K-1}$.*

2. *There exist sequences p_n, q_n with disjoint supports such that $\mathrm{KL}(p_n \,\|\, q_n) = +\infty$ for all n, while $W_1(p_n, q_n) \to 0$.*

Proof. **(1)** For any coupling $\gamma \in \Pi(p,q)$, $\sum_{k,k'} |k-k'|\,\gamma(k,k') \le (K-1)\sum_{k,k'} \gamma(k,k') = K-1$. Thus, $W_1(p,q) \le K-1$. Since the CDF form $W_1(p,q) = \sum_{k=1}^{K-1} |F_p(k) - F_q(k)|$ is a finite sum of continuous functions of p, q, it is continuous on the product simplex.

(2) Take $p_n = \delta_{k_0}$ and $q_n = \delta_{k_0+\epsilon_n}$ with $\epsilon_n \downarrow 0$ along a refined grid (or, in the discrete case, adjacent coordinates that get closer in some refinement). Since the supports are disjoint, $\mathrm{KL}(p_n \,\|\, q_n)$ contains a term $p_n(k_0)\log(p_n(k_0)/0) = +\infty$. CDFs F_{p_n}, F_{q_n} differ only in a vanishing interval, then $W_1(p_n, q_n) \to 0$. □

Proposition 2 (Non-vanishing gradient at distributional separation)**.** *Let $p_\theta \in \Delta^{K-1}$ be smoothly parametrised by $\theta \in \mathbb{R}^d$ and let $q \in \Delta^{K-1}$ be fixed. Suppose* $\mathrm{supp}(p_{\theta_0}) \cap \mathrm{supp}(q) = \emptyset$.

1. $\mathrm{KL}(q \,\|\, p_{\theta_0}) = +\infty$ *(its gradient is undefined), and* $\mathrm{KL}(p_{\theta_0} \,\|\, q)$ *is finite only if p_{θ_0} vanishes wherever q does, a condition that fails under disjoint support. When the latter is finite, its gradient depends only on the overlapping support and carries no information about the spatial separation between* $\mathrm{supp}(p_{\theta_0})$ *and* $\mathrm{supp}(q)$.

2. *The partial gradient of W_1 with respect to $p(j)$ has the explicit form*

$$\frac{\partial W_1(p,q)}{\partial p(j)} \;=\; \sum_{k=j}^{K-1} \mathrm{sign}\big(F_p(k) - F_q(k)\big), \tag{14}$$

which is non-zero whenever $p \neq q$, and whose magnitude grows with the number of indices at which the CDFs disagree, i.e. with the spatial separation between $\mathrm{supp}(p)$ *and* $\mathrm{supp}(q)$.

Proof. **(1)** Under disjoint support, there exists k with $q(k) > 0$ and $p_{\theta_0}(k) = 0$, so $\mathrm{KL}(q \,\|\, p_{\theta_0}) = +\infty$ and the gradient is undefined. For the reverse direction, finiteness requires $p_{\theta_0}(k) = 0$ wherever $q(k) = 0$, which fails when supports are disjoint; when finite, the gradient $\partial_\theta \mathrm{KL}(p_{\theta_0} \,\|\, q) = \sum_k (\log(p_{\theta_0}(k)/q(k)) + 1)\, \partial_\theta p_{\theta_0}(k)$ depends only on the ratio p_{θ_0}/q on the overlap, with no dependence on the geometric distance between the two supports.

(2) Differentiating $W_1(p,q) = \sum_{k=1}^{K-1} |F_p(k) - F_q(k)|$ with respect to $p(j)$ and using $\partial F_p(k)/\partial p(j) = \mathbb{1}[j \le k]$ yields equation 14. Each surviving term contributes ± 1, so the magnitude is the number of CDF disagreements, which grows with the support separation. □

5 Experiments

5.1 Experimental Setup

Datasets. We conduct experiments on three fine-grained image classification benchmarks that admit a natural multi-level taxonomy, following the Semantic Shift Benchmark (SSB) protocol Vaze et al. (2022b) **CUB-200-2011** Wah et al. (2011) (200 bird species), **Stanford Cars** Krause et al. (2013) (196 car models), and **FGVC-Aircraft** Maji et al. (2013) (100 aircraft variants). We follow the standard GCD split Vaze et al. (2022a): half of the classes are designated as known $\mathcal{Y}_l$, and 50% of the images from each known class form the labelled set $\mathcal{D}_l$; all remaining images, from both known and novel classes, constitute the unlabelled set $\mathcal{D}_u$. A summary of the three datasets is given in Table 1.

Evaluation metrics. We follow the standard GCD evaluation protocol Vaze et al. (2022a). At test time, predictions on the unlabelled set are matched to ground-truth labels via the Hungarian algorithm, which

Table 1: **Overview of datasets.** $|\mathcal{Y}_l|, |\mathcal{Y}_u|$ denote the number of classes in the labelled and unlabelled sets, and $|\mathcal{D}_l|, |\mathcal{D}_u|$ the corresponding number of images. *Balance* indicates whether the class distribution is balanced.

Dataset	**Balance**	$\lvert\mathcal{D}_l\rvert$	$\lvert\mathcal{Y}_l\rvert$	$\lvert\mathcal{D}_u\rvert$	$\lvert\mathcal{Y}_u\rvert$
CUB-200-2011 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

resolves the unknown correspondence between predicted and true novel classes. We then report classification accuracy (ACC), defined as

$$\text{ACC} \;=\; \frac{1}{|\mathcal{D}_u|}\sum_{i=1}^{|\mathcal{D}_u|} \mathbb{1}\big(\, y_i = \mathbf{h}(\hat{y}_i)\,\big), \tag{15}$$

where $\hat{y}_i$ is the prediction for sample x_i, y_i is its ground-truth label, and $\mathbf{h}$ is the optimal one-to-one mapping between predicted and ground-truth class indices obtained by the Hungarian algorithm. We report ACC on three subsets of the unlabelled set: **All** (the full unlabelled set $\mathcal{D}_u$), **Old** (unlabelled images from known classes $\mathcal{Y}_l$), and **New** (unlabelled images from novel classes $\mathcal{Y}_n = \mathcal{Y}_u \setminus \mathcal{Y}_l$). Reporting Old and New separately is important since GCD methods can trade off accuracy across the two splits, and the All metric alone hides such imbalances.

Implementation details. We use a DINOv2 ViT-B/14 Oquab et al. (2023) backbone and fine-tune only the last transformer block, following common GCD practice He et al. (2025); Liu & Han (2025). The MLP projection head produces H per-level features of dimension $d' = 768$, all ℓ_2-normalised before similarity computation. We train for 200 epochs on a single NVIDIA RTX A6000 GPU with batch size $B = 128$, using AdamW with a cosine learning rate schedule. The taxonomic-similarity weights are $(\alpha, \beta, \gamma) = (0.6, 0.3, 0.1)$ on three-level datasets (CUB and FGVC-Aircraft), and analogous values on two-level Stanford Cars. For the diffusion similarity, we use $t = 2$ steps with bandwidths σ_W, σ_D set to the median pairwise distance inside the mini-batch, and a hybrid blend coefficient $\lambda = 0.5$. The Wasserstein consistency weight is set to $\nu = 0.1$, which approximately matches the magnitude of the contrastive and supervised terms; we discuss the sensitivity of ν in Section 5.3.

5.2 Main Results

Table 2 compares GeoGCD against representative GCD methods on the SSB benchmark with the DINOv2 ViT-B/14 backbone. GeoGCD obtains the best All accuracy on CUB and Stanford Cars, improving the SEAL baseline by +3.23 All on CUB and +2.40 All on Stanford Cars. Moreover, the significant result achieves **87.73** on CUB with Old setting, higher than the SOTA DebGCD by **6.93**.

The largest improvements appear on the **New** split, which is the primary target of category discovery and the regime where most GCD methods struggle. On CUB, GeoGCD reaches 79.93 All / 87.73 Old / 72.20 New, and on Stanford Cars 80.10 All / 88.70 Old / 67.11 New. The consistent gains across both datasets, despite their very different visual statistics, suggest that the geometric signals introduced by GeoGCD generalise well across domains.

5.3 Ablation Study

We isolate the contribution of each module on CUB and Stanford Cars by progressively adding the diffusion-based similarity (**+ Diff**) and the Wasserstein cross-granularity loss (**+ W**) to the soft contrastive baseline that uses taxonomic similarity and KL consistency. Results are reported in Table 3.

Effect of the diffusion similarity. Adding the diffusion-based signal to the taxonomic similarity (*+ Diff*) substantially improves Old accuracy on both datasets (+9.29 on CUB, +4.77 on Stanford Cars), which indicates that the manifold-aware similarity helps the model exploit the cluster structure of known classes

Table 2: **Comparison of GCD methods on the SSB benchmark** Vaze et al. (2022b) with the DINOv2 ViT-B/14 backbone. Accuracy (%) is reported on the All / Old / New splits; the rightmost column is the average All across the three datasets. The highest and second-highest scores in each column are in **bold** and underlined. Results for prior methods are taken from He et al. (2025).

		CUB			**Stanford Cars**			**FGVC-Aircraft**			**Avg.**
Method	**Venue**	All	Old	New	All	Old	New	All	Old	New	All
k-means MacQueen (1967)	–	67.6	60.6	71.1	29.4	24.5	31.8	18.9	16.9	19.9	38.6
GCD Vaze et al. (2022a)	CVPR'22	71.9	71.2	72.3	65.7	67.8	64.7	55.4	47.9	59.2	64.3
CiPR Hao et al. (2023)	TMLR'24	78.3	73.4	**80.8**	66.7	77.0	61.8	59.2	65.0	56.3	68.1
SimGCD Wen et al. (2023)	ICCV'23	71.5	78.1	68.3	71.5	81.9	66.6	63.9	69.9	60.9	69.0
μGCD Vaze et al. (2023)	NeurIPS'23	74.0	75.9	73.1	76.1	91.0	68.9	66.3	68.7	65.1	72.1
SPTNet Wang et al. (2024b)	ICLR'24	76.3	79.5	74.6	–	–	–	–	–	–	–
DebGCD Liu & Han (2025)	ICLR'25	77.5	80.8	75.8	75.4	87.7	69.5	71.9	76.0	69.8	74.9
SEAL He et al. (2025)	NeurIPS'25	76.7	78.3	75.9	77.7	88.7	**72.4**	**74.6**	73.2	**75.3**	76.3
GeoGCD (Ours)	–	**79.93**	**87.73**	72.20	**80.10**	**93.57**	67.11	69.01	**79.92**	58.08	**76.3**

Table 3: **Ablation on CUB-200-2011, Stanford Cars and FGVC-Aircraft** Starting from the soft contrastive baseline with taxonomic similarity and KL consistency, we progressively add the diffusion-based similarity to the contrastive supervision and replace the KL term by the 1-Wasserstein objective.

	CUB-200-2011			**Stanford Cars**			**FGVC-Aircraft**		
Variant	**All**	**Old**	**New**	**All**	**Old**	**New**	**All**	**Old**	**New**
Baseline (taxonomic + KL)	76.70	78.30	**75.90**	77.70	88.70	**72.40**	**74.60**	73.20	**75.30**
+ Diff	76.44	87.59	65.40	78.47	93.47	64.01	73.51	**81.77**	65.23
+ Diff + W **(GeoGCD)**	**79.93**	**87.73**	72.20	**80.10**	**93.57**	67.11	69.01	79.92	58.08

more effectively: the random-walk distances on the batch graph capture density modes that the taxonomy tree cannot directly express, sharpening intra-cluster similarity for known classes. The downside, however, is consistent across both datasets: New accuracy drops compared to the baseline (-10.50 on CUB, -8.39 on Stanford Cars), which suggests that the diffusion signal alone, when paired with the standard KL consistency, can over-commit the representation to the known-class geometry: novel classes are still poorly clustered early in training, so the data-driven similarity inherits this noise and pushes the consistency loss towards reinforcing known-class structure rather than discovering new modes.

Effect of the Wasserstein consistency. Replacing the KL term by the 1-Wasserstein objective (*+ Diff + W*) yields the best result on every metric across both datasets. Compared to the baseline, All accuracy improves by +3.23 on CUB and +2.40 on Stanford Cars; compared to *+ Diff*, the gain on the New split is +6.80 on CUB and +3.10 on Stanford Cars. On Stanford Cars, the Wasserstein objective also slightly improves Old accuracy further over *+ Diff* ($93.47 \rightarrow 93.57$), showing that the alignment signal does not trade off known-class accuracy when stabilised this way. The recovery on the New split is particularly meaningful: it shows that swapping KL for W_1 does not merely stabilise training, but specifically restores the ability of the cross-level alignment to benefit novel classes, even when the upstream contrastive signal is biased towards known-class geometry. This pattern is consistent with the disjoint-support analysis of Lemma 1 and Propositions 1 and 2: when predictions over novel classes are noisy and the predicted and inferred coarse distributions concentrate on nearly disjoint regions of the label simplex, KL is either undefined or carries gradients that ignore the spatial separation of supports, while W_1 remains bounded and provides a non-vanishing gradient that shrinks the geometric distance between the two distributions. The Wasserstein term thus complements the diffusion similarity by allowing the cross-level alignment signal to remain informative exactly in the regime where novel classes are most fragile.

Sensitivity to the consistency weight. The Wasserstein term has a different magnitude than KL, so its loss weight ν has to be rescaled accordingly. Setting $\nu = 1.0$ destabilises training; rescaling to $\nu = 0.1$, which approximately matches the contrastive-term magnitude, recovers stable training and yields the gains reported above. A full sensitivity sweep over ν is provided in the appendix.

Summary. The two contributions are complementary: the diffusion similarity strengthens the representation of known classes by injecting data-driven manifold structure, while the Wasserstein consistency ensures that this geometric signal propagates correctly across granularities, especially on novel classes whose coarse-level distributions are still poorly aligned with the pseudo-coarse ones. Their combination yields the best result on every reported split across both CUB-200-2011 and Stanford Cars.

6 Conclusion

We presented GeoGCD, a geometry-guided hierarchical framework for Generalized Category Discovery built around the principle that the global geometry of semantics, the shape that taxonomic relations carve in feature and label space, should be a first-class learning target rather than an auxiliary signal. The framework realises this principle through two components: a hybrid soft-label matrix that fuses taxonomic similarity with a manifold-aware diffusion similarity for hierarchical contrastive learning, and a Wasserstein cross-granularity consistency loss that aligns predictions across levels with formal guarantees of finiteness, continuity, and gradient informativeness under disjoint support, the regime that arises with overwhelming probability at random initialisation. Empirically, GeoGCD attains state-of-the-art accuracy on CUB and Stanford Cars, with the largest gains concentrated on the splits that prior methods find hardest. Sensitivity analysis further shows that the geometric framework, rather than careful hyperparameter tuning, is what drives these improvements.

Limitations. Three limitations of the present work are worth noting. First, our theoretical analysis of the 1-Wasserstein term characterises the disjoint-support regime that is typical at random initialisation, but does not provide a tight description of the intermediate phase of training where the predicted and pseudo-coarse distributions begin to overlap. Second, GeoGCD assumes that a curated semantic taxonomy is available at training time and is not designed for settings where the hierarchy itself must be inferred jointly with the representation. Third, the diffusion similarity is computed within each mini-batch, so the random-walk structure that informs the hybrid soft-label matrix is local relative to the batch size; extending the construction to a global graph defined over a memory bank could strengthen the manifold-aware component at the cost of additional bookkeeping.

Acknowledgments

TODO: Acknowledgments (hidden in anonymous submission).

References

Shabib Aftab and Samia Akhtar. Diabetic retinopathy severity classification using data fusion and ensemble transfer learning. *Journal of Software Engineering and Applications*, 18(1):1–23, 2025.

Tal Argov and Tal Wagner. Spex: A spectral approach to explainable clustering. *arXiv preprint arXiv:2511.00885*, 2025.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. Pmlr, 2017.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51, 04 2014. doi: 10.1007/s10851-014-0506-3.

Fernando Julio Cendra, Bingchen Zhao, and Kai Han. Promptccd: Learning gaussian mixture prompt pool for continual category discovery. In *European conference on computer vision*, pp. 188–205. Springer, 2024.

Dongliang Chang, Kaiyue Pang, Yixiao Zheng, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Your" flamingo" is my" bird": Fine-grained, or not. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11476–11485, 2021.

Ruoyi Du, Jiyang Xie, Zhanyu Ma, Dongliang Chang, Yi-Zhe Song, and Jun Guo. Progressive learning of category-consistent multi-granularity features for fine-grained visual classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9521–9535, 2021.

Tyler Farghly, Peter Potaptchik, Samuel Howard, George Deligiannidis, and Jakiw Pidstrigach. Diffusion models and the manifold hypothesis: Log-domain smoothing is geometry adaptive. *arXiv preprint arXiv:2510.02305*, 2025.

Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8401–8409, 2019.

Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. *arXiv preprint arXiv:2002.05714*, 2020.

Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel: Automatically discovering and learning novel visual categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6767–6781, 2021.

Shaozhe Hao, Kai Han, and Kwan-Yee K Wong. Cipr: An efficient framework with cross-instance positive relations for generalized category discovery. *arXiv preprint arXiv:2304.06928*, 2023.

Zhenqi He, Yuanpei Liu, and Kai Han. Seal: Semantic-aware hierarchical learning for generalized category discovery. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Xuhui Jia, Kai Han, Yukun Zhu, and Bradley Green. Joint representation learning and novel category discovery on single-and multi-modal data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 610–619, 2021.

KJ Joseph, Sujoy Paul, Gaurav Aggarwal, Soma Biswas, Piyush Rai, Kai Han, and Vineeth N Balasubramanian. Novel class discovery without forgetting. In *European Conference on Computer Vision*, pp. 570–586. Springer, 2022.

Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue. *(No Title)*, 2018.

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.

Yuanpei Liu and Kai Han. Debgcd: Debiased learning with distribution guidance for generalized category discovery. *arXiv preprint arXiv:2504.04804*, 2025.

J MacQueen. Multivariate observations. In *Proceedings ofthe 5th Berkeley symposium on mathematical statisticsand probability*, volume 1, pp. 281–297. University of California press Oakland, CA, USA, 1967.

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

Boaz Nadler, Stephane Lafon, Ioannis Kevrekidis, and Ronald Coifman. Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. *Advances in neural information processing systems*, 18, 2005.

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.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11:355–206, 02 2019. doi: 10.1561/2200000073.

Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7579–7588, 2023.

Yanyun Qu, Li Lin, Fumin Shen, Chang Lu, Yang Wu, Yuan Xie, and Dacheng Tao. Joint hierarchical category structure learning and large-scale image classification. *IEEE Transactions on Image Processing*, 26(9):4331–4346, 2016.

Sarah Rastegar, Hazel Doughty, and Cees GM Snoek. Learn to categorize or categorize to learn? self-coding for generalized category discovery supplementary materials.

S. S. Vallender. Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786, 1974. doi: 10.1137/1118101. URL `https://doi.org/10.1137/1118101`.

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*, pp. 7492–7501, 2022a.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. The semantic shift benchmark. In *ICML 2022 Shift Happens Workshop*, 2022b.

Sagar Vaze, Andrea Vedaldi, and Andrew Zisserman. No representation rules them all in category discovery. *Advances in Neural Information Processing Systems*, 36:19962–19989, 2023.

Cédric Villani. *Optimal transport – Old and new*, volume 338, pp. xxii+973. 01 2008. doi: 10.1007/978-3-540-71050-9.

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

Binxu Wang and Cengiz Pehlevan. An analytical theory of spectral bias in the learning dynamics of diffusion models. *arXiv preprint arXiv:2503.03206*, 2025.

Hongjun Wang, Sagar Vaze, and Kai Han. Hilo: A learning framework for generalized category discovery robust to domain shifts. *arXiv preprint arXiv:2408.04591*, 2024a.

Hongjun Wang, Sagar Vaze, and Kai Han. Sptnet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. *arXiv preprint arXiv:2403.13684*, 2024b.

Yu Wang, Zhun Zhong, Pengchong Qiao, Xuxin Cheng, Xiawu Zheng, Chang Liu, Nicu Sebe, Rongrong Ji, and Jie Chen. Discover and align taxonomic context priors for open-world semi-supervised learning. *Advances in Neural Information Processing Systems*, 36:19015–19028, 2023.

Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8927–8948, 2021.

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*, pp. 16590–16600, 2023.

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*, pp. 3479–3488, 2023.

Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10867–10875, 2021.

A Analysis on Generic Datasets

To assess whether the geometric framework of GeoGCD generalises beyond the standard fine-grained natural-image benchmarks, we evaluate it on two domain-specific datasets that exhibit very different visual statistics from CUB or Stanford Cars: a clinical retinal imaging dataset with strong class imbalance and an ordinal severity structure, and a histopathology dataset with balanced texture-like classes. Unlike the SSB benchmarks, both datasets exhibit a shallow 2-level taxonomy: a single domain-level root grouping the leaf classes, with no intermediate family structure. This setting tests whether GeoGCD remains effective when the hierarchical prior degenerates and the diffusion similarity has to carry most of the structural signal.

Datasets. **APTOS 2019** Aftab & Akhtar (2025) is a diabetic retinopathy (DR) grading benchmark of 3,662 retinal fundus images annotated with five severity grades (No DR, Mild, Moderate, Severe, Proliferative). The hierarchy is shallow: all five grades share a common root *DR Severity*. We treat the three earliest grades (No DR, Mild, Moderate; 3,174 images, 86.7%) as known classes $\mathcal{Y}_l$ and the two most severe grades (Severe, Proliferative; 488 images, 13.3%) as novel classes $\mathcal{Y}_n$. The dataset poses two distinctive challenges relative to the SSB benchmark: a strong class imbalance and an ordinal label space, where adjacent severity grades are visually similar.

NCT-CRC-HE-100K Kather et al. (2018) is a colorectal histopathology dataset of 100,000 patches at 224×224 resolution, annotated into nine tissue types (adipose, background, debris, lymphocytes, mucus, smooth muscle, normal mucosa, stroma, tumour). All nine classes share a single root *Tissue*. We designate the first five classes (ADI, BACK, DEB, LYM, MUC; 52,938 images, 52.9%) as known and the remaining four (MUS, NORM, STR, TUM; 47,062 images, 47.1%) as novel. Compared to APTOS, the class distribution is approximately balanced (each class contains between 8.7K and 14.3K images) and the visual content is texture-dominated, which is closer to the regime in which diffusion-based similarity is expected to be informative. A class-level breakdown of both datasets is given in Table 4.

Results. Table 5 reports the best All accuracy obtained by GeoGCD with and without the Wasserstein cross-granularity term, together with a no-diffusion baseline on NCT-CRC-HE-100K.

Discussion. On NCT-CRC-HE-100K, GeoGCD attains the highest accuracy on both the All and New splits (95.32 All / 90.38 New), improving the no-diffusion baseline by +1.89 All and the diffusion-only variant by +0.42 All, with consistent gains on novel classes (+3.94 over no-diffusion). The Old accuracy remains essentially saturated above 99.6 across all variants, indicating that the additional geometric signal does not harm known-class performance even when the latter is already very strong. The fact that GeoGCD continues to improve under the shallow *Tissue* root, where the taxonomic prior carries no discriminative information beyond a single grouping, confirms that the manifold-aware diffusion signal is the main contributor in this regime, and that the Wasserstein term propagates this signal across granularities without destabilising training.

On APTOS 2019, the diffusion-only variant achieves the best overall and known-class accuracy (78.62 All, 84.72 Old), while GeoGCD with the Wasserstein term recovers a notably higher accuracy on novel classes (44.06 vs. 38.93, +5.13). The reversed ranking on the All metric is a direct consequence of the strong class imbalance (86.7% known): a gain on the small novel set (13.3%) is more than compensated by a small drop on the dominant known set when predictions are summed over $\mathcal{D}_u$. Importantly, the trend on the

Table 4: **Class-level overview of the generic datasets.** For each class we report the total image count and its split into labelled ($\mathcal{D}_l$) and unlabelled ($\mathcal{D}_u$) sets. Known classes contribute 80% of their images to $\mathcal{D}_l$ and the remaining 20% to $\mathcal{D}_u$; novel classes contribute their entire content to $\mathcal{D}_u$. The full unlabelled pool is used for evaluation.

Cls	**Name**	**Total**	$\lvert\mathcal{D}_l\rvert$	$\lvert\mathcal{D}_u\rvert$	**Type**
APTOS 2019					
0	No DR	1,805	1,444	361	Known
1	Mild	370	296	74	Known
2	Moderate	999	799	200	Known
3	Severe	193	0	193	Novel
4	Proliferative	295	0	295	Novel
	Total	3,662	2,539	1,123	
NCT-CRC-HE-100K					
0	ADI (adipose)	10,407	8,326	2,081	Known
1	BACK (background)	10,566	8,453	2,113	Known
2	DEB (debris)	11,512	9,210	2,302	Known
3	LYM (lymphocytes)	11,557	9,246	2,311	Known
4	MUC (mucus)	8,896	7,117	1,779	Known
5	MUS (smooth muscle)	13,536	0	13,536	Novel
6	NORM (normal mucosa)	8,763	0	8,763	Novel
7	STR (stroma)	10,446	0	10,446	Novel
8	TUM (tumour)	14,317	0	14,317	Novel
	Total	100,000	42,352	57,648	

Table 5: **Results on generic datasets.** "Diffusion" refers to the soft contrastive baseline with hybrid diffusion-taxonomic similarity and KL cross-granularity consistency; "Diff + W" is GeoGCD with the 1-Wasserstein consistency replacing KL; "No Diff" uses purely taxonomic similarity. Best All in **bold**.

Dataset	**Variant**	**All**	**Old**	**New**
APTOS 2019	Diffusion	**78.62**	**84.72**	38.93
	Diff + W (**GeoGCD**)	76.73	81.76	**44.06**
NCT-CRC-HE-100K	No Diff	93.43	99.64	86.44
	Diffusion	94.90	**99.78**	89.41
	Diff + W (**GeoGCD**)	**95.32**	99.71	**90.38**

New split, the regime that category discovery is ultimately concerned with, follows the same direction as on the SSB benchmarks: replacing KL by W_1 improves novel-class performance, which is consistent with the disjoint-support analysis of Lemma 1 and Propositions 1 and 2, and is particularly relevant on the ordinal label space of DR severity, where the 1-Wasserstein distance respects the natural ordering of severity grades.

B Analysis on Multiple Settings

We conduct sensitivity analysis on three key hyperparameters of GeoGCD: the hybrid blend coefficient λ controlling the balance between taxonomic and diffusion similarities, the number of diffusion steps t, and the Gaussian kernel bandwidth σ_W used in the diffusion graph. All ablations are run on CUB-200-2011, Stanford Cars, and FGVC-Aircraft with the default configuration ($\lambda = 0.5$, $t = 2$, $\sigma_W = 1.0$) as the reference; results are reported in Tables 6, 7, and 8.

Table 6: **Sensitivity to the hybrid blend coefficient λ.** $\lambda = 0$ uses only the diffusion similarity while $\lambda = 1$ uses only the taxonomic similarity; the default value $\lambda = 0.5$ is reported in Table 2. Best results per dataset in **bold**.

	CUB-200-2011			**Stanford Cars**			**FGVC-Aircraft**		
$\boldsymbol{\lambda}$	All	Old	New	All	Old	New	All	Old	New
0.00	**78.87**	**86.55**	**71.27**	80.05	**92.50**	68.04	**77.14**	81.12	**73.15**
0.25	77.11	85.92	68.38	–	–	–	75.94	81.18	70.69
0.75	77.82	85.75	69.97	**80.34**	92.00	**69.09**	–	–	–
1.00	77.93	86.03	69.90	79.49	**92.50**	66.94	–	–	–

Table 7: **Sensitivity to the number of diffusion steps t.** The default value $t = 2$ is reported in Table 2. Best results per dataset in **bold**.

	CUB-200-2011			**Stanford Cars**			**FGVC-Aircraft**		
$\boldsymbol{t}$	All	Old	New	All	Old	New	All	Old	New
1	**79.72**	**87.07**	**72.44**	**80.84**	**94.55**	**67.60**	74.68	78.72	**70.63**
3	78.79	86.79	70.86	–	–	–	**74.83**	**79.08**	70.57
5	78.65	86.37	71.00	79.59	92.98	66.67	71.86	80.22	63.48

Effect of the hybrid blend λ. Table 6 shows that the optimal λ depends on the visual structure of the dataset. On CUB, $\lambda = 0$ (pure diffusion) achieves the best result (78.87 All), with the gap to other values around 1–1.8 points; this suggests that for fine-grained biological categories with elongated, density-varying clusters, the manifold-aware diffusion signal is sufficient and the taxonomic prior is partially redundant. On Stanford Cars, the trend reverses: $\lambda = 0.75$ (heavy taxonomic weight) yields the best result (80.34 All), indicating that the rigid visual differences across car models are better captured by the taxonomy than by the feature manifold. On Aircraft, $\lambda = 0$ again outperforms larger values, similar to CUB. Overall, $\lambda \in [0, 0.75]$ is a safe range, and the choice within this range adapts to the geometry of each domain.

Effect of the diffusion steps t. Table 7 reveals a consistent preference for $t = 1$ across CUB and Stanford Cars. On CUB, $t = 1$ achieves 79.72 All, decreasing monotonically as t grows; on Stanford Cars, $t = 1$ similarly dominates (80.84 All vs. 79.59 at $t = 5$). This is consistent with the spectral interpretation of diffusion maps (Section 3.3): larger t progressively erases fine local structure as the random walk approaches the stationary distribution, weakening the discriminative signal that the contrastive loss relies on. On Aircraft, $t = 1$ and $t = 3$ are comparable, with $t = 5$ clearly worse, supporting the same conclusion.

Effect of the Gaussian kernel bandwidth σ_W. Table 8 shows that GeoGCD is robust across $\sigma_W \in \{0.5, 2.0, 5.0\}$ on CUB, with All accuracy varying within $\sim$1.5 points. The best result is obtained at $\sigma_W = 5.0$ (79.43 All), although the smaller $\sigma_W = 0.5$ gives the highest New accuracy (72.99). This suggests a small trade-off between intra-cluster sharpness (small σ_W, favouring novel-class discovery) and global connectivity (larger σ_W, favouring known-class consolidation), and indicates that the median-distance heuristic used in our default configuration is a reasonable compromise.

Summary. The three hyperparameters control different aspects of the geometric signal: λ balances prior versus data-driven similarity, t controls the locality scale of the random walk, and σ_W governs the connectivity of the batch graph. Across all three, GeoGCD remains within a narrow performance band, with no single setting dominating in absolute terms. This stability suggests that the geometric framework itself, rather than careful hyperparameter tuning, is what drives the improvements reported in Section 5.2.

Table 8: **Sensitivity to the Gaussian kernel bandwidth σ_W on CUB-200-2011.** The default value $\sigma_W = 1.0$ is reported in Table 2. Best results in **bold**.

$\boldsymbol{\sigma_W}$	**All**	**Old**	**New**
0.5	78.75	84.57	**72.99**
2.0	77.89	85.40	70.45
5.0	**79.43**	**86.23**	72.68