# PROTOTAIL: PROTOTYPE-GUIDED REPRESENTATION AND PSEUDO-LABELING FOR LONG-TAILED FINE-GRAINED GENERALIZED CATEGORY DISCOVERY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Generalized Category Discovery (GCD) in fine-grained recognition is challenging because categories often differ only in subtle local cues, while long-tailed training provides uneven evidence for learning category-specific representations. When category support is limited, flexible local representations can introduce weakly supported cues, and unreliable predictions may be reinforced through pseudo supervision. We propose PROTOTAIL, a part-aware GCD framework that couples support-aware local representation learning with reliability-aware pseudo-label selection. ProtoTail uses class-conditioned prototypes over shared latent parts to capture category-specific local evidence, while category-wise gating regularizes effective local-part capacity according to estimated category support. In parallel, separate known- and novel-category selection rules progressively incorporate reliable unlabeled samples using the validation signals available to each group. The two components complement each other: improved local representations support more reliable sample selection, while accepted samples further refine category-specific representations. Under long-tailed training, ProtoTail achieves **75.42%** All accuracy on CUB and **68.29%** on Stanford Cars, with its clearest benefits on low-frequency categories; under balanced training, it reaches **79.58%** and **76.30%**, respectively. These results support support-aware local regularization and progressive pseudo supervision as complementary mechanisms for fine-grained GCD under uneven category support. Our code is available at: `https://anonymous.4open.science/r/ProtoTail-FC6E`

## 1 INTRODUCTION

Generalized Category Discovery (GCD) learns from labeled samples of known categories and an unlabeled pool containing both known and novel categories, with the goal of recognizing the former and discovering the latter (Vaze et al., 2022). SimGCD provides a strong parametric baseline, yet mixed known/novel data still induce prediction bias and noisy pseudo supervision (Wen et al., 2023). Two properties make this problem particularly challenging in realistic settings. First, fine-grained categories often differ only in subtle, localized cues that global image representations may suppress. Second, unlabeled data can be highly imbalanced: under a long-tailed distribution, tail categories provide limited observations, while their frequencies are not directly observed.

We argue that these challenges are coupled through the reliability of category-specific evidence. Fine-grained discrimination benefits from multiple local cues, which part-based GCD methods capture through latent parts shared across categories (Dai et al., 2025; Wang et al., 2026). However, determining how these cues support a particular category requires category-specific statistics, whose reliability depends on the available observations. Under long-tailed training, allowing equally flexible local representations across categories can therefore introduce weakly supported degrees of freedom for tail categories. Importantly, this does not assume that rare categories are intrinsically simpler; rather, their category-specific local evidence is estimated from fewer observations and thus warrants stronger regularization.

The same reliability issue arises in pseudo supervision. Known-category predictions can be assessed against labeled examples, whereas novel-category predictions lack such direct validation; errors may



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

therefore be persistently reinforced once promoted as supervision. Empirically, with the same base model, our pseudo-label selection improves All accuracy on CUB by 6.3% under a balanced split but only 1.2% under a long-tailed split, and can even reduce accuracy when applied alone to a long-tailed base model (Tables 1 and 4). This suggests that pseudo-label selection is not independently sufficient under uneven category support, but depends on the representation from which its reliability signals are derived.

ProtoTail couples support-aware local representation learning with reliability-aware pseudo supervision. Class-conditioned prototypes over shared latent parts provide category-specific local evidence, while category-wise gates select relevant cues and define effective local-part capacity, which is regularized by estimated category support. In parallel, dual-branch pseudo-label selection uses labeled-set reliability for known categories and clustering stability, compactness, and prediction agreement for novel categories. Accepted samples refine the prototypes, while improved local representations support more reliable subsequent selection. Together, these mechanisms provide complementary benefits under long-tailed fine-grained GCD, particularly for low-frequency categories.

**Contributions.** (i) We introduce class-conditioned local representations that combine shared latent parts with category-specific prototypes and gating to capture selective fine-grained evidence; (ii) we propose support-aware regularization of effective local-part capacity, constraining category-specific representations according to the evidence available for estimating them; and (iii) we couple the representation with reliability-aware known/novel pseudo-label selection, and show that the two mechanisms are complementary under long-tailed GCD, with the strongest benefits on low-frequency categories.

## 2 Related Work

### 2.1 Generalized Category Discovery

Novel Category Discovery (NCD) transfers knowledge from labeled known classes to unlabeled novel classes. Deep Transfer Clustering learns a transferable representation before clustering novel categories (Han et al., 2019). AutoNovel transfers ranking structure from labeled data to unlabeled discovery (Han et al., 2020). Neighborhood Contrastive Learning exploits local neighborhood relations to improve the embedding (Zhong et al., 2021), while UNO unifies supervised and unsupervised category prediction (Fini et al., 2021).

GCD relaxes the assumption that all unlabeled samples belong to novel classes. DPN uses decoupled prototypical modeling (An et al., 2023), while TAN and KTN transfer information between known and novel categories (An et al., 2024; Shi et al., 2024). Contrastive Mean-Shift Learning refines category structure through local density modes (Choi et al., 2024). ProtoGCD combines unified prototype learning with adaptive pseudo-labeling (Ma et al., 2025), and AllGCD studies broader use of unlabeled data during training (Cao et al., 2025). ProtoTail is complementary to these approaches: it keeps the underlying GCD representation objective and adds class-conditioned local evidence together with a separate pseudo-label refinement mechanism.

### 2.2 Fine-grained Category Discovery

Fine-grained discovery is particularly sensitive to local evidence because inter-class differences may be confined to small regions. Dual Ranking Statistics combines global and local ranking information (Zhao & Han, 2021), while RAPL aligns discriminative regions for ultra-fine-grained discovery (Liu et al., 2024). Prior-Constrained Association Learning uses known-category priors and cross-instance relations to improve fine-grained semantic grouping (Wang et al., 2025).

APL introduces adaptive part learning with shared part queries for fine-grained GCD (Dai et al., 2025). PartGCD explicitly decomposes category representations into discriminative semantic parts (Wang et al., 2026), while ConGCD constructs primitive-oriented representations through self-deconstruction (Tang et al., 2025). ProtoTail differs by organizing latent parts with a *class-conditioned prototype bank* and category-specific gates, then fusing the resulting local evidence with the global prediction.

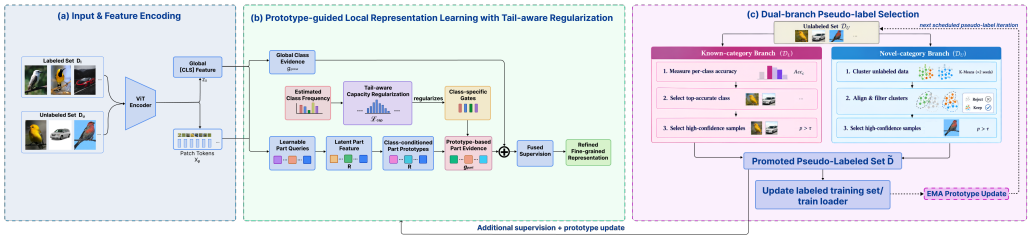



Figure 1: Overview of the proposed **ProtoTail** framework. The encoder extracts a global representation and patch-level features. Prototype-guided local representation learning captures complementary class-conditioned local evidence, while tail-aware regularization controls the effective part capacity across categories. In parallel, dual-branch pseudo-label selection identifies reliable known- and novel-category samples and progressively incorporates them into training.

### 2.3 Long-tailed Learning and Pseudo-labeling

Class imbalance introduces an additional difficulty because the category distribution of the unlabeled set is not directly observed. Bootstrap Your Own Prior estimates class priors for distribution-agnostic novel category discovery (Yang et al., 2023). BaCon formulates distribution-agnostic GCD under long-tailed open-world data (Bai et al., 2023). Long-tailed GCD methods further study imbalanced category frequencies and their effect on representation learning and sample selection (Li et al., 2024; Hoang, 2025). More recent work directly analyzes GCD under long-tailed category distributions (Zhao & Han, 2025).

ProtoTail uses distribution information at the representation-capacity level: the estimated category-frequency profile regularizes the effective local-part capacity assigned to each category. Separately, its pseudo-label module distinguishes known from novel candidates so that the two groups can be evaluated using different evidence.

## 3 Method

### 3.1 Problem Formulation

We consider generalized category discovery (GCD) with a labeled set $\mathcal{D}_{\mathrm{L}}$ and an unlabeled set $\mathcal{D}_{\mathrm{U}}$:

$$\mathcal{D}_{\mathrm{L}} = \{(x_i, y_i)\}_{i=1}^{N_{\mathrm{L}}}, \qquad \mathcal{D}_{\mathrm{U}} = \{x_j\}_{j=1}^{N_{\mathrm{U}}}. \tag{1}$$

The labeled set contains samples from a set of known categories $\mathcal{C}_{\mathrm{K}}$, whereas the unlabeled set may contain samples from both $\mathcal{C}_{\mathrm{K}}$ and a disjoint set of novel categories $\mathcal{C}_{\mathrm{N}}$, with $\mathcal{C}_{\mathrm{K}} \cap \mathcal{C}_{\mathrm{N}} = \varnothing$. The objective is to learn representations that preserve discrimination among known categories while discovering the underlying novel categories without access to ground-truth category identities for unlabeled samples.

### 3.2 Overall Framework

We propose **ProtoTail**, a part-aware GCD framework comprising two complementary components: *prototype-guided local representation learning with tail-aware regularization* and *dual-branch pseudo-label selection*. The first component complements the global image representation with class-conditioned local evidence extracted from patch tokens. The second selects reliable unlabeled samples through separate known- and novel-category branches and progressively incorporates the selected samples into training. The complete workflow is illustrated in Figure 1.

### 3.3 Prototype-guided Local Representation Learning

Fine-grained recognition requires more than extracting local regions: the local representation must make the same latent coordinates comparable across images while allowing their discriminative role

to vary across categories. A local pattern that is decisive for one category may be common or even misleading for another. We therefore construct the local representation in three stages: a shared latent-part basis establishes cross-image correspondence, class-conditioned prototypes assign category meaning to each basis element, and class-specific gates control how much local representation capacity is allocated to each category.

**Shared latent-part basis.** Following latent-part representation learning (Dai et al., 2025), the patch-token sequence of image $x_i$ is summarized by $M$ shared latent slots $R_i = [r_{i,1}, \ldots, r_{i,M}]$, obtained by attention pooling with $M$ learnable queries and $\ell_2$ normalization. The slots are not given semantic part labels. Instead, they define a common coordinate system for local evidence: slot $m$ occupies the same representational role across images, so evidence extracted at that coordinate can be compared across samples. This shared basis alone is class-agnostic, however, and therefore cannot express that the same local cue may have different meanings for different categories.

**Class-conditioned prototype geometry.** For every category $c$ and latent slot $m$, we maintain a prototype $P_{c,m}$ that represents the center of the slot-specific features currently associated with that category. Conceptually, $P_{c,m} \approx \mathbb{E}[r_{i,m} \mid Y_i = c]$ under the assignments available during training. The prototype is therefore not an additional part detector; it is a category-conditioned anchor in the shared local space.

Given a slot feature and its category anchor, we define their compatibility by

$$s_{i,c,m} = \mathrm{sim}(r_{i,m}, P_{c,m}) \,, \tag{2}$$

where $\mathrm{sim}(\cdot,\cdot)$ is the cosine similarity. The resulting score answers a representation-level question: *how consistent is the local evidence carried by slot $m$ with category $c$?* Because the slot extractor is shared but the anchor is category-specific, the same local feature can support different categories by different amounts.

The prototype bank is a running estimate of these category-slot centers, updated by an exponential moving average $P_{c,m} \leftarrow \mathrm{norm}(\mu P_{c,m} + (1-\mu)\bar{r}_{c,m})$, where $\bar{r}_{c,m}$ is the mini-batch mean of slot-$m$ features assigned to category $c$. Labeled samples are assigned by their labels, while unlabeled samples contribute only when their predicted category is sufficiently confident.

**Selective class-conditioned evidence.** Not every latent slot should contribute equally to every category. We associate category $c$ with a gate vector $a_c = [a_{c,1}, \ldots, a_{c,M}] \in [0,1]^M$, parameterized as $a_{c,m} = \sigma(u_{c,m})$ with learnable logits $u_{c,m}$, and define its effective local capacity as

$$\kappa_c = \sum_{m=1}^{M} a_{c,m}. \tag{3}$$

For $\kappa_c > 0$, the normalized contribution of slot $m$ is $\bar{a}_{c,m} = a_{c,m}/\kappa_c$, and the class-conditioned local evidence is

$$g_{i,c}^{\mathrm{part}} = \sum_{m=1}^{M} \bar{a}_{c,m}\, s_{i,c,m}. \tag{4}$$

This decomposition separates two roles that are otherwise conflated. The normalized weights $\bar{a}_{c,m}$ determine *which* local coordinates a category relies on, whereas the gate mass $\kappa_c$ measures *how much* local representational capacity is assigned to that category. Normalization also prevents categories with more active slots from receiving a systematically larger score solely because of scale.

**Why gate regularization is required.** The task objective alone does not determine a useful gate structure. In particular, the normalized evidence in Eq. (4) is invariant to a common positive rescaling of the gates, so relative slot importance can be learned without identifying a meaningful total gate mass. Moreover, a soft gate near $0.5$ does not express a clear decision about whether a slot should participate in the representation.

We therefore regularize the gates as

$$\mathcal{L}_{\mathrm{gate}} = \frac{1}{CM} \sum_{c=1}^{C} \sum_{m=1}^{M} a_{c,m}(1 - a_{c,m}) + \lambda_{\mathrm{b}} \frac{1}{C} \sum_{c=1}^{C} \left[ M_{\min} - \kappa_c \right]_+ . \tag{5}$$

The first term is minimized at the boundary values $a_{c,m} \in \{0,1\}$ and therefore discourages ambiguous intermediate activations. By itself, however, that term also admits the trivial all-zero solution.

The second term removes this degenerate minimum by imposing a minimum amount of active local evidence. Hence, $\mathcal{L}_{\text{gate}}$ regularizes the *structure* of the local representation: it encourages selective, near-discrete slot usage while preserving at least one non-trivial local pathway.

**Tail-aware representation capacity.** Selective gating determines relative slot usage, but fine-grained categories are not supported by equal amounts of training evidence. Using the same effective capacity for every category can therefore impose the same representation complexity on classes with very different statistical support. We instead regularize $\kappa_c$ according to a category-support statistic $\widehat{\pi}_c$ estimated from the current training representation. The method does not depend on a particular estimator: it only requires a non-negative support measure whose ordering reflects the relative amount of evidence available for the categories, without requiring ground-truth novel labels.

Let

$$\rho_c = \phi(\widehat{\pi}_c) \in [0,1], \qquad \frac{\partial \phi}{\partial \widehat{\pi}_c} \geq 0, \tag{6}$$

where $\phi$ is a bounded monotone mapping. We then define the desired category-wise local capacity as

$$\kappa_c^\star = 1 + (M-1)\rho_c \tag{7}$$

and penalize deviations from this target:

$$\mathcal{L}_{\text{cap}} = \frac{1}{C}\sum_{c=1}^{C} \left(\kappa_c - \kappa_c^\star\right)^2. \tag{8}$$

This term regularizes representation *complexity*, rather than deciding which slots a category must use. Evidence-rich categories are permitted a larger effective local capacity, allowing multiple complementary cues to contribute, whereas categories with limited support are biased toward a more compact local representation. Importantly, the constraint acts only on the aggregate gate mass; the task objective and prototype compatibility still determine which specific slots carry the useful evidence. Thus the tail-aware prior controls capacity without prescribing semantic parts.

**Coupling local and global representations.** The local representation is used as complementary evidence rather than as a replacement for the global image representation. For category $c$, the fused logit is $g_{i,c}^{\text{fused}} = g_{i,c}^{\text{global}} + \lambda_{\text{part}} g_{i,c}^{\text{part}}$. Supervising this fused prediction couples prototype-aligned local structure to category discrimination while preserving the global pathway as the primary image-level representation.

### 3.4 Dual-branch Pseudo-label Selection

Pseudo-labeling in GCD faces an asymmetric validation problem: known-category predictions can be checked against labeled examples, whereas novel-category predictions have no direct labeled reference. Applying a single acceptance rule can therefore promote erroneous assignments into persistent supervision. To address this gap, we use separate known and novel selection branches, so each group is validated using the evidence available to it before accepted samples are added to the supervised pool.

Let $\widetilde{\mathcal{D}}_{\text{L}}^{(t)}$ denote the supervised training pool at selection round $t$. The samples accepted by the two branches form a set $\widehat{\mathcal{D}}^{(t)}$, which is added as

$$\widetilde{\mathcal{D}}_{\text{L}}^{(t+1)} = \widetilde{\mathcal{D}}_{\text{L}}^{(t)} \cup \widehat{\mathcal{D}}^{(t)}. \tag{9}$$

Each sample is selected at most once and remains in the supervised pool in subsequent epochs. Because promotion is persistent, both branches control errors at the point of entry: the known branch expands only one category per round, and the novel branch accepts a cluster only when several independent criteria hold simultaneously.

**Known-category branch.** For known categories, labeled supervision allows us to measure the current class-wise prediction accuracy. Let $p_t(c \mid x)$ be the current global prediction probability. On the current supervised pool, which includes previously promoted samples, we define

$$\text{acc}_c^{(t)} = \frac{\sum_{(x_i,y_i)\in\widetilde{\mathcal{D}}_{\text{L}}^{(t)}} \mathbb{I}[y_i = c]\,\mathbb{I}[\arg\max_k p_t(k \mid x_i) = c]}{\sum_{(x_i,y_i)\in\widetilde{\mathcal{D}}_{\text{L}}^{(t)}} \mathbb{I}[y_i = c] + \epsilon}, \qquad c_t^\star = \arg\max_{c\in\mathcal{C}_{\text{K}}} \text{acc}_c^{(t)}. \tag{10}$$

Let $\mathcal{S}_{<t}$ denote the unlabeled samples promoted in earlier rounds. Among the remaining unlabeled samples predicted as $c_t^\star$, we retain the most confident ones:

$$\widehat{\mathcal{D}}_{\mathrm{K}}^{(t)} = \mathrm{TopK}_{x \in \mathcal{D}_{\mathrm{U}} \setminus \mathcal{S}_{<t}} \left( p_t(c_t^\star \mid x) \;\middle|\; \arg\max_k p_t(k \mid x) = c_t^\star \right), \tag{11}$$

where the number of retained candidates is limited by the prescribed selection ratio and per-round cap. Thus, the known branch first selects the known category with the highest labeled-set accuracy and then promotes only its highest-confidence unlabeled predictions. This yields an easy-first curriculum: categories whose decision regions are already best validated by labeled data are expanded first, while less reliable categories are deferred to later rounds, when the representation has improved.

**Novel-category branch.** For novel categories, no labeled examples are available to directly validate a prediction. We therefore assess novel candidates using both representation structure and prediction consistency.

At each selection round, we partition the unlabeled representations into $|\mathcal{C}_{\mathrm{K}}| + |\mathcal{C}_{\mathrm{N}}|$ clusters, so that each cluster can be read as a candidate category, and repeat the partition under independent initializations to assess its stability. Labeled known samples are assigned to the resulting centroids, and Hungarian matching (Kuhn, 1955) is used to identify clusters associated with known categories. These known-aligned clusters are excluded, leaving the remaining clusters as novel candidates.

For each candidate cluster, we first measure stability across clustering initializations using Jaccard overlap (Jaccard, 1901) between corresponding membership sets. A cluster is retained only if its membership is sufficiently stable, its mean Silhouette score (Rousseeuw, 1987) is at least the median over all clusters, and a sufficient fraction of its members agree on the same novel output dimension, which then serves as their pseudo label. We additionally require a minimum cluster size. The conditions are combined conjunctively because each targets a distinct failure mode: stability rejects clusters that are artifacts of a particular initialization, compactness rejects poorly separated clusters, prediction agreement rejects clusters the classifier does not recognize consistently, the minimum size ensures these statistics are estimated from enough samples, and the temporal checks below prevent global degradation or remapping of an established novel output dimension.

**Temporal consistency.** The novel branch also checks consistency across successive selection rounds. A Davies–Bouldin index (Davies & Bouldin, 1979) circuit breaker suppresses a round when the global cluster structure deteriorates substantially relative to the previous accepted round. In addition, a no-remap constraint rejects large changes in the sample population associated with an already established novel output dimension. These checks reduce the risk of turning transient cluster assignments into persistent pseudo supervision.

All selection criteria use only training representations, model predictions, and the labels available for known-category training samples. Ground-truth novel-category identities are not used for pseudo-label selection.

**Interaction with prototype memory.** Prototype-memory updates and pseudo-label selection operate at different reliability levels. Confidence-filtered provisional assignments can update the prototype memory before a sample satisfies the stricter selection criteria above. Once a sample is selected and added to the supervised pool, it contributes to subsequent supervised learning and continues to reinforce the corresponding prototype. This interaction allows local category evidence and pseudo supervision to improve progressively over training.

### 3.5 Training Objective

ProtoTail retains the original supervised and unsupervised objective $\mathcal{L}_{\mathrm{base}}$ of the underlying GCD model, computed on the global logits. The fused prediction provides auxiliary supervision through $\mathcal{L}_{\mathrm{fused}}$, the cross-entropy of $g^{\mathrm{fused}}/\tau_c$ on labeled samples, while $\mathcal{L}_{\mathrm{gate}}$ encourages selective and non-degenerate part activation and $\mathcal{L}_{\mathrm{cap}}$ constrains the aggregate part capacity according to the estimated category frequency. Pseudo-labeled samples do not introduce a separate loss; once selected, they are added to the supervised pool and participate in the same supervised objectives as the originally labeled samples. The prototype bank is updated separately through EMA. The complete training procedure is summarized in Algorithm 1.

**Algorithm 1** ProtoTail training

Initialize $\widetilde{\mathcal{D}}_{\mathrm{L}} \leftarrow \mathcal{D}_{\mathrm{L}}$
**for** each training epoch **do**
  Compute $g^{\mathrm{global}}$, $g^{\mathrm{part}}$, and $g^{\mathrm{fused}}$
  Optimize $\mathcal{L}_{\mathrm{base}} + \lambda_{\mathrm{f}}\mathcal{L}_{\mathrm{fused}} + \lambda_{\mathrm{g}}\mathcal{L}_{\mathrm{gate}} + \lambda_{\mathrm{cap}}\mathcal{L}_{\mathrm{cap}}$
  Update the prototype bank by EMA
  **if** pseudo-label selection is scheduled **then**
    Select $\widehat{\mathcal{D}}_{\mathrm{K}}$ and $\widehat{\mathcal{D}}_{\mathrm{N}}$
    $\widetilde{\mathcal{D}}_{\mathrm{L}} \leftarrow \mathcal{D}_{\mathrm{L}} \cup \widehat{\mathcal{D}}_{\mathrm{K}} \cup \widehat{\mathcal{D}}_{\mathrm{N}}$
  **end if**
**end for**

# 4 Experiments

## 4.1 Experimental Setup

**Datasets.** We evaluate on three fine-grained image classification benchmarks: **CUB-200-2011 (CUB)** (Wah et al., 2011), **Stanford Cars** (Krause et al., 2013), and **FGVC-Aircraft** (Maji et al., 2013). CUB contains 200 bird categories, which are divided into 100 known and 100 novel categories. Stanford Cars contains 196 vehicle categories, divided into 98 known and 98 novel categories, and FGVC-Aircraft contains 100 aircraft categories, divided into 50 known and 50 novel categories. Component and long-tailed analyses are conducted on CUB and Stanford Cars, while FGVC-Aircraft is additionally used for comparison with prior methods in the balanced setting. Following the GCD setting, labeled training samples are available only for known categories, whereas the unlabeled training set contains samples from both known and novel categories.

**Balanced and long-tailed settings.** Experiments are conducted under both balanced and long-tailed training distributions. For the balanced setting, we follow the SimGCD protocol (Wen et al., 2023) and use the corresponding balanced training split. For the long-tailed setting, we construct the training split using the standard exponential long-tail sampling scheme (Cui et al., 2019) with an imbalance ratio of 10, defined as the ratio between the largest and smallest class sizes, while preserving the same known–novel category partition. We use the same precomputed data splits across compared methods so that differences are not caused by changes in class identities or sample distributions. The evaluation set is kept unchanged between the balanced and long-tailed settings.

**Evaluation metrics.** We report clustering accuracy on **All**, **Old**, and **New** categories. **All** measures accuracy over the complete category set, while **Old** and **New** report performance separately on known and novel categories. Predicted category assignments are aligned with ground-truth labels using Hungarian matching before computing accuracy.

**Implementation details.** All methods use DINOv2 ViT-B/14 (Oquab et al., 2024) as the visual backbone and fine-tune the final transformer block; results of prior methods in Table 3 are those reported in the literature with the same backbone. Models are trained for 200 epochs with a batch size of 128 using SGD and cosine learning-rate decay (Loshchilov & Hutter, 2017). The initial learning rate is $0.1$, weight decay is $5 \times 10^{-5}$, and the supervised-loss weight is $0.35$. Images are processed at $224 \times 224$ resolution using ImageNet-style augmentation; further details are given in Appendix B.

## 4.2 Results

**Long-tailed benchmarks.** Table 1 evaluates the components of ProtoTail with BaCon under the long-tailed setting. On CUB, the full model raises All accuracy from 73.72% to 75.42%, mainly through an improvement on Old classes (72.68% to 75.58%) while preserving strong New-class performance. On Stanford Cars, All, Old, and New accuracy increase from 66.86/61.67/71.99% to 68.29/63.59/72.94%, respectively. The full model thus improves all three metrics on both datasets. In contrast, dual selection alone slightly lowers All accuracy (73.72→73.40% on CUB and 66.86→65.70% on Stanford Cars), indicating that its benefit under skewed frequencies depends on the improved representation provided by the prototype branch.

Table 1: Component analysis of ProtoTail under long-tailed (BaCon) and balanced (SimGCD) fine-grained GCD settings. Blue shading marks All accuracy; bold and underlined values denote the best and second-best results within each setting, respectively.

| Base Method | Mode | CUB | | | Stanford Cars | | |
|---|---|---|---|---|---|---|---|
| | | All | Old | New | All | Old | New |
| **BaCon** *(LT)* | Baseline | 73.72 | 72.68 | 74.75 | 66.86 | 61.67 | 71.99 |
| | Dual selection | 73.40 | 72.65 | 74.15 | 65.70 | 62.10 | 69.30 |
| | Prototype (+Tail-aware) | 74.23 | 72.66 | **75.75** | 68.10 | 63.50 | 72.20 |
| | Full (+Tail-aware) | **75.42** | **75.58** | 75.26 | **68.29** | **63.59** | **72.94** |
| **SimGCD** *(balanced)* | Baseline | 71.50 | 78.10 | 68.30 | 71.50 | 81.90 | 66.60 |
| | Dual selection | 77.75 | **85.70** | 69.88 | 72.56 | **88.56** | 64.83 |
| | Prototype (+Tail-aware) | 69.53 | 77.59 | 65.50 | 74.87 | 87.66 | 68.69 |
| | Full (+Tail-aware) | **79.58** | 83.47 | **75.68** | **76.30** | 84.16 | **72.57** |

Table 2: Component and frequency-group analysis under the BaCon long-tailed setting with imbalance ratio 10. Following Bai et al. (2023), known (K-) and novel (U-) categories are each ranked by their number of training samples and evenly divided into Many/Med/Few groups. Blue shading marks All and Few-group accuracy; bold and underlined values denote the best and second-best results within each dataset block.

| Dataset | Configuration | Accuracy | | | Known frequency | | | Unlabeled frequency | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | All | Old | New | Many | Med | Few | Many | Med | Few |
| **CUB** | Baseline | 73.72 | 72.68 | 74.75 | 73.67 | **73.57** | 70.38 | **82.94** | 73.41 | 68.78 |
| | Dual selection | 73.40 | 72.65 | 74.15 | 74.32 | 72.59 | 70.86 | 80.73 | 72.87 | 69.63 |
| | Prototype (w/o Tail-aware) | 74.30 | 74.70 | 74.00 | 75.70 | 73.40 | 68.60 | 79.20 | **74.40** | 68.60 |
| | Prototype (+Tail-aware) | 74.23 | 72.66 | **75.75** | 73.29 | 72.99 | 71.53 | 80.97 | 71.83 | **76.13** |
| | Full (w/o Tail-aware) | 73.20 | 72.30 | 74.10 | 78.70 | 69.70 | 68.90 | 75.50 | 73.10 | 68.90 |
| | Full (+Tail-aware) | **75.42** | **75.58** | 75.26 | **78.84** | 72.94 | **75.52** | 82.22 | 73.90 | 70.48 |
| **Stanford Cars** | Baseline | 66.86 | 61.67 | 71.99 | 70.31 | 61.13 | 53.81 | **85.72** | **76.73** | 52.94 |
| | Dual selection | 65.70 | 62.10 | 69.30 | 68.30 | 61.90 | 53.20 | 81.10 | 66.80 | 57.40 |
| | Prototype (w/o Tail-aware) | 64.00 | 59.50 | 68.50 | 67.50 | 62.30 | 48.60 | 82.30 | 72.10 | 50.70 |
| | Prototype (+Tail-aware) | 68.10 | 63.50 | 72.20 | **75.90** | **63.50** | 53.90 | 84.20 | 72.20 | 60.50 |
| | Full (w/o Tail-aware) | 67.40 | 62.80 | 72.00 | 75.50 | 58.80 | 54.80 | 80.80 | 73.50 | 52.00 |
| | Full (+Tail-aware) | **68.29** | **63.59** | **72.94** | 74.63 | 57.59 | **59.38** | 80.74 | 76.27 | **61.36** |

Table 2 decomposes the long-tailed results by component and by category frequency. The two w/o Tail-aware rows isolate the tail-aware capacity regularizer. Adding it to the full model consistently improves the low-frequency groups: K-Few/U-Few increase from 68.90/68.90% to 75.52/70.48% on CUB and from 54.80/52.00% to 59.38/61.36% on Stanford Cars. The same trend holds when it is added to the prototype branch alone, where U-Few improves from 68.60% to 76.13% on CUB and from 50.70% to 60.50% on Stanford Cars. These gains come at a moderate cost on some medium-frequency groups, while the full model attains the best All accuracy on both datasets.

**Comparison with prior GCD methods.** Table 3 compares ProtoTail with representative GCD methods on balanced fine-grained benchmarks. ProtoTail obtains the best All (79.58%) and Old (83.47%) accuracy on CUB and the best New accuracy on Stanford Cars (72.57%). The recent SEAL attains a slightly higher average All accuracy (76.3% vs. 76.1%), owing to Stanford Cars and FGVC-Aircraft; on FGVC-Aircraft, ProtoTail ranks second in All and New accuracy.

**Component analysis in the balanced setting.** The SimGCD block of Table 1 decomposes the balanced results. On CUB, the full model improves All, Old, and New accuracy from 71.50/78.10/68.30% to 79.58/83.47/75.68%, respectively. On Stanford Cars, it improves All accuracy from 71.50% to 76.30%, with gains on both Old (81.90% to 84.16%) and New (66.60% to 72.57%) classes. Dual selection alone raises All accuracy on CUB by 6.25 points here, in contrast to its negative effect under the long-tailed setting, consistent with pseudo-label reliability depending on category support. The weaker Prototype (+Tail-aware) result on CUB shows that the gain is not obtained by simply adding a prototype branch; the complete formulation is required for the strongest overall performance.

Table 3: Comparison with state-of-the-art GCD methods on balanced benchmarks. Blue shading marks All accuracy, the green row highlights our method, and bold and underlined values denote the best and second-best reported results in each column, respectively.

| Method | Venue | CUB | | | Stanford Cars | | | FGVC-Aircraft | | | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | All | Old | New | All | Old | New | All | Old | New | All |
| GCD (Vaze et al., 2022) | CVPR22 | 71.90 | 71.20 | 72.30 | 65.70 | 67.80 | 64.70 | 55.40 | 47.90 | 59.20 | 64.3 |
| CiPR (Hao et al., 2024) | TMLR24 | <u>78.30</u> | 73.40 | **80.80** | 66.70 | 77.00 | 61.80 | 59.20 | 65.00 | 56.30 | 68.1 |
| SimGCD (Wen et al., 2023) | ICCV23 | 71.50 | 78.10 | 68.30 | 71.50 | 81.90 | 66.60 | 63.90 | 69.90 | 60.90 | 69.0 |
| $\mu$GCD (Vaze et al., 2023) | NeurIPS23 | 74.00 | 75.90 | 73.10 | 76.10 | **91.00** | 68.90 | 66.30 | 68.70 | 65.10 | 72.1 |
| DebGCD (Liu & Han, 2025) | ICLR25 | 77.50 | 80.80 | 75.80 | 75.40 | 87.70 | 69.50 | 71.90 | <u>76.00</u> | 69.80 | 74.9 |
| ProtoGCD+ (Ma et al., 2025) | TPAMI25 | 75.70 | <u>81.50</u> | 72.90 | <u>77.60</u> | <u>90.50</u> | 71.50 | 71.10 | **76.30** | 68.50 | 74.8 |
| SEAL (He et al., 2025) | NeurIPS25 | 76.70 | 78.30 | <u>75.90</u> | **77.70** | 88.70 | <u>72.40</u> | **74.60** | 73.20 | **75.30** | **76.3** |
| **Ours** | - | **79.58** | **83.47** | 75.68 | 76.30 | 84.16 | **72.57** | <u>72.49</u> | 69.15 | <u>74.15</u> | <u>76.1</u> |

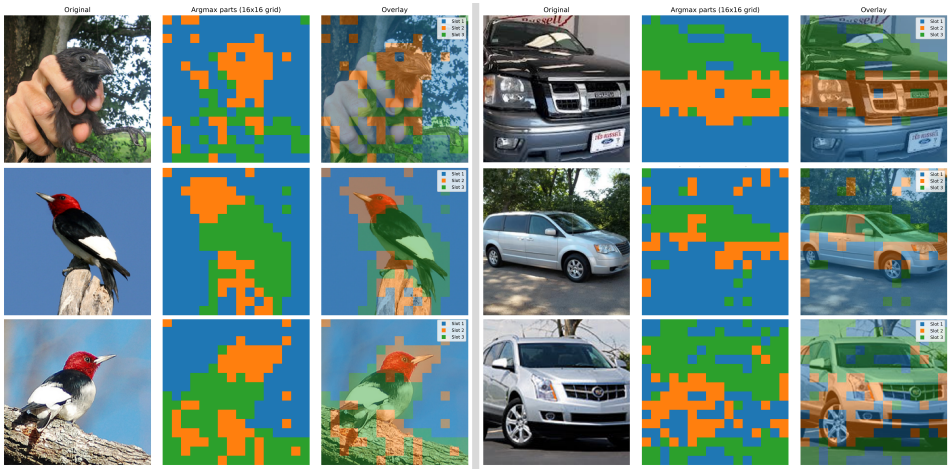



Figure 2: Spatial assignments of the learned latent slots on representative CUB (left) and Stanford Cars (right) examples. For each image, the columns show the original image, the patch-wise argmax assignment on the $16 \times 16$ grid, and the overlay. Blue, orange, and green denote Slots 1–3, respectively; no semantic part labels are provided during training.

**Spatial specialization of latent parts.** Figure 2 visualizes the spatial specialization learned by the prototype-guided part representations. Without any part annotations, different latent slots consistently capture complementary local regions: for birds, the orange slot focuses mainly on the head and upper body and largely avoids the human hand in the first example; for cars, the green slot tends to capture the windshield and upper body, while the orange slot focuses on frontal structures such as the grille and headlamps. This suggests that the learned latent slots capture localized fine-grained cues and can de-emphasize irrelevant regions such as the occluding hand.

## 5 Conclusion

We presented **ProtoTail**, a part-aware framework for fine-grained generalized category discovery under both balanced and long-tailed training distributions. ProtoTail combines prototype-guided local representation learning with tail-aware capacity regularization and dual-branch pseudo-label selection. The local branch complements global image representations with class-conditioned fine-grained evidence, while the pseudo-labeling module separately handles known and novel categories according to the supervision available to each group. Experiments on CUB and Stanford Cars show improvements under both balanced and long-tailed training distributions, with the full model reaching 79.58% and 76.30% All accuracy under the balanced setting, and 75.42% and 68.29% under the long-tailed setting, respectively. These results indicate that class-conditioned local representations and progressive pseudo supervision are complementary for fine-grained GCD. A remaining limitation is the reliance on periodic clustering and fixed selection thresholds; future work may investigate adaptive selection and frequency estimation under more diverse category distributions.

### AI USE STATEMENT

Generative AI tools were used during manuscript preparation for language refinement, LaTeX formatting, programming/debugging support, and assistance in methodology/experimental-design discussions and interpretation of experimental diagnostics. They were not used to fabricate or directly generate experimental measurements. All methodological choices, implementation decisions, experiments, reported results, analyses, and conclusions were reviewed and verified by the authors, who take full responsibility for the final manuscript.

### ETHICS STATEMENT

This work uses established public fine-grained image classification benchmarks and does not introduce new human-subject data or collect private information. As with other representation-learning and pseudo-labeling methods, the proposed approach may inherit biases present in the training data, and erroneous pseudo-label assignments may reinforce existing model errors. Our experiments focus on evaluating the proposed methodology under controlled balanced and long-tailed category distributions. The method is intended for research on generalized category discovery, and its behavior should be carefully evaluated before deployment in applications involving sensitive or high-stakes decisions.

### REPRODUCIBILITY STATEMENT

We provide the datasets, known–novel category partitions, evaluation protocol, backbone, training configuration, and major hyperparameters in Section 4.1, with the complete loss weights, prototype-update schedule, and pseudo-label selection criteria in Appendix B. This includes the balanced and long-tailed data construction, optimization settings, latent-part configuration, pseudo-label selection schedules, and selection thresholds. The evaluation procedure, including Hungarian matching for All, Old, and New clustering accuracy, is also specified. The same precomputed data splits are used across compared methods to avoid differences caused by category identities or sample distributions. Additional experimental details, results, ablations, and visualizations are provided in the appendix.

## REFERENCES

Wenbin An, Feng Tian, Qinghua Zheng, Wei Ding, Qianying Wang, and Ping Chen. Generalized Category Discovery with Decoupled Prototypical Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 12527–12535, 2023.

Wenbin An, Feng Tian, Wenkai Shi, Yan Chen, Yaqiang Wu, Qianying Wang, and Ping Chen. Transfer and Alignment Network for Generalized Category Discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 10856–10864, 2024.

Jianhong Bai, Zuozhu Liu, Hualiang Wang, Ruizhe Chen, Lianrui Mu, Xiaomeng Li, Joey Tianyi Zhou, Yang Feng, Jian Wu, and Haoji Hu. Towards Distribution-Agnostic Generalized Category Discovery. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Xinzi Cao, Ke Chen, Feidiao Yang, Xiawu Zheng, Yonghong Tian, and Yutong Lu. AllGCD: Leveraging All Unlabeled Data for Generalized Category Discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3293–3303, 2025.

Sua Choi, Dahyun Kang, and Minsu Cho. Contrastive Mean-Shift Learning for Generalized Category Discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-Balanced Loss Based on Effective Number of Samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9268–9277, 2019.

Qiyuan Dai, Hanzhuo Huang, Yu Wu, and Sibei Yang. Adaptive Part Learning for Fine-Grained Generalized Category Discovery: A Plug-and-Play Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25444–25453, 2025.

David L. Davies and Donald W. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979.

Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A Unified Objective for Novel Class Discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9284–9292, 2021.

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 (ICCV)*, 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. In *International Conference on Learning Representations (ICLR)*, 2020.

Shaozhe Hao, Kai Han, and Kwan-Yee K. Wong. CiPR: An Efficient Framework with Cross-instance Positive Relations for Generalized Category Discovery. *Transactions on Machine Learning Research*, 2024.

Zhenqi He, Yuanpei Liu, and Kai Han. SEAL: Semantic-Aware Hierarchical Learning for Generalized Category Discovery. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Cuong Manh Hoang. Long-Tailed Learning for Generalized Category Discovery. *Neurocomputing*, 650:130745, 2025.

Paul Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37(142):547–579, 1901.

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 (ICCV) Workshops*, pp. 554–561, 2013.

Harold W. Kuhn. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

Ziyun Li, Christoph Meinel, and Haojin Yang. Generalized Categories Discovery for Long-tailed Recognition. *arXiv preprint arXiv:2401.05352*, 2024.

Yu Liu, Yaqi Cai, Qi Jia, Binglin Qiu, Weimin Wang, and Nan Pu. Novel Class Discovery for Ultra-Fine-Grained Visual Categorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17679–17688, 2024.

Yuanpei Liu and Kai Han. DebGCD: Debiased Learning with Distribution Guidance for Generalized Category Discovery. In *International Conference on Learning Representations (ICLR)*, 2025.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations*, 2017.

Shijie Ma, Fei Zhu, Xu-Yao Zhang, and Cheng-Lin Liu. ProtoGCD: Unified and Unbiased Prototype Learning for Generalized Category Discovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):6022–6038, 2025.

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 2024.

Peter J. Rousseeuw. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

Wenkai Shi, Wenbin An, Feng Tian, Yan Chen, Yaqiang Wu, Qianying Wang, and Ping Chen. A Unified Knowledge Transfer Network for Generalized Category Discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18961–18969, 2024.

Luyao Tang, Kunze Huang, Chaoqi Chen, Yuxuan Yuan, Chenxin Li, Xiaotong Tu, Xinghao Ding, and Yue Huang. Dissecting Generalized Category Discovery: Multiplex Consensus under Self-Deconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 297–307, 2025.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized Category Discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7492–7501, 2022.

Sagar Vaze, Andrea Vedaldi, and Andrew Zisserman. No Representation Rules Them All in Category Discovery. In *Advances in Neural Information Processing Systems*, volume 36, pp. 19962–19989, 2023.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Enguang Wang, Zhimao Peng, Zhengyuan Xie, Haori Lu, Fei Yang, and Xialei Liu. Learning Part Knowledge to Facilitate Category Understanding for Fine-Grained Generalized Category Discovery. *IEEE Transactions on Multimedia*, 2026.

Menglin Wang, Zhun Zhong, and Xiaojin Gong. Prior-Constrained Association Learning for Fine-Grained Generalized Category Discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21162–21170, 2025.

Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric Classification for Generalized Category Discovery: A Baseline Study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16590–16600, 2023.

Muli Yang, Liancheng Wang, Cheng Deng, and Hanwang Zhang. Bootstrap Your Own Prior: Towards Distribution-Agnostic Novel Class Discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3459–3468, 2023.

Bingchen Zhao and Kai Han. Novel Visual Category Discovery with Dual Ranking Statistics and Mutual Knowledge Distillation. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Bingchen Zhao and Kai Han. Generalized Category Discovery under the Long-Tailed Distribution. *arXiv preprint arXiv:2506.12515*, 2025.

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 (CVPR)*, pp. 10867–10875, 2021.

# A Additional Results

## A.1 Long-Tailed SimGCD

On Stanford Cars, prototype-guided representation learning increases All accuracy from 70.79% to 72.39%, while the full model further reaches 73.80% All and 57.50% New accuracy. The effect is more pronounced on CUB, where the Prototype (+Tail-aware) configuration raises All accuracy from 60.55% to 65.87% and New accuracy from 47.60% to 56.01%. Combining the local representation with dual-branch pseudo-label selection further improves the full model to 68.06% All and 76.09% Old accuracy, supporting the complementary roles of representation structuring and progressive supervision in ProtoTail.

Table 4: Component results under the SimGCD setting with imbalance ratio 10.

| Dataset | Configuration | All | Old | New |
|---|---|---|---|---|
| **CUB** | Baseline | 60.55 | 73.44 | 47.60 |
| | Dual selection | 61.74 | 74.58 | 48.84 |
| | Prototype (+Tail-aware) | 65.87 | 72.30 | 56.01 |
| | Full (+Tail-aware) | **68.06** | **76.09** | **56.14** |
| **Stanford Cars** | Baseline | 70.79 | 87.65 | 54.03 |
| | Dual selection | 70.66 | **88.92** | 52.52 |
| | Prototype (+Tail-aware) | 72.39 | 84.85 | 54.58 |
| | Full (+Tail-aware) | **73.80** | 85.20 | **57.50** |

### A.2 Additional Qualitative Visualizations

To complement Figure 2 in the main text, we provide additional latent-slot visualizations on CUB and Stanford Cars. These examples show that the learned slots produce consistent coarse spatial decompositions across diverse images without any part annotations. While the slots are not assigned fixed semantic labels, they repeatedly separate complementary regions such as foreground structure, object-centered detail, and surrounding context. This behavior supports the interpretation that ProtoTail learns reusable latent coordinates for localized evidence rather than predefined semantic parts.

Across both datasets, the slot assignments are instance-adaptive yet visually structured. Some slots tend to cover broader silhouette or contextual regions, whereas others focus more strongly on localized object regions. On CUB, the assignments frequently separate the bird body from surrounding background and other spatial regions of the image. On Stanford Cars, the slots similarly partition complementary vehicle regions across different instances and viewpoints. Importantly, these patterns emerge without semantic part annotations and should therefore be interpreted as learned latent spatial specialization rather than fixed semantic part identities.

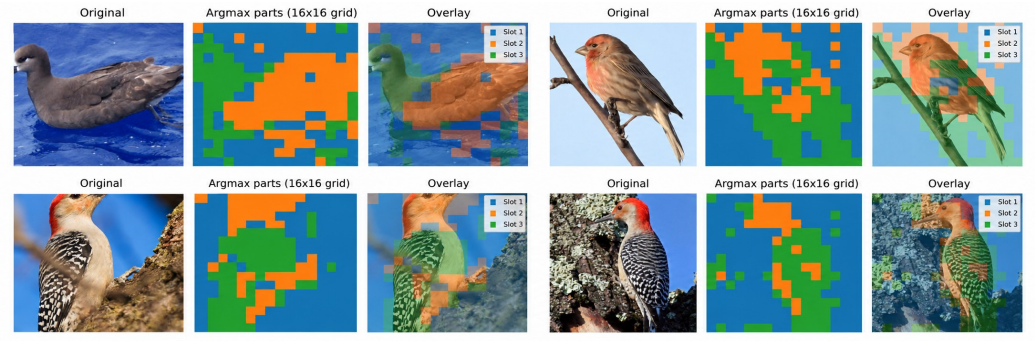
This qualitative behavior is consistent with the role of the local branch in ProtoTail: the shared latent slots provide a common coordinate system for localized evidence, while their category-specific relevance is determined by the prototype-guided representation rather than by manually predefined parts. The additional examples therefore complement the main-text visualization by showing that the learned slot structure persists across a broader range of images from both fine-grained benchmarks.

### A.3 Representation-Space Ablation and Prototype Alignment

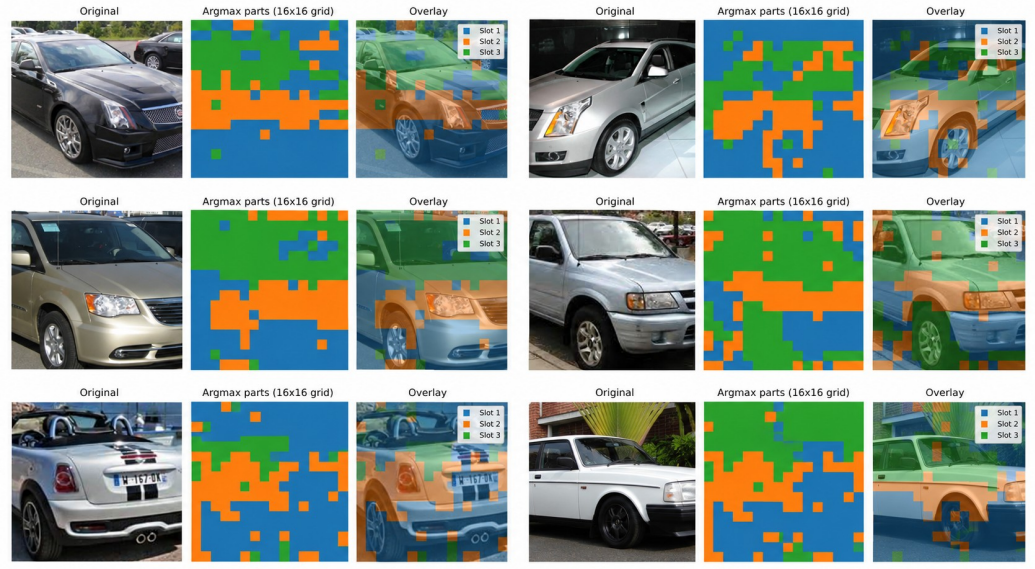
To complement the quantitative ablation results, we visualize the embedding geometry before and after introducing the full ProtoTail formulation in Figure 4. For each dataset, t-SNE is applied to the top-20 selected classes. Circles and triangles denote samples from known and novel categories, respectively. In the ProtoTail visualizations, stars denote the learned class-conditioned prototypes.

Compared with the corresponding baselines, ProtoTail produces a more explicitly structured class-conditioned embedding in both datasets. Most sample groups remain compact and well separated, while the learned prototypes are generally located within or close to their associated sample regions. This behavior is especially visible in the CUB visualization, where the prototype markers closely track the centers of multiple compact clusters, and in Stanford Cars, where prototypes similarly follow the geometry of both known and novel category groups.

The comparison provides qualitative support for the role of the prototype-guided local branch. Rather than introducing a separate geometry disconnected from the underlying representation, the prototype bank follows the organization of the sample embeddings and provides category-conditioned anchors within that space. Together with the quantitative gains reported in Table 2, these visualizations support the view that the prototype and tail-aware components improve representation structure while preserving separation between discovered categories.

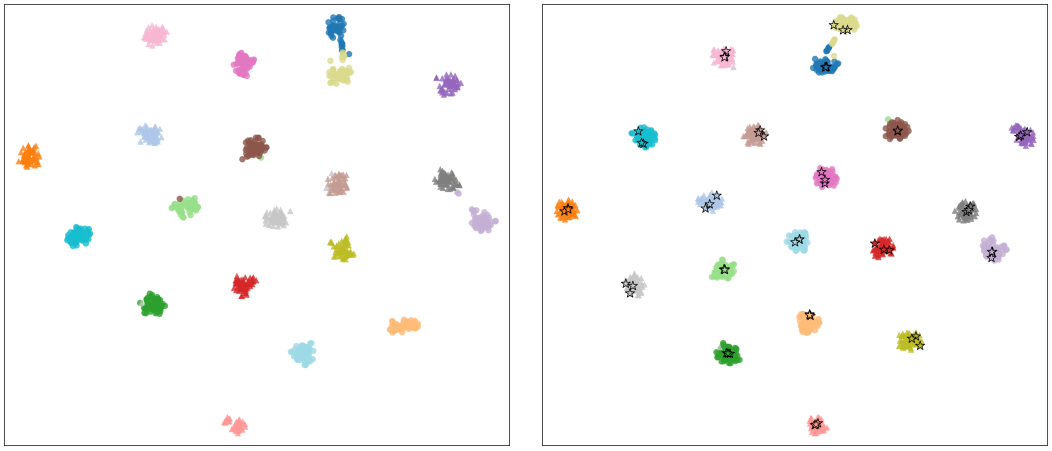



**(a) CUB**

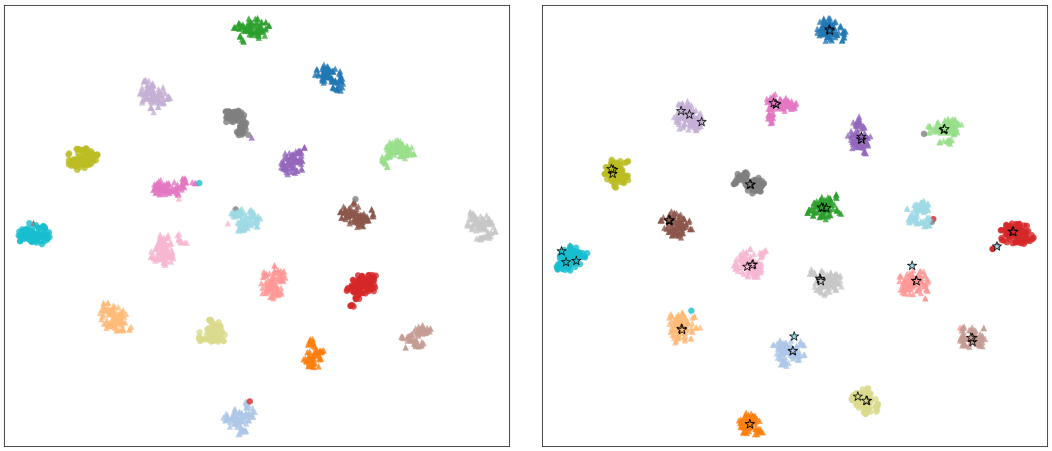



**(b) Stanford Cars**

Figure 3: Additional latent-slot visualizations on CUB and Stanford Cars. For each example, the columns show the original image, the patch-wise argmax slot assignment on the $16 \times 16$ grid, and the corresponding overlay. Blue, orange, and green denote Slots 1–3, respectively.



**(a) CUB: baseline** **(b) CUB: ProtoTail**



**(c) Stanford Cars: baseline** **(d) Stanford Cars: ProtoTail**

Figure 4: Representation-space comparison between the baseline and ProtoTail on CUB and Stanford Cars. Each panel visualizes the top-20 selected classes using t-SNE. Circles and triangles denote known- and novel-category samples, respectively; stars in the ProtoTail panels denote learned class-conditioned prototypes.

### A.4 Additional CIFAR100-LT Ablation

To further examine whether the tail-aware capacity mechanism generalizes beyond the two fine-grained benchmarks used in the main component analysis, we perform an additional ablation on CIFAR100-LT using the BaCon setting.

Table 5: Additional component and frequency-group analysis on CIFAR100-LT. Blue shading marks All and Few-group accuracy.

| Configuration | Accuracy | | | Known frequency | | | Unlabeled frequency | | |
|---|---|---|---|---|---|---|---|---|---|
| | **All** | **Old** | **New** | **Many** | **Med** | **Few** | **Many** | **Med** | **Few** |
| Baseline | 77.50 | 76.90 | 79.80 | **86.70** | 77.60 | 66.50 | 86.20 | 80.90 | 72.20 |
| Dual selection | 80.70 | 79.70 | 84.50 | **86.70** | 84.00 | 68.10 | 91.50 | **89.80** | 70.30 |
| Prototype (w/o Tail-aware) | 79.40 | 78.00 | 85.10 | 85.60 | 82.00 | 63.20 | 87.70 | 87.20 | 79.70 |
| Prototype (+Tail-aware) | 81.10 | 80.30 | 84.10 | 86.50 | 78.80 | 70.70 | 75.80 | 89.10 | **86.20** |
| Full (w/o Tail-aware) | 82.00 | **82.30** | 80.70 | 85.90 | **84.00** | 66.80 | 92.00 | 87.60 | 60.20 |
| Full (+Tail-aware) | **83.80** | 79.10 | **87.80** | 86.20 | 78.60 | **72.50** | **92.70** | 88.20 | 82.20 |

The CIFAR100-LT results provide an additional check of the tail-aware mechanism outside CUB and Stanford Cars. Comparing the two full configurations, adding tail-aware capacity regularization increases All accuracy from 82.00% to 83.80%, while K-Few improves from 66.80% to 72.50% and U-Few increases substantially from 60.20% to 82.20%. The same low-frequency trend is visible in the prototype branch alone: K-Few rises from 63.20% to 70.70%, while U-Few improves from 79.70% to 86.20%.

These results reinforce the main long-tailed analysis in Table 2: tail-aware capacity allocation provides its clearest benefit on categories with limited support. Although some medium- and head-frequency groups do not improve uniformly, the full model obtains the highest overall accuracy and markedly strengthens the rare-category regime, which is the behavior targeted by the proposed capacity regularizer.

## B Implementation Details

**Local representation.** The $M = 3$ latent slots are computed by scaled dot-product attention between orthogonally initialized learnable queries and the patch tokens of the last transformer block, with attention temperature $0.07$. Prototypes are orthogonally initialized, $\ell_2$-normalized, and updated only by EMA with momentum $\mu = 0.99$ from epoch 30 onward. Unlabeled samples update the prototype of their predicted category only when the maximum softmax probability of the logits scaled by $\tau_c$ exceeds $0.7$; to avoid self-reinforcing drift, these unlabeled updates are restricted to epochs 30–60. Gate logits $u_{c,m}$ are optimized with SGD (learning rate $0.01$) and the slot queries with learning rate $0.05$, both with cosine decay.

**Loss weights and fusion.** We use $\lambda_{\mathrm{g}} = \lambda_{\mathrm{cap}} = \lambda_{\mathrm{b}} = 0.05$ and $M_{\min} = 1$. The fused cross-entropy weight is $\lambda_{\mathrm{f}} = 1$ with BaCon and $0.5$ with SimGCD. The local evidence is fused with weight $\lambda_{\mathrm{part}} = 0.5$, and the classification temperature is $\tau_c = 0.1$.

**Support statistic and capacity target.** The support statistic $\widehat{\pi}_c$ follows the distribution estimator of BaCon (Bai et al., 2023): category sizes are read from a clustering of the training representations, in which known categories receive the sizes of their aligned clusters and the remaining cluster sizes are assigned to novel categories. The estimate is refreshed every 10 epochs. The bounded monotone mapping $\phi$ in Eq. (6) is a log-scale min–max normalization,

$$\rho_c = \frac{\log \widehat{\pi}_c - \log \widehat{\pi}_{\min}}{\log \widehat{\pi}_{\max} - \log \widehat{\pi}_{\min}},$$

and the resulting target capacity $\kappa_c^\star$ is clipped to $[1, M]$.

**Known-category branch.** Class-wise accuracy is computed on the current supervised pool. Among unused unlabeled samples predicted as $c_t^\star$, the top $80\%$ by confidence are promoted, up to 500 samples per round, for at most 20 rounds.

**Novel-category branch.** Clustering uses the backbone `[CLS]` representation with one cluster per category. Two partitions are obtained with different random seeds (10 restarts each, fitted on at most

15,000 unlabeled samples and then assigned to all samples). Known-aligned clusters are identified by assigning labeled samples to their nearest centroid and applying Hungarian matching. A candidate cluster is accepted if it contains at least 10 samples, its best Jaccard overlap with a cluster of the second partition is at least $0.6$, its mean Silhouette score (computed on a random subset of 2,000 samples) is at least the median over all clusters, and at least $70\%$ of its members share the same majority novel output dimension. A round is skipped if the Davies–Bouldin index exceeds that of the previous accepted round by more than $5\%$, and a cluster is rejected if its membership has Jaccard overlap below $0.3$ with the samples previously promoted to the same output dimension. From each accepted cluster, up to 200 of the most confident members are promoted with the majority output dimension as their label. Novel selection runs for at most six rounds (epochs 50–100).