# CALIBRATED CATEGORY DISCOVERY: KNITTING CALIBRATION FROM THE KNOWN INTO THE NOVEL

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

## ABSTRACT

Generalized Category Discovery (GCD) aims to recognize old classes while discovering novel ones in unlabeled data, yet the practical value of such discoveries hinges on whether the model's confidence can be trusted. Existing GCD methods optimize clustering accuracy alone, leaving their predictions on novel classes markedly miscalibrated. We argue that this deficiency cannot be resolved by conventional post-hoc calibration. First, miscalibration in GCD is inherently cluster-level, whereas conventional methods apply a single sample-level mapping to all clusters. Second, novel classes provide no labels for calibrating confidence. To address this, we formulate Calibrated Category Discovery (CalCD) and introduce a cluster-conditional expected calibration error (cluECE) that assesses calibration within each predicted cluster, preventing the deviations of different clusters from cancelling as they do in the standard expected calibration error. We further propose Known-to-Novel calIbration Transfer (KNIT), grounded in the premise that the trustworthiness of a cluster may follow a class-agnostic pattern. KNIT characterizes cluster reliability through the model's self-assessment and an independent semantic check by a vision-language model, estimates this pattern on labeled old-class clusters, and transfers it to novel-class clusters, correcting confidence without altering predictions. Extensive experiments comprising 220 comparisons across 4 GCD methods, 11 datasets spanning generic, fine-grained and multi-domain images, and 5 calibration baselines demonstrate that KNIT significantly reduces cluECE for both novel and old classes.

## 1 INTRODUCTION

Deep neural networks have achieved remarkable success in visual recognition (He et al., 2016; Dosovitskiy et al., 2021) but can only recognize classes predefined for training. In practice, however, novel classes keep emerging and are mixed with known ones: whether a newly collected image belongs to a known or a novel class is often unknown in advance, and manually annotating the continuously arriving data one by one requires substantial human effort and expertise. Generalized Category Discovery (GCD) (Vaze et al., 2022) was proposed for this situation: given some labeled samples of known classes, the model must classify and cluster unlabeled samples of both known and novel classes, i.e., recognize known classes while discovering novel ones. Existing GCD methods fine-tune pretrained models with supervision on labeled samples and contrastive learning or pseudo-labeling on unlabeled ones (Wen et al., 2023); recent methods further use the text modality of CLIP (Radford et al., 2021) to distinguish visually similar classes (Zheng et al., 2024).

Although GCD has greatly improved clustering accuracy, high-stakes applications require a model not only to predict correctly but also to be well calibrated, i.e., its confidence should match the probability of being correct (Guo et al., 2017). In medical imaging, for example, a GCD model must assign new cases to known diseases or group rare, not-yet-annotated lesions into new classes. Physicians need to rely on confidence to decide which predictions can be accepted and which cases must be re-examined, so whether confidence faithfully reflects prediction accuracy is critical: once the model is overconfident on novel classes, misclassified cases skip re-examination, leading to missed diagnoses or misdiagnoses. However, existing GCD methods mostly focus only on clustering accuracy (Vaze et al., 2022; Wen et al., 2023; Wang et al., 2025; Caselli et al., 2026), and experiments



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.

show that these methods are very poorly calibrated (Table 1). Building a well-calibrated GCD model is therefore a pressing problem for bringing GCD into practical use.

The most direct way to calibrate a GCD model on novel classes is to reuse post-hoc methods such as temperature scaling (TS) (Guo et al., 2017), Platt scaling (Platt, 1999), isotonic regression (Zadrozny & Elkan, 2002), and Dirichlet calibration (Kull et al., 2019), but they are of limited effectiveness on novel classes for two reasons. First, GCD's goal is "cluster-level", whereas these methods are sample-level. GCD separates novel-class samples by semantics, so the model should give trustworthy confidence for each discovered novel class. However, GCD miscalibration varies across clusters, with some clusters overconfident and others underconfident, whereas these methods apply the same mapping to all samples and cannot correct clusters individually. Second, novel classes are unlabeled, so novel-cluster accuracy is unobservable, and the mapping can only be fitted on known classes and copied to novel ones. Only a sample-level mapping is copied, while the more valuable information in known-class labels, namely "what kind of cluster deserves trust", is discarded. Calibrating novel classes therefore requires correcting confidence at the cluster level and learning, from labeled known clusters, a calibration pattern that transfers to novel clusters.

To this end, our core idea is that "what kind of cluster deserves trust" may be a class-agnostic pattern. A known cluster that is more confident than other known clusters and whose members are semantically more consistent tends to be more accurate; the same holds for novel clusters. Neither cue needs labels, so both are computable on novel clusters. This pattern can thus be learned on labeled known clusters and transferred to unlabeled novel clusters.

Specifically, we first propose the Calibrated Category Discovery (CalCD) task, which requires confidence to match the probability of a correct prediction within every predicted cluster at minimal cost in GCD accuracy, with calibration labels available only for known classes. For evaluation, ECE pools all clusters' samples into shared bins, so deviations of different clusters cancel out; we therefore propose the cluster-conditional expected calibration error (cluECE), which bins each cluster independently and is reported separately for known and novel clusters. In this setting, we propose the Known-to-Novel calIbration Transfer (KNIT) method. On top of TS, KNIT adds a per-cluster offset that adjusts only confidence, leaving predictions unchanged. The offset is set by the transferable pattern above according to whether a cluster deserves trust, which KNIT characterizes from internal and external views. The internal view is the model's self-assessment, i.e., how confident it is on a cluster's samples and whether the cluster tends to collapse; the external view uses an independent vision–language model to check the semantic consistency within a cluster. Since known and novel clusters receive different supervision, their overall confidence levels do not transfer, so each keeps its average confidence at the TS level.

Our main contributions are summarized as follows: **(1)** We propose the CalCD task and the cluECE metric, revealing the cluster-level miscalibration of GCD models on novel classes that ECE masks. **(2)** We propose KNIT, which transfers the pattern "what kind of cluster deserves trust" learned on known clusters to novel clusters, and calibrates via cluster-level offsets without changing predictions. **(3)** Across 220 comparisons (4 GCD methods $\times$ 11 datasets $\times$ 5 calibration methods), KNIT significantly reduces the cluECE of both novel and known classes.

## 2 Related Work

**Generalized category discovery.** Generalized Category Discovery (Vaze et al., 2022) relaxes the assumption of novel class discovery (Han et al., 2019; 2020; Fini et al., 2021) that unlabeled data contain only novel classes, and requires classifying and clustering unlabeled data that mix known and novel classes. Existing methods fall roughly into three groups: non-parametric, clustering learned representations (Vaze et al., 2022; Pu et al., 2023; Zhang et al., 2023; Choi et al., 2024; Zhao et al., 2023); parametric, represented by SimGCD (Wen et al., 2023), jointly training representations and per-class classification heads by self-distillation (Cao et al., 2024; Liu & Han, 2025; Wang et al., 2024; Yu et al., 2026); and multimodal, using CLIP (Radford et al., 2021) text to distinguish visually similar classes (Ouldnoughi et al., 2023; Zheng et al., 2024; Wang et al., 2025; Caselli et al., 2026). They focus on clustering accuracy, while novel-class calibration has not been separately evaluated; we evaluate it, find existing methods clearly miscalibrated, and accordingly propose the Calibrated Category Discovery (CalCD) task.

**Probabilistic calibration.** Calibration requires a model's confidence to reflect its probability of being correct (Guo et al., 2017). Training-time methods suppress overconfidence through the training objective (Müller et al., 2019; Mukhoti et al., 2020; Thulasidasan et al., 2019; Kumar et al., 2018), while post-hoc methods fit a shared or class-wise transformation on held-out labeled data (Platt, 1999; Zadrozny & Elkan, 2002; Naeini et al., 2015; Guo et al., 2017; Zhang et al., 2020; Kull et al., 2019), or estimate tail-class calibration from head classes under long tails (Chen & Su, 2023). Both require labels for every class and calibrate per sample, and thus cannot handle unlabeled novel clusters defined by clustering; moreover, confidence-binned ECE (Naeini et al., 2015; Guo et al., 2017) lets the deviations of different clusters cancel out within a bin, whereas we measure and correct confidence per cluster.

Other work calibrates without labels: Calibrated Deep Clustering (CDC) (Jia et al., 2025) targets fully unlabeled deep clustering with no labeled classes to draw on, while PseudoCal (Hu et al., 2024) and other cross-domain methods (Tomani et al., 2021; Wang et al., 2020) transfer calibration across domains with the same classes and involve no novel classes. Neither applies directly to CalCD; our KNIT instead transfers calibration knowledge from labeled known classes to unlabeled novel classes.

# 3 Calibrated Category Discovery

This section reviews GCD and confidence calibration (Section 3.1), analyzes why the conventional calibration setting falls short for GCD, and accordingly defines the CalCD task and its metric cluECE (Section 3.2).

## 3.1 Preliminaries

**Generalized category discovery.** In GCD (Vaze et al., 2022), the labeled set $\mathcal{D}_l$ contains only known classes $\mathcal{Y}_k$, while the unlabeled set $\mathcal{D}_u$ contains both $\mathcal{Y}_k$ and novel classes $\mathcal{Y}_n$ ($\mathcal{Y}_k \cap \mathcal{Y}_n = \varnothing$); using both, a model must be trained to assign each sample $x$ in $\mathcal{D}_l \cup \mathcal{D}_u$ to one of $K$ clusters. GCD is evaluated by accuracy under the optimal cluster-to-class mapping (Vaze et al., 2022; Wen et al., 2023):

$$\mathrm{ACC} = \max_{\pi' \in \Pi} \frac{1}{|\mathcal{D}_u|} \sum_{(x,y)\in\mathcal{D}_u} \mathbf{1}\big[\pi'(\hat{y}(x)) = y\big], \tag{1}$$

where $\hat{y}(x)$ is the predicted cluster of $x$ and $\Pi$ is the set of one-to-one cluster-to-class mappings. The maximizer $\pi$ is found by the Hungarian algorithm; $g(x, y) = \mathbf{1}[\pi(\hat{y}(x)) = y]$ indicates whether $(x, y)$ is correctly predicted under $\pi$. Clusters with $\pi(j) \in \mathcal{Y}_k$ are known clusters; the rest are novel.

**Confidence calibration.** In fixed-class classification, a model gives a confidence $c(x) \in [0, 1]$ for each prediction $\hat{y}(x)$. For a test sample $(X, Y)$ with $\hat{Y} = \hat{y}(X)$ and $C = c(X)$, calibration requires confidence to equal the probability of being correct (Guo et al., 2017):

$$\Pr\Big(Y = \hat{Y} \,\Big|\, C = c\Big) = c, \qquad \forall c \in [0, 1]. \tag{2}$$

Calibration is usually performed on a labeled validation set covering all classes.

## 3.2 Problem Formulation and Evaluation Metric

The most direct way to study GCD calibration is to adopt the conventional setting: take Eq. (2) as the objective and calibrate on a labeled validation set. However, this setting cannot effectively support the study of GCD model calibration, for two reasons.

**First, ECE cannot reflect cluster-level calibration.** GCD's goal of discovering novel classes, i.e., separating novel-class samples from one another by semantics, is cluster-level, so a well-calibrated GCD model should give trustworthy confidence for each discovered novel class. After TS, GET (Wang et al., 2025) has an ECE of only 2.8% on novel-class samples (Fig. 1(a)), yet examining the novel clusters one by one, some are underconfident and others overconfident, and for half of the clusters the average confidence and the accuracy differ by more than 11 percentage points (Fig. 1(b)). ECE bins samples of all clusters together, so their miscalibrations cancel out; hence, ECE cannot directly evaluate GCD calibration.

**Second, GCD lacks the all-class labeled validation set required by conventional calibration.** In GCD, novel classes are both unlabeled and unknown. A realistic configuration thus has a labeled

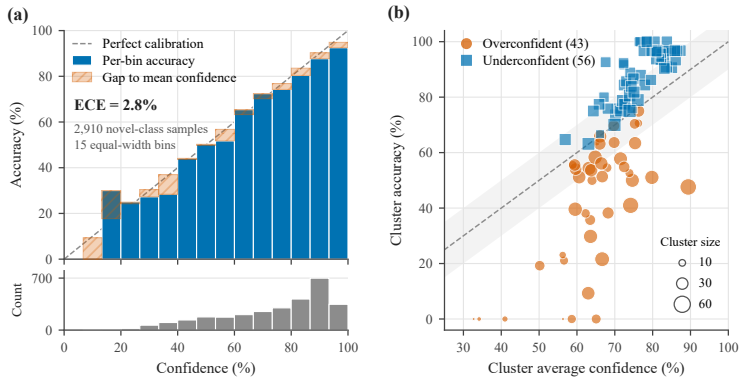



Figure 1: Novel-class predictions of GET on CUB after TS. (a) Reliability diagram of all novel-class samples (15 equal-width bins): the ECE is low. (b) Accuracy and average confidence of all novel clusters: some are overconfident, others underconfident.

validation set $\mathcal{D}^l_{\text{val}}$ of known classes only, identically distributed with and independent of the training set, plus an unlabeled set $\mathcal{D}^u_{\text{val}}$ of known and novel classes.

Accordingly, we propose and formally define the Calibrated Category Discovery (CalCD) task and its metric.

**Calibrated Category Discovery.** Suppose a GCD model gives each sample $x$ a predicted cluster $\hat{y}(x)$ and a confidence $c(x)$, and that, beyond the training data, the additional data available are a labeled set $\mathcal{D}^l_{\text{val}}$ containing only known-class samples and an unlabeled set $\mathcal{D}^u_{\text{val}}$ containing both known and novel classes. CalCD requires the confidence $C$ to satisfy cluster-level calibration while preserving, as far as possible, the accuracy defined in Eq. (1):

$$\Pr\Big(Y = \pi(\hat{Y}) \,\Big|\, \hat{Y} = j,\ C = c\Big) = c, \qquad \forall j \in \{1, \ldots, K\},\ c \in [0, 1]. \tag{3}$$

As $\mathcal{D}^l_{\text{val}}$ has no novel-class samples, novel-cluster accuracy is unobservable during calibration, so the core challenge of CalCD is transferring calibration learned on known clusters to novel ones.

**Cluster-conditional expected calibration error.** To quantify deviation from Eq. (3), we follow ECE but first partition samples by predicted cluster and then bin within each cluster independently, avoiding cross-cluster cancellation; this yields the cluster-conditional expected calibration error (cluECE). Let $S_j$ be the test samples (out of $N$) predicted to cluster $j$; cluECE is the weighted sum of per-cluster ECEs:

$$\text{cluECE} = \sum_j \frac{|S_j|}{N} \text{ECE}(S_j), \qquad \text{ECE}(S) = \sum_{B \in \mathcal{B}(S)} \frac{|B|}{|S|} \big|\text{acc}(B) - \text{conf}(B)\big|, \tag{4}$$

where $\mathcal{B}(S)$ partitions $S$ into confidence bins (Section A), and $\text{acc}(B)$ and $\text{conf}(B)$ are within-bin accuracy and average confidence. Correctness is judged after mapping clusters to classes via $\pi$ (i.e., $g$). For brevity, a class $y$ hereafter also denotes its cluster $\pi^{-1}(y)$; e.g., $p_y$ is that cluster's probability, and $Y = \hat{Y}$ means $Y = \pi(\hat{Y})$. Computing Eq. (4) over all clusters gives $\text{cluECE}_{\text{all}}$; computing it on the known and novel clusters (split by $\pi$) gives $\text{cluECE}_{\text{old}}$ and $\text{cluECE}_{\text{new}}$. $\text{cluECE}_{\text{new}}$ is the main metric of CalCD. Estimation is detailed in Section 5.1 and Section A.

## 4 Known-to-Novel calIbration Transfer

This section presents KNIT. Section 4.1 starts from temperature scaling and introduces cluster-level offsets as the form of correction; Section 4.2 formalizes calibration as learning a transferable pattern and explains the roles of the two kinds of calibration data in fitting and inference; Section 4.3 constructs class-transferable cluster descriptors and determines the level constants, so that novel-cluster offsets follow from the pattern learned on known clusters. Parameter estimation and algorithm details are given in Section B. Figure 2 gives an overview.

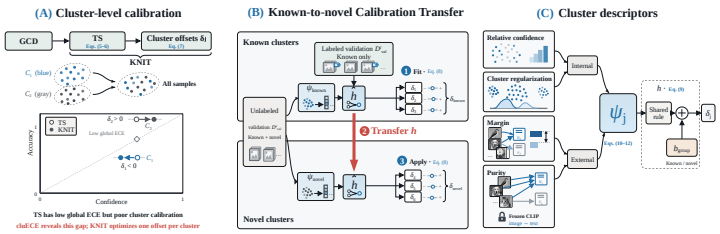



Figure 2: **Overview of KNIT.** (A) On top of TS, KNIT adds one offset $\delta_j$ per cluster (Eq. (7)), correcting the cluster-level miscalibration that a low ECE can hide. (B) The transferable pattern $h$ is fitted on known clusters with the labels of $\mathcal{D}^l_{\text{val}}$ (Eq. (8)) and applied to novel clusters. (C) Descriptors $\boldsymbol{\psi}_j$ from the internal and external views (Eqs. (10) to (12)) are mapped to $\delta_j$ by $h$ (Eq. (9)).

### 4.1 From Temperature Scaling to Cluster-Level Offsets

**Temperature scaling for CalCD.** Temperature scaling (TS) (Guo et al., 2017) is the most widely used and an effective post-hoc calibration method; KNIT builds on it. TS adjusts the logit distribution with a single temperature, optimizes it on a validation set, and reuses it on test samples. Specifically, for the logits $z(x)$ of a GCD model over $K$ clusters, the predicted probability at temperature $\tau_{\text{cal}}$ is

$$p_j(x;\tau_{\text{cal}}) = \frac{\exp(z_j(x)/\tau_{\text{cal}})}{\sum_{k=1}^{K}\exp(z_k(x)/\tau_{\text{cal}})}, \tag{5}$$

and $\tau_{\text{cal}} = 1$ gives the uncalibrated output. The temperature is optimized for the negative log-likelihood (NLL) on the labeled validation set $\mathcal{D}^l_{\text{val}}$:

$$\tau^\star_{\text{cal}} = \arg\min_{\tau_{\text{cal}}} \ \text{NLL}\left(\tau_{\text{cal}};\, \mathcal{D}^l_{\text{val}}\right), \tag{6}$$

where NLL, a common calibration objective, is defined in Section B. The resulting $\tau^\star_{\text{cal}}$ is applied uniformly to all clusters, including novel clusters without labeled samples. Let $p(x) = p(x;\tau^\star_{\text{cal}})$; its largest entry is the TS-calibrated confidence $c(x)$, and its index is the predicted cluster $\hat{y}(x)$.

**From a global temperature to cluster-level offsets.** With a single temperature, TS can only adjust confidence globally; yet GCD miscalibration varies across clusters (Fig. 1), and Eq. (3) requires each cluster to be calibrated separately. KNIT therefore adds on top of TS an offset $\delta_j$ for each cluster $j$, shared by all its samples:

$$\tilde{c}(x) = \sigma\Big(\sigma^{-1}\big(c(x)\big) + \delta_{\hat{y}(x)}\Big), \tag{7}$$

where $\sigma$ is the sigmoid function and $\sigma^{-1}$ its inverse; non-predicted class probabilities are rescaled proportionally to sum to 1. The offset changes confidence but not the predicted cluster, so classification accuracy equals that of TS; the minimal-change and prediction-preservation properties of this form are described in Section B. Calibration thus reduces to determining a set of cluster-level offsets $\{\delta_j\}$.

### 4.2 Calibration as Learning Transferable Patterns

The difficulty in determining these offsets lies in novel clusters: without validation labels, their accuracy is unobservable and their offsets cannot be fitted directly. To address this, the core idea of KNIT is that "what kind of cluster deserves trust" may be a class-agnostic pattern. As long as a known cluster is more confident and its members semantically more consistent than other known clusters, its accuracy tends to be higher; the same holds for novel clusters. KNIT therefore learns this pattern on labeled known clusters and applies it to novel clusters.

**Formalization of KNIT.** Let $\boldsymbol{\psi}_j$ be the label-independent descriptor of cluster $j$ (Section 4.3), computed from the $\mathcal{D}^u_{\text{val}}$ samples predicted to cluster $j$. KNIT assumes that known and novel clusters share one transferable pattern $h : \boldsymbol{\psi}_j \mapsto \delta_j$, mapping a cluster's descriptor to its offset. The transfer process is

$$\hat{h} = \arg\min_{h} \ \text{NLL}\left(h;\, \mathcal{D}^l_{\text{val}}\right), \qquad \hat{\delta}_j = \hat{h}(\boldsymbol{\psi}_j), \quad j \in \mathcal{Y}_k \cup \mathcal{Y}_n, \tag{8}$$

with NLL as in Eq. (6). Since $\pi$ is unavailable during calibration, KNIT distinguishes known and novel clusters by whether their output units were supervised by labels during training (this split does not exactly coincide with the split by $\pi$ in Section 3.1; see Section B). The first expression learns $\hat{h}$

only from $\mathcal{D}^l_{\text{val}}$ samples predicted to known clusters; the second applies $\hat{h}$ to all clusters, including the unlabeled novel clusters. This learn-then-generalize process is exactly what KNIT means by calibration transfer from known to novel classes (Known-to-Novel calIbration Transfer).

**Form of the mapping.** Transfer requires $h$ to capture only the pattern shared by known and novel clusters, but the more complex $h$ is, the more easily it fits patterns specific to known clusters. KNIT therefore takes $h$ to be linear:

$$h(\boldsymbol{\psi}_j) = \boldsymbol{\theta}^\top \boldsymbol{\psi}_j + b_j, \tag{9}$$

where $\boldsymbol{\theta}$ captures the pattern linking descriptors to miscalibration, and $b_j$ is a level constant, $b_k$ for known clusters and $b_n$ for novel ones. In practice, the first three coefficients of $\boldsymbol{\theta}$ are estimated once by pooling the known clusters of 18 combinations and are then frozen, and only the coefficient $\lambda$ of the fourth descriptor is fitted for each model (Section B). Section 4.3 is devoted to the design of the descriptors and level constants.

**Data and inference.** Overall, the calibration process of KNIT consists of Eqs. (6), (8) and (9), with different stages using different data. Fitting the temperature and $\boldsymbol{\theta}$ uses $\mathcal{D}^l_{\text{val}}$ labels, whereas computing descriptors and level constants uses only model outputs on $\mathcal{D}^u_{\text{val}}$. At inference, each test sample takes its predicted cluster's offset and obtains its calibrated confidence via Eq. (7). The whole process uses no novel-class labels and trains no network; Algorithm 1 in Section B gives the complete algorithm.

### 4.3 Transferable Cluster Descriptors

Whether Eq. (9) can apply the pattern learned on known clusters to novel clusters depends on how the descriptor $\boldsymbol{\psi}_j$ is designed. KNIT assesses cluster reliability from an internal and an external view. The internal view is the model's self-assessment: how confident it is on the cluster's samples and whether the cluster tends to collapse; the external view uses an independent vision–language model to check intra-cluster semantic consistency, so as to uncover errors that the model itself can hardly detect. Accordingly, KNIT constructs four descriptors and, with the level constants, writes out the complete form of Eq. (9).

**Descriptor 1: relative confidence.** The model's average confidence on a cluster is its self-assessment of the cluster's reliability. For estimating model accuracy on unlabeled data, average confidence is the most basic and common cue (Garg et al., 2022; Guillory et al., 2021). Its ranking across clusters usually matches that of accuracy, though their scales may not align. KNIT takes its log-odds:

$$\psi_j^{\text{conf}} = \sigma^{-1}\Big(\operatorname*{mean}_{u \in \mathcal{U}_j} c(u)\Big), \tag{10}$$

where $\mathcal{U}_j$ is the set of $\mathcal{D}^u_{\text{val}}$ samples predicted to cluster $j$.

**Descriptors 2 and 3: semantic purity.** The semantic consistency of a cluster's samples is an external check of its reliability. Intuitively, members of a cluster corresponding to a single class should be describable by one concept that also distinguishes the cluster from others, which self-assessment cannot check. KNIT checks these two points with a frozen CLIP (Radford et al., 2021) on a generic vocabulary $\mathcal{V}$ (Zheng et al., 2024; Caselli et al., 2026):

$$\psi_j^{\text{marg}} = \text{Margin}(\mathcal{V}, \mathcal{U}_j), \qquad \psi_j^{\text{pur}} = \text{Purity}(\mathcal{V}, \mathcal{U}_j), \tag{11}$$

where $\text{Margin}$ is the gap between the largest and second-largest average image–text similarity of cluster members to the concepts, and $\text{Purity}$ is the cluster purity computed from the zero-shot classification results; Section B gives detailed definitions.

**Descriptor 4: cluster regularization.** To avoid the degenerate solution of cluster collapse, GCD algorithms introduce a mean-entropy maximization regularizer (Wen et al., 2023; Wang et al., 2025; Caselli et al., 2026; He et al., 2026). Likewise, KNIT introduces this regularization term inside $h$:

$$\psi_j^{\text{reg}} = \log \frac{K\,|\mathcal{U}_j|}{|\mathcal{D}^u_{\text{val}}|}. \tag{12}$$

**Level constants.** Known and novel clusters receive different supervision, so their overall confidence levels cannot be transferred. Hence $b_k$ and $b_n$ are not estimated by Eq. (8); instead, they keep the average confidence of known and novel clusters, respectively, at the TS level (Section B).

**Complete form.** Standardizing each descriptor separately within known and within novel clusters and substituting into Eq. (9) gives

$$\delta_j = \boldsymbol{\theta}^\top \boldsymbol{\psi}_j + b_j = \theta_1 \, \psi_j^{\mathrm{conf}} + \theta_2 \, \psi_j^{\mathrm{marg}} + \theta_3 \, \psi_j^{\mathrm{pur}} + \lambda \, \psi_j^{\mathrm{reg}} + b_j, \tag{13}$$

where the first three coefficients, shared across datasets and GCD methods, are $(0.3, 0.2, 0.25)$, estimated once by pooled fitting and frozen (Eq. (16) in Section B); $\lambda$ is estimated per model on its known clusters (Eq. (17) in Section B).

# 5 Experiments

## 5.1 Experimental Setup

**Datasets.** The experimental data fall into four categories by their nature. (1) The generic datasets are CIFAR-10, CIFAR-100, and ImageNet-100. (2) The fine-grained datasets are CUB-200-2011, Stanford Cars, FGVC-Aircraft, and Flowers-102 (Nilsback & Zisserman, 2008). These two categories are the standard evaluation data of existing GCD methods, and we follow their commonly used splits (Vaze et al., 2022; Wen et al., 2023; Zheng et al., 2024). (3) The domain-shifted datasets are the four domains of Office-Home (Venkateswara et al., 2017), whose image styles differ markedly across domains. It is less commonly used for GCD evaluation, and we use it to evaluate the robustness of calibration methods under different image styles (Section C). (4) The long-tailed dataset is CIFAR-100 resampled according to a long-tailed curve, which is commonly used in long-tailed GCD research (Bai et al., 2023). The splits of all datasets and the construction of the long-tailed data are given in Section C.

**Compared methods.** (1) Base GCD methods. We adopt four GCD methods, SimGCD (Wen et al., 2023), GET (Wang et al., 2025), SpectralGCD (Caselli et al., 2026), and SSR$^2$-GCD (He et al., 2026), all of which use CLIP (Radford et al., 2021) ViT-B/16 as the backbone; following the existing GCD setting, the number of clusters is set to the ground-truth number of classes (Vaze et al., 2022; Wen et al., 2023), and the final model is used. (2) Calibration baselines. These include the uncalibrated outputs, temperature scaling (TS) (Guo et al., 2017), ensemble temperature scaling (ETS) (Zhang et al., 2020), histogram binning (Zadrozny & Elkan, 2001), isotonic regression (Zadrozny & Elkan, 2002), and PseudoCal (Hu et al., 2024). None of the methods changes the predicted clusters; implementation details are given in Section C.

**Evaluation metrics and data splits.** The primary metric is cluECE$_{\text{new}}$ defined in Eq. (4); we also report cluECE$_{\text{old}}$ and cluECE$_{\text{all}}$, and the conventional ECE metric is also used in the analyses. The clustering accuracy of GCD is computed by Eq. (1), i.e., the accuracy under the optimal cluster-to-class mapping found by the Hungarian algorithm (Vaze et al., 2022). Each dataset contains a training set, a validation set, and a test set. (1) The training set follows the GCD setting and is divided into the labeled $\mathcal{D}_l$ (containing only known classes) and the unlabeled $\mathcal{D}_u$ (containing known and novel classes). (2) The validation set is used for calibration and is likewise divided into the labeled $\mathcal{D}_{\text{val}}^l$ (containing only known classes) and the unlabeled $\mathcal{D}_{\text{val}}^u$ (containing known and novel classes) (Section A). (3) The test set contains known and novel classes and is used to compute the above metrics. The whole calibration process uses no novel-class labels.

**Implementation details.** The shared coefficients of the relative confidence and the two semantic purity measures are $(0.3, 0.2, 0.25)$, obtained by pooled fitting on the known clusters of the three GCD methods on CIFAR-10, CIFAR-100, ImageNet-100, CUB-200-2011, Stanford Cars, and FGVC-Aircraft (Eq. (16)); the prior of the cluster regularization coefficient $\lambda$ is $\mathcal{N}(0, 0.5^2)$ (Eq. (17)). The vocabulary $\mathcal{V}$ follows the Open Images image-level labels used by the GCD method SpectralGCD (Caselli et al., 2026); images and texts are encoded with the pretrained CLIP ViT-B/16 used by the base GCD methods, and the encoder is kept frozen (Section C).

## 5.2 Main Results

As shown in Table 1, KNIT achieves the best cluECE$_{\text{new}}$ in almost all cases across the three image categories and the four base methods. Over all 44 combinations, the average cluECE$_{\text{new}}$ and

Table 1: cluECE (%, lower is better) of each calibration method on three image types. Generic images are averaged over CIFAR-10, CIFAR-100, and ImageNet-100; domain-shifted images are averaged over the four Office-Home domains; fine-grained images are averaged over CUB-200-2011, Stanford Cars, FGVC-Aircraft, and Flowers-102. Bold marks the column-wise minimum within the same base method; full per-dataset results are given in Tables C4 to C6 in the appendix.

| Base method | Calibration | Generic images | | | Domain-shifted images | | | Fine-grained images | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | All | Old | New | All | Old | New | All | Old | New |
| SimGCD (Wen et al., 2023) | Uncalibrated | 11.3 | 8.8 | 16.8 | 23.6 | **17.6** | 28.1 | 21.8 | 20.5 | 22.6 |
| | TS (Guo et al., 2017) | 5.1 | 3.5 | 7.1 | 22.2 | 21.8 | 22.5 | 15.8 | 11.9 | 18.7 |
| | Isotonic regression (Zadrozny & Elkan, 2002) | 5.8 | **3.1** | 10.3 | 22.7 | 23.0 | 22.4 | 15.0 | 12.4 | 17.0 |
| | PseudoCal (Hu et al., 2024) | 9.1 | 8.2 | 8.9 | 34.2 | 20.0 | 45.0 | 27.8 | 13.6 | 38.2 |
| | KNIT | **4.4** | 3.3 | **5.7** | **20.3** | 20.0 | **20.5** | **12.7** | **10.3** | **14.4** |
| GET (Wang et al., 2025) | Uncalibrated | 9.7 | 7.8 | 14.7 | 23.4 | **17.5** | 27.8 | 22.7 | 20.9 | 24.2 |
| | TS | 5.4 | 4.0 | 9.2 | 22.1 | 19.4 | 24.2 | 14.3 | 11.3 | 16.8 |
| | Isotonic regression | 6.3 | **3.9** | 11.6 | 23.0 | 22.2 | 23.6 | 14.4 | 11.5 | 16.9 |
| | PseudoCal | 9.2 | 8.4 | 9.9 | 33.9 | 19.1 | 44.9 | 22.2 | 12.5 | 29.6 |
| | KNIT | **5.1** | **3.9** | **8.0** | **20.4** | 17.7 | **22.4** | **11.3** | **10.2** | **12.3** |
| SpectralGCD (Caselli et al., 2026) | Uncalibrated | 11.2 | 8.9 | 16.4 | 19.7 | 17.3 | 21.8 | 25.6 | 24.7 | 26.5 |
| | TS | 4.7 | 3.2 | 7.8 | 19.0 | 18.2 | 19.8 | 12.7 | 9.9 | 15.0 |
| | Isotonic regression | 4.9 | **3.1** | 8.7 | 19.6 | 19.5 | 19.8 | 11.9 | **8.5** | 14.6 |
| | PseudoCal | 8.3 | 7.4 | 8.4 | 26.8 | **16.4** | 35.6 | 20.4 | 13.3 | 26.0 |
| | KNIT | **4.3** | **3.1** | **6.9** | **17.1** | 17.0 | **17.3** | **10.6** | 8.9 | **11.8** |
| SSR$^2$-GCD (He et al., 2026) | Uncalibrated | 5.3 | 4.2 | 8.2 | 17.0 | 14.0 | 20.1 | 15.3 | 14.1 | 16.4 |
| | TS | 4.9 | 3.8 | 7.4 | 17.8 | 17.1 | 18.5 | 10.8 | 9.7 | 11.7 |
| | Isotonic regression | 4.6 | **3.3** | **6.7** | 18.3 | 16.5 | 20.0 | **9.6** | **8.0** | **11.0** |
| | PseudoCal | 5.7 | 4.1 | 7.6 | 20.7 | **12.4** | 29.1 | 16.9 | 13.8 | 19.5 |
| | KNIT | **4.4** | 3.4 | **6.7** | **16.4** | 15.3 | **17.5** | 9.9 | 8.4 | 11.1 |

Table 2: Ablation study. ✓ indicates that the item is included; values are the average cluECE over the 44 combinations (%, lower is better).

| Method | $\psi^{\text{conf}}$ equation 10 | $\psi^{\text{marg}}$ equation 11 | $\psi^{\text{pur}}$ equation 11 | $\psi^{\text{reg}}$ equation 12 | $b_j$ equation 13 | $\text{cluECE}_{\text{old}}$ | $\text{cluECE}_{\text{new}}$ |
|---|---|---|---|---|---|---|---|
| TS (Guo et al., 2017) | – | – | – | – | – | 11.83 | 15.53 |
| | ✓ | × | × | × | ✓ | 11.38 | 14.93 |
| | × | ✓ | ✓ | ✓ | ✓ | 10.85 | 13.51 |
| | ✓ | × | ✓ | ✓ | ✓ | **10.74** | 13.59 |
| | ✓ | ✓ | × | ✓ | ✓ | 11.01 | 13.95 |
| | ✓ | ✓ | ✓ | × | ✓ | 11.00 | 14.13 |
| | ✓ | ✓ | ✓ | ✓ | × | 11.13 | 13.77 |
| KNIT | ✓ | ✓ | ✓ | ✓ | ✓ | **10.74** | **13.44** |

Table 3: cluECE on long-tailed CIFAR-100 (%, lower is better; averaged over GET and SpectralGCD). $\rho$ is the imbalance ratio, i.e., the ratio between the numbers of training images of the most- and least-populated classes; the construction is given in Section C. Bold marks the column-wise minimum.

| Calibration | $\rho = 20$ | | | $\rho = 50$ | | | $\rho = 100$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | All | Old | New | All | Old | New | All | Old | New |
| Uncalibrated | 21.5 | 20.4 | 25.5 | **19.5** | **19.1** | 21.0 | **19.8** | **18.7** | 24.6 |
| TS (Guo et al., 2017) | 22.3 | 21.4 | 26.0 | 22.6 | 23.2 | 20.5 | 23.5 | 23.4 | **23.6** |
| KNIT | **20.8** | **19.7** | **25.0** | 21.3 | 21.5 | **20.4** | 22.4 | 21.8 | 24.5 |

$\text{cluECE}_{\text{old}}$ of KNIT are 13.44 and 10.74, those of TS are 15.53 and 11.83, and none of the other calibration methods is better than TS. Per combination, KNIT is better than TS on novel classes in 39 combinations and is significantly worse than TS in none. We attribute this to KNIT transferring the pattern of "what kind of cluster deserves trust". The complete results are given in Tables C4 to C8 in the appendix.

### 5.3 Analysis

**Ablation study.** Table 2 shows that every design component of KNIT contributes. Removing any descriptor or the level constants makes $\text{cluECE}_{\text{new}}$ worse than that of the full KNIT (13.44), with removing cluster regularization and Purity having the largest effect. Keeping only the relative confidence gives 14.93, which is still better than the 15.53 of TS.

**ECE masks cluster-level miscalibration.** Conventional ECE masks cluster-level miscalibration, which confirms the analysis in Section 3.2. On the 22 combinations of SimGCD and GET, the ECE of novel-class samples after TS is only 6.34, whereas $\text{cluECE}_{\text{new}}$ is 17.18. KNIT is close to TS in ECE (6.83), whereas in $\text{cluECE}_{\text{new}}$ it drops to 14.54 (Table C9 in the appendix). This is because ECE bins the samples of all clusters together, so the deviations between clusters cancel each other out.

**Miscalibration occurs cluster by cluster.** Correcting only the overall confidence level of novel clusters is insufficient to calibrate novel classes, which confirms the motivation for introducing cluster-level offsets in Section 4.1. The best single temperature or single offset selected with novel-class labels can only reduce $\text{cluECE}_{\text{new}}$ from the 15.53 of TS to 14.35 and 14.43, respectively, whereas per-cluster correction can reduce it to 9.64.

**The miscalibration pattern can be transferred across classes.** The correction of novel clusters by the pattern transferred from known clusters is comparable, in both magnitude and direction, to that by

Table 4: Correction of novel clusters by transfer and by fitting with novel-class labels. $\text{cluECE}_{\text{new}}$ values are averages (%, lower is better); transfer efficiency is (TS − transfer) / (TS − fitting with novel-class labels). Fitting with novel-class labels uses the test labels of novel classes and is used only for diagnosis.

| Data group | $\text{cluECE}_{\text{new}}$ | | | Transfer efficiency | Novel clusters with consistent direction (%) | Offset rank correlation (median) |
|---|---|---|---|---|---|---|
| | TS | Transfer (KNIT) | Fitting with novel-class labels (diagnosis) | | | |
| Generic images | 7.85 | 6.83 | 6.53 | 0.77 | 85.9 | 0.90 |
| Domain-shifted images | 21.26 | 19.42 | 19.93 | 1.38 | 84.0 | 0.85 |
| Fine-grained images | 15.57 | 12.41 | 12.91 | 1.19 | 86.3 | 0.87 |
| All 44 combinations | 15.53 | 13.44 | 13.72 | 1.16 | 85.6 | 0.87 |

Table 5: Sensitivity of KNIT to $\boldsymbol{\theta}$ and $\lambda$ (average cluECE over the 44 combinations, %, lower is better). When one of $\theta_1$, $\theta_2$, and $\theta_3$ is multiplied by the multiple shown in the header, $\lambda$ and the level constants are estimated as usual; the $\lambda$ row multiplies the $\lambda$ estimated for each model by this multiple. The multiple 1 is the default setting. The $\text{cluECE}_{\text{new}}$ and $\text{cluECE}_{\text{old}}$ of TS are 15.53 and 11.83.

| Metric | Coefficient | ×0 | ×0.5 | ×1 | ×1.5 | ×2 |
|---|---|---|---|---|---|---|
| $\text{cluECE}_{\text{new}}$ | $\theta_1$ (relative confidence) | 13.70 | 13.51 | 13.44 | 13.45 | 13.54 |
| | $\theta_2$ (Margin) | 13.72 | 13.52 | 13.44 | 13.45 | 13.53 |
| | $\theta_3$ (Purity) | 14.00 | 13.68 | 13.44 | 13.27 | 13.16 |
| | $\lambda$ (cluster regularization) | 14.13 | 13.65 | 13.44 | 13.56 | 13.81 |
| $\text{cluECE}_{\text{old}}$ | $\theta_1$ (relative confidence) | 10.99 | 10.84 | 10.74 | 10.69 | 10.67 |
| | $\theta_2$ (Margin) | 10.82 | 10.76 | 10.74 | 10.76 | 10.81 |
| | $\theta_3$ (Purity) | 11.07 | 10.89 | 10.74 | 10.62 | 10.52 |
| | $\lambda$ (cluster regularization) | 11.00 | 10.84 | 10.74 | 10.69 | 10.69 |

the pattern fitted directly with novel-class labels, which confirms the hypothesis in Section 4.2 that "what kind of cluster deserves trust" is class-agnostic. As shown in Table 4, over the 44 combinations, the transfer efficiency (the ratio of the two improvements relative to TS) is 1.16, and for 85.6% of novel clusters the two either both raise or both lower the confidence. We attribute this to the fact that the linear $h$ and within-type standardization capture only the relative reliability of a cluster within its type, while the overall difference between the types is absorbed by the level constants $b_k$ and $b_n$.

**Class imbalance.** We evaluate the calibration methods on long-tailed CIFAR-100 with imbalance ratios of 20, 50, and 100 (Table 3). Overall, KNIT remains consistently better than TS under different imbalance ratios.

**Sensitivity and computational cost.** KNIT is insensitive to the values of $\boldsymbol{\theta}$ and $\lambda$ (Table 5). When any coefficient varies between 0.5 and 1.5 times its default value, $\text{cluECE}_{\text{new}}$ lies between 13.27 and 13.68 and $\text{cluECE}_{\text{old}}$ between 10.62 and 10.89, always clearly better than the 15.53 and 11.83 of TS. However, setting any coefficient to zero makes $\text{cluECE}_{\text{new}}$ worse than that of the default setting (13.70 to 14.13), with $\theta_3$ and $\lambda$ having the largest effect, consistent with the ablation study. Increasing $\theta_3$ also yields a slight improvement (13.16 at twice the default value), which shows that the default coefficients were not tuned on the test set. When the shared coefficients are refitted after removing one dataset or one base method, $\text{cluECE}_{\text{new}}$ differs from the 13.44 of pooled fitting by less than 0.05. Fitting separately on each dataset is significantly worse than pooled fitting (14.11); hence, the shared coefficients need to be fitted by pooling across datasets. The effects of the other hyperparameters are given in Section C. In addition, KNIT does not train any network, and a single combination takes 0.3 to 2.5 seconds on a CPU.

**Limitations.** The improvement of KNIT on novel clusters is limited on heavily long-tailed data. On long-tailed CIFAR-100 with $\rho = 100$, the $\text{cluECE}_{\text{new}}$ of KNIT is 24.5, worse than the 23.6 of TS (Table 3). This stems from the level constants (Section 4.3). Long-tailed training makes the clusters without label supervision underconfident overall, whereas the level constants keep the overall level of TS and do not correct this bias (Table C10 in the appendix). After the level constants are removed, the average $\text{cluECE}_{\text{new}}$ on long-tailed CIFAR-100 drops from 23.30 to 22.84, but over the 44 combinations it rises from 13.44 to 13.77 (Table 2).

## 6 Conclusion

We point out that the miscalibration of GCD occurs cluster by cluster, that novel classes lack the labels required for calibration, and that conventional calibration can hardly correct it; accordingly, we propose the Calibrated Category Discovery (CalCD) task and the cluster-conditional expected calibration error cluECE. Based on the idea that "what kind of cluster deserves trust" is class-agnostic, we propose KNIT, which characterizes the reliability of a cluster from two views, the model's self-assessment and an external semantic check, learns this pattern on known clusters, and transfers it to novel clusters. In the 220 experimental comparisons formed by 4 GCD methods, 11 datasets, and 5 calibration methods, KNIT significantly reduces the cluECE of both novel and known classes. Future work may further correct the overall deviation of novel clusters on long-tailed data and explore achieving cluster-level calibration during training.

## AI USE STATEMENT

In the article, we only used LLMs to polish our writing and did not use them for any other assistance.

## REFERENCES


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 (NeurIPS)*, volume 36, pp. 58625–58647, 2023.

Xinzi Cao, Xiawu Zheng, Guanhong Wang, Weijiang Yu, Yunhang Shen, Ke Li, Yutong Lu, and Yonghong Tian. Solving the catastrophic forgetting problem in generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16880–16889, 2024.

Lorenzo Caselli, Marco Mistretta, Simone Magistri, and Andrew Bagdanov. SpectralGCD: Spectral concept selection and cross-modal representation learning for generalized category discovery. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=PyfV9tFmdR.

Jiahao Chen and Bing Su. Transfer knowledge from head to tail: Uncertainty calibration under long-tailed distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19978–19987, 2023.

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)*, pp. 23094–23104, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

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.

Saurabh Garg, Sivaraman Balakrishnan, Zachary C. Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations (ICLR)*, 2022.

Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1134–1144, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017.

Chirag Gupta and Aaditya Ramdas. Top-label calibration and multiclass-to-binary reductions. In *International Conference on Learning Representations (ICLR)*, 2022.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Wei He, Xianghan Meng, Zhiyuan Huang, Xianbiao Qi, Rong Xiao, and Chun-Guang Li. Multi-modal representation learning via semi-supervised rate reduction for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39637–39646, 2026.

Dapeng Hu, Jian Liang, Xinchao Wang, and Chuan-Sheng Foo. Pseudo-Calibration: Improving predictive uncertainty estimation in unsupervised domain adaptation. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19304–19326. PMLR, 2024.

Yuheng Jia, Jianhong Cheng, Hui Liu, and Junhui Hou. Towards calibrated deep clustering network. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=JvH4jDDcG3.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2805–2814. PMLR, 2018.

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

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 15288–15299, 2020.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 29, pp. 2901–2907, 2015.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Sixth Indian Conference on Computer Vision, Graphics & Image Processing (ICVGIP)*, pp. 722–729, 2008.

Rabah Ouldnoughi, Chia-Wen Kuo, and Zsolt Kira. CLIP-GCD: Simple language guided generalized category discovery. *arXiv preprint arXiv:2305.10420*, 2023.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans (eds.), *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

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 (CVPR)*, pp. 7579–7588, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A. Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Christian Tomani, Sebastian Gruber, Muhammed Ebrar Erdem, Daniel Cremers, and Florian Buettner. Post-hoc uncertainty calibration for domain drift scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10124–10132, 2021.

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.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5018–5027, 2017.

Enguang Wang, Zhimao Peng, Zhengyuan Xie, Fei Yang, Xialei Liu, and Ming-Ming Cheng. GET: Unlocking the multi-modal potential of CLIP for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20296–20306, 2025.

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

Ximei Wang, Mingsheng Long, Jianmin Wang, and Michael I. Jordan. Transferable calibration with lower bias and variance in domain adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 19212–19223, 2020.

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.

Jiawei Yu, Zijian Gao, Tianjiao Wan, Xuan Liu, Cheng Yang, and Kele Xu. Reliable confidence alignment for generalized category discovery. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://openreview.net/forum?id=ztvohV1CJl.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, pp. 609–616. Morgan Kaufmann, 2001.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 694–699. ACM, 2002.

Jize Zhang, Bhavya Kailkhura, and T. Yong-Jin Han. Mix-n-Match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11117–11128. PMLR, 2020.

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 (CVPR)*, pp. 3479–3488, 2023.

Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised Gaussian mixture models for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16623–16633, 2023.

Haiyang Zheng, Nan Pu, Wenjing Li, Nicu Sebe, and Zhun Zhong. Textual knowledge matters: Cross-modality co-teaching for generalized visual class discovery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 15110 of *Lecture Notes in Computer Science*, pp. 41–58. Springer, 2024.

# A Evaluation Details

**ECE in Table C9.** The ECE in Table C9 is computed on the test samples whose ground-truth class is a novel class: these samples are first divided by confidence into 15 equal-mass bins, and the differences between the accuracy and the average confidence of the bins are then summed with weights.

**Within-cluster binning.** The test samples of each cluster $j$ are sorted by confidence and divided into $M_j = \max(1, \lfloor |S_j|/50 \rfloor)$ equal-mass bins, i.e., about 50 samples per bin. This draws on the practice of Gupta & Ramdas (2022) in top-label histogram binning, which fixes the number of samples per bin rather than the number of bins: the clusters of fine-grained datasets usually have only a few dozen test samples, so with a fixed number of bins most bins would contain only a few samples and the estimate would be dominated by sampling noise; with a fixed number of samples per bin, a cluster with fewer than 100 samples has only one bin, while clusters with many samples (e.g., the clusters of CIFAR-10 and CIFAR-100) are subdivided within the cluster.

**Source of the validation sets and cross-fitting.** $\mathcal{D}^u_{\text{val}}$ is taken to be the unlabeled training set $\mathcal{D}_u$ of GCD, and calibration uses only the model outputs on it. $\mathcal{D}^l_{\text{val}}$ is drawn from the known-class samples of the test set; hence the results on known classes are reported with two-fold cross-fitting: in each fold, the calibration quantities are fitted on one half of $\mathcal{D}^l_{\text{val}}$ and evaluated on the other half, which avoids using the same samples for both fitting and evaluation. The calibration quantities applied to novel classes are fitted on the whole of $\mathcal{D}^l_{\text{val}}$.

# B KNIT Details

This appendix supplements the content omitted from Section 4: the definition of the negative log-likelihood, the properties of cluster-level offsets and prediction preservation, as well as the computation of the semantic purity measures, parameter estimation, and the complete algorithm.

**Negative log-likelihood.** For a labeled sample set $\mathcal{D}$, $\text{NLL}(\cdot\,; \mathcal{D}) = -\,\text{mean}_{(x,y)\in\mathcal{D}} \log p_y(x)$, where $p$ is the predicted probability determined by the first argument: in Eq. (6) it is $p(x; \tau_{\text{cal}})$ at temperature $\tau_{\text{cal}}$; in Eqs. (8), (16) and (17) it is the probability obtained by determining the offsets from that argument via Eq. (9) and then correcting according to Eq. (7). For the latter, $p_y = \tilde{c}$ when the prediction is correct, and when the prediction is wrong $p_y$ is $(1 - \tilde{c})$ multiplied by a factor independent of the offset; hence the NLL differs from the binary cross-entropy of prediction correctness only by a constant independent of the offset.

**Properties of cluster-level offsets.** The form of Eq. (7) is the cluster-level correction that changes TS the least. Given the expected number of correct predictions $E_j$ of cluster $j$, the probability vectors closest to the TS output (measured by the KL divergence) under this constraint are

$$\min_{q} \sum_{u\in\mathcal{U}_j} \text{KL}\big(q(u) \,\|\, p(u)\big) \quad \text{s.t.} \quad \sum_{u\in\mathcal{U}_j} q_j(u) = E_j \qquad \Longrightarrow \qquad q_k(u) \propto p_k(u)\, e^{\delta_j \mathbf{1}[k=j]}, \tag{14}$$

where $\delta_j$ is the Lagrange multiplier of the constraint and is uniquely determined by $E_j$. After normalization, the largest entry of $q$ is exactly $\tilde{c}$ in Eq. (7), and the probabilities of the non-predicted classes are $p_k(1 - \tilde{c})/(1 - c)$, which preserves the ratios among the TS probabilities. Therefore, whatever value the cluster-level target takes, the correction that changes TS the least is a single offset $\delta_j$ shared within each cluster, i.e., Eq. (7). In addition, $\sigma$ is strictly monotonic and the samples of the same cluster receive the same $\delta_j$, so the order of confidences within a cluster is unchanged; when $\delta_j = 0$, $\tilde{c} = c$ and KNIT reduces to TS.

**Prediction preservation.** To keep the predicted cluster unchanged, $\tilde{c}(x)$ is clipped to $[\ell(x) + \epsilon,\ \max(1-\epsilon, c(x))]$, where $\ell(x) = p_{(2)}(x)/(1-c(x)+p_{(2)}(x))$, $p_{(2)}$ is the second-largest probability, and $\epsilon = 10^{-6}$. The condition $\tilde{c}(x) \geq \ell(x)$ exactly guarantees that the corrected second-largest probability does not exceed $\tilde{c}(x)$; hence both the predictions and the classification accuracy are the same as those of TS.

**Partition into known and novel clusters.** Since $\pi$ is unavailable during calibration, the implementation distinguishes known clusters from novel clusters by whether the output unit corresponding to a cluster was supervised by labels during training. This partition does not exactly coincide with the partition by $\pi$ in Section 3.1, and the difference is large on long-tailed data (Table C10).

**Semantic purity measures.** Let $s(u, v)$ be the image–text cosine similarity between sample $u$ and concept $v$, let $\bar{s}_j(v)$ be its mean over $\mathcal{U}_j$, and let the name of cluster $j$ be $v_j = \arg\max_v \bar{s}_j(v)$. The two semantic purity measures are

$$\mathrm{Margin}(\mathcal{V},\mathcal{U}_j) = \bar{s}_j^{(1)} - \bar{s}_j^{(2)}, \qquad \mathrm{Purity}(\mathcal{V},\mathcal{U}_j) = \operatorname*{mean}_{u\in\mathcal{U}_j} \mathbf{1}\Big[\arg\max_{j'} \; s(u, v_{j'}) = j\Big], \tag{15}$$

where $\bar{s}_j^{(1)}$ and $\bar{s}_j^{(2)}$ are the largest and second-largest values of $\bar{s}_j$ over the vocabulary, and $\arg\max_{j'}$ is taken only over the names of valid clusters. In the implementation, each sample keeps only the 20 concepts with the highest image–text similarity, $\bar{s}_j(v)$ accumulates the part exceeding that sample's 20th-highest similarity, and the remaining concepts are counted as 0. The two measures are standardized and then clipped to $[-3, 3]$.

**Parameter estimation.** $\boldsymbol{\theta}$ in Eq. (8) is estimated in two steps. A single model has only 5 to 100 usable known clusters, so estimating the coefficients of relative confidence and of the two semantic purity measures per model, or even per dataset, is unstable (Section 5.3). We assume that these three coefficients (denoted $\boldsymbol{\theta}_s$) are mainly determined by how the GCD model is trained; hence they are estimated by pooling over the set of combinations $\mathcal{C}$, and are then fixed and used for all models. $\mathcal{C}$ consists of the 18 combinations of SimGCD, GET, and SpectralGCD on the six datasets from CIFAR-10 to FGVC-Aircraft (Section 5.1), and the estimator is

$$\hat{\boldsymbol{\theta}}_s = \arg\min_{\boldsymbol{\theta}_s} \; \frac{1}{|\mathcal{C}|} \sum_{\mathcal{C}} \mathrm{NLL}\left(\boldsymbol{\theta}_s;\, \mathcal{D}_{\mathrm{val}}^{l}\right), \tag{16}$$

where $\lambda = 0$ during estimation and each combination is computed with its own temperature, descriptors, and $\mathcal{D}_{\mathrm{val}}^{l}$. The direction of the cluster-regularization coefficient $\lambda$ depends on the class distribution of the data (Section 4.3); hence $\lambda$ is estimated for each model on its own known clusters with $\hat{\boldsymbol{\theta}}_s$ fixed, and a prior is added so that it tends to 0 when the evidence is insufficient:

$$\hat{\lambda} = \arg\min_{|\lambda|\le 1} \; J \cdot \mathrm{NLL}\left(\lambda;\, \mathcal{D}_{\mathrm{val}}^{l}\right) + \frac{\lambda^2}{2 \cdot 0.5^2}, \tag{17}$$

where $J$ is the number of known clusters used in the fit. Samples of the same cluster share the same cluster-regularization quantity, so the effective sample size is the number of clusters rather than the number of samples, and the likelihood term is therefore weighted by $J$; when $J < 10$, $\lambda = 0$. In both equations, $\mathcal{D}_{\mathrm{val}}^{l}$ includes only the samples predicted to valid known clusters that have at least 5 samples in $\mathcal{D}_{\mathrm{val}}^{l}$, and $b_k = b_n = 0$ in both estimation steps.

**Level constants.** $b_k$ and $b_n$ are determined by constraints on the known and the novel clusters, respectively:

$$\operatorname*{mean}_{j} \operatorname*{mean}_{u\in\mathcal{U}_j} \sigma\big(\sigma^{-1}\big(c(u)\big) + \delta_j\big) = \operatorname*{mean}_{j} \operatorname*{mean}_{u\in\mathcal{U}_j} c(u), \tag{18}$$

where the mean over clusters $j$ is taken over the valid clusters of the corresponding type. The left-hand side is strictly monotonic in $b_k$ or $b_n$, so each of them is uniquely determined by one-dimensional root finding.

**Algorithm and information.** The complete procedure of KNIT is given in Algorithm 1. Each step involves only direct computation, low-dimensional convex optimization, and one-dimensional root finding. In the cross-fitting of Section A, $\tau_{\mathrm{cal}}^{\star}$, relative confidence, $\hat{\lambda}$, $b_k$, and $b_n$ are recomputed on the fitting samples of each fold, while $\hat{\boldsymbol{\theta}}_s$ is unchanged. The information used by KNIT consists of the labels of $\mathcal{D}_{\mathrm{val}}^{l}$, the model outputs on $\mathcal{D}_{\mathrm{val}}^{u}$, the frozen CLIP encoder, and the generic vocabulary. It does not use the labels of $\mathcal{D}_{\mathrm{val}}^{u}$, the samples of $\mathcal{D}_l$, or per-class counts, nor does it take any class name as input.

**Algorithm 1** KNIT

**Input:** Logits of the GCD model over $K$ clusters; labeled validation set $\mathcal{D}^l_{\text{val}}$ (known classes only); unlabeled set $\mathcal{D}^u_{\text{val}}$; shared coefficients $\hat{\boldsymbol{\theta}}_s$ obtained by pooled fitting on the 18 combinations (Section B); frozen CLIP encoder and generic vocabulary $\mathcal{V}$.

**Output:** Calibrated confidence $\tilde{c}(x)$ of test samples, with the same predicted clusters as TS.

1: **Temperature scaling:** fit $\tau^\star_{\text{cal}}$ within $[0.05, 20]$ according to Eq. (6), obtaining the confidence $c(x)$ and the predicted cluster $\hat{y}(x)$ of each sample.
2: **Cluster partition:** partition the clusters into known and novel clusters by whether the output unit was supervised by labels during training, and take the clusters with $|\mathcal{U}_j| \geq 5$ on $\mathcal{D}^u_{\text{val}}$ as valid clusters.
3: **Descriptors:** for each valid cluster, compute the raw quantities of relative confidence, Margin, Purity, and cluster regularization on $\mathcal{U}_j$, and robustly standardize them among the valid clusters of the same type to obtain $\boldsymbol{\psi}_j$.
4: **Pattern:** set the first three entries of $\boldsymbol{\theta}$ to $\hat{\boldsymbol{\theta}}_s$; if the number of known clusters used in the fit satisfies $J \geq 10$, estimate $\hat{\lambda}$ on $\mathcal{D}^l_{\text{val}}$ according to Eq. (17); otherwise set $\hat{\lambda} = 0$.
5: **Level:** solve for $b_k$ and $b_n$ separately according to Eq. (18). Types with fewer than 10 valid clusters are not corrected.
6: **Offsets:** obtain the offset $\delta_j = h(\boldsymbol{\psi}_j)$ of each valid cluster from Eq. (9); the remaining clusters take $\delta_j = 0$.
7: **Inference:** for a test sample $x$, compute $\tilde{c}(x)$ according to Eq. (7), clip it according to prediction preservation, and rescale the probabilities of the non-predicted classes proportionally.

Table C1: Dataset splits. $K_k$ and $K_n$ are the numbers of known and novel classes; $|\mathcal{D}_l|$ and $|\mathcal{D}_u|$ are the training-set sizes (thousands of images); the last two columns are the numbers of known-class (i.e., $\mathcal{D}^l_{\text{val}}$) and novel-class images in the test set.

| Dataset | Type | $K_k$ | $K_n$ | $\lvert\mathcal{D}_l\rvert$ | $\lvert\mathcal{D}_u\rvert$ | $\lvert\mathcal{D}^l_{\text{val}}\rvert$ | Novel-class test images |
|---|---|---|---|---|---|---|---|
| CIFAR-10 | generic | 5 | 5 | 12.5 | 37.5 | 5000 | 5000 |
| CIFAR-100 | generic | 80 | 20 | 20.0 | 30.0 | 8000 | 2000 |
| ImageNet-100 | generic | 50 | 50 | 31.9 | 95.3 | 2500 | 2500 |
| CUB-200-2011 | fine-grained | 100 | 100 | 1.5 | 4.5 | 2884 | 2910 |
| Stanford Cars | fine-grained | 98 | 98 | 2.0 | 6.1 | 3948 | 4093 |
| FGVC-Aircraft | fine-grained | 50 | 50 | 1.7 | 5.0 | 1668 | 1665 |
| Flowers-102 | fine-grained | 51 | 51 | 0.3 | 0.8 | 2473 | 3676 |
| Office-Home Art | domain-shifted | 33 | 32 | 0.5 | 1.1 | 492 | 314 |
| Office-Home Clipart | domain-shifted | 33 | 32 | 0.7 | 2.2 | 740 | 711 |
| Office-Home Product | domain-shifted | 33 | 32 | 0.8 | 2.2 | 789 | 691 |
| Office-Home Real-World | domain-shifted | 33 | 32 | 0.8 | 2.1 | 785 | 667 |

# C Experimental Details

This appendix supplements the experimental details omitted from Section 5.

**Data splits.** Table C1 lists the splits of all 11 datasets. Office-Home previously had no single-domain GCD split; we treat each domain as a separate dataset, take the first 33 classes as known classes, split each class into a training set and a test set at a ratio of 2:1 in a stratified manner, and put 50% of the known-class training images into $\mathcal{D}_l$. In addition, we previously evaluated an early version of KNIT on long-tailed CIFAR-100 with class-ordering seed 0 and on Herbarium19; these results are no longer reported. Long-tailed CIFAR-100 is constructed by shuffling the 100 classes with a class-ordering seed; the class ranked $r$ keeps $\lfloor 500\,\rho^{-r/99} \rfloor$ training images, the test set remains balanced, and the split into known and novel classes is the same as for CIFAR-100. Class-ordering seed 1 uses $\rho = 20, 50, 100$, and seed 2 uses $\rho = 50, 100$.

**Clustering accuracy.** Table C2 lists the $\mathcal{D}_u$ clustering accuracy of the first three base methods on each dataset together with the values reported in the original papers, and Table C3 lists the corresponding results of SSR$^2$-GCD.

Table C2: Clustering accuracy on $\mathcal{D}_u$ (All, %) of the first three base methods on each dataset. "Ours" is the accuracy of the models used in the evaluation, and "Original" is the accuracy reported in each method's paper or in the SSR$^2$-GCD paper (He et al., 2026); "–" indicates that no original value is available for comparison.

| Dataset | SimGCD | | GET | | SpectralGCD | |
|---|---|---|---|---|---|---|
| | Ours | Original | Ours | Original | Ours | Original |
| CIFAR-10 | 96.9 | 97.0 | 97.1 | 97.2 | 98.5 | 98.5 |
| CIFAR-100 | 82.4 | 81.1 | 77.0 | 82.1 | 85.4 | 86.1 |
| ImageNet-100 | 90.8 | 90.8 | 91.8 | 91.7 | 92.7 | 93.4 |
| CUB | 70.4 | 71.7 | 75.4 | 77.0 | 79.4 | 79.2 |
| Stanford Cars | 67.2 | 70.0 | 80.1 | 78.5 | 88.3 | 89.1 |
| FGVC-Aircraft | 55.7 | 54.3 | 60.8 | 58.9 | 60.2 | 63.0 |
| Flowers-102 | 82.1 | 75.3 | 84.6 | 85.5 | 86.9 | – |
| Office-Home Art | 62.0 | – | 61.8 | – | 61.2 | – |
| Office-Home Clipart | 52.8 | – | 51.9 | – | 61.6 | – |
| Office-Home Product | 73.1 | – | 75.0 | – | 83.3 | – |
| Office-Home Real-World | 69.1 | – | 72.0 | – | 80.0 | – |

Table C3: Clustering accuracy on $\mathcal{D}_u$ (All, %) of SSR$^2$-GCD on each dataset. The accuracy is computed from the average of the predicted probabilities of the image and text classification heads; "Original" is taken from Tables 1 and 2 of the SSR$^2$-GCD paper (He et al., 2026); "–" indicates that no original value is available for comparison.

| Dataset | Ours | Original |
|---|---|---|
| CIFAR-10 | 98.5 | 98.5 |
| CIFAR-100 | 83.3 | 86.4 |
| ImageNet-100 | 79.8 | 92.1 |
| CUB | 75.6 | 78.3 |
| Stanford Cars | 87.4 | 89.2 |
| FGVC-Aircraft | 47.0 | – |
| Flowers-102 | 84.3 | 93.5 |
| Office-Home Art | 54.0 | – |
| Office-Home Clipart | 62.0 | – |
| Office-Home Product | 79.4 | – |
| Office-Home Real-World | 75.7 | – |

**Implementation of the baselines.** Histogram binning and isotonic regression fit a mapping from confidence to accuracy on $\mathcal{D}^l_{\text{val}}$ and then apply it unchanged to novel classes; they represent flexible mappings fitted on known classes. PseudoCal synthesizes a pseudo-labeled calibration set on $\mathcal{D}^u_{\text{val}}$; it represents calibration that does not rely on target labels. It can only perform mixup at the logit level, so a mixed sample is almost always predicted as the class of the dominant sample; the accuracy of the pseudo calibration set is therefore inflated (0.89–0.995), and the fitted temperature is too low, making all samples overconfident. The uncalibrated output is the probability of Eq. (5) with $\tau_{\text{cal}} = 1$, and temperature scaling is fitted according to Eq. (6). Ensemble temperature scaling (ETS) (Zhang et al., 2020) forms a convex combination of the temperature-scaled probability, the uncalibrated probability, and the uniform distribution:

$$q^{\text{ETS}}(x) = w_1\, p(x;t) + w_2\, p(x;1) + w_3\, \frac{1}{K}\mathbf{1}, \qquad w_1, w_2, w_3 \geq 0, \quad w_1 + w_2 + w_3 = 1, \quad (19)$$

where $p(x;t)$ is the probability vector of Eq. (5) at temperature $t$, and $\mathbf{1}$ is the $K$-dimensional all-ones vector. Following the reference implementation of the original method, $t$ is first fitted within $[0.05, 5]$ with the squared error, and $w$ is then fitted with $t$ fixed; the squared error is taken between $q^{\text{ETS}}(x)$ and the one-hot vector of $y$ and computed on $\mathcal{D}^l_{\text{val}}$, with the same cross-fitting as for TS. On the 18 combinations of the six datasets from CIFAR-10 to FGVC-Aircraft, the $t$ fitted on the whole $\mathcal{D}^l_{\text{val}}$ ranges from 0.517 to 1.722 and differs from $\tau^\star_{\text{cal}}$ of Eq. (6) by $-0.087$ to $+0.206$; $w_2$ is 0 in all cases, $w_3$ is at most 0.022, and $w = (1, 0, 0)$ on 13 of the combinations. The first two components have the same ordering as the logits and the third component is the same for all clusters, so the predicted cluster does not change when $w_1 + w_2 > 0$; ETS is class-agnostic and can be applied directly to novel clusters. Histogram binning uses 15 equal-mass bins on the uncalibrated maximum

class confidence, and isotonic regression is fitted on the same score; both keep the predicted cluster unchanged. PseudoCal performs mixup at the logit level on $\mathcal{D}^u_{\text{val}}$ (mixing coefficient 0.65) and then fits the temperature with the negative log-likelihood. All calibration methods use the same cross-fitting scheme and the same predictions.

**Text features.** The text prompt is "a photo of a $v$." To save storage, only the 20 concepts with the highest similarity are kept for each image, and the similarities of the remaining concepts are replaced by the value of the 20th-ranked concept.

**Ablation variants.** In Table 2, the variants that remove relative confidence, that remove Margin or Purity, and that keep only relative confidence all refit the shared coefficients on the known clusters of the 18 combinations. The variants that remove cluster regularization and that remove the level constants reuse the coefficients of KNIT. In the two-stage paired bootstrap pooled over the 44 combinations, the $\text{cluECE}_{\text{new}}$ of every variant is significantly worse than that of the full KNIT, except for removing relative confidence ($+0.07$).

**Per-combination results.** Tables C4 to C6 give the cluECE of all calibration methods on each dataset of generic images, domain-shifted images, and fine-grained images, respectively; the value of each image category in Table 1 is the arithmetic mean over the corresponding datasets. Table C7 gives the difference between KNIT and TS on each combination together with its significance, and Table C8 summarizes the per-combination comparison of each calibration method with TS.

**Sensitivity.** Besides the default setting, we examine 16 settings: the width (0.25–$\infty$), center, and truncation of the prior on $\lambda$; the number of concepts kept in the vocabulary (5–20); truncation of the semantic purity measures; the minimum number of valid clusters per type (5–15); and the minimum number of samples per cluster (3–10). The $\text{cluECE}_{\text{new}}$ over the 44 combinations lies between 13.39 and 13.82, and for 14 of the settings it differs from that of the default setting (13.44) by less than 0.15. The two exceptions both come from the width of the prior on $\lambda$: a narrower prior ($\sigma = 0.25$) and the widest prior ($\sigma = \infty$) raise $\text{cluECE}_{\text{new}}$ to 13.64 and 13.82, respectively. A prior centered at a negative value gives 13.50, which is not better than the default setting, and it is 0.28 worse than the default setting on long-tailed CIFAR-100 with class-ordering seed 1. Selecting the best setting per dataset using the test labels gives 13.19. On the 44 combinations, no dataset or base method has its average $\text{cluECE}_{\text{new}}$ made worse by KNIT (Table C7). When the minimum number of valid clusters per type is lowered to 5, CIFAR-10 is corrected, and its $\text{cluECE}_{\text{new}}$ is 1.75, versus 1.72 for TS. All calibration quantities of KNIT are obtained by direct computation, low-dimensional convex optimization, and one-dimensional root finding.

**Comparison of ECE and cluECE.** Table C9 gives the average ECE and $\text{cluECE}_{\text{new}}$ of each calibration method for SimGCD and GET on the 11 datasets, corresponding to the analysis in Section 5.3.

**Overall level of each kind of cluster on long-tailed data.** Table C10 gives the accuracy and average confidence of each kind of cluster after TS and after KNIT on long-tailed CIFAR-100 with class-ordering seed 1. After TS, novel-type clusters are underconfident overall by 21.32 percentage points, in all 8 combinations without exception, far more than known-type clusters (8.64), whereas KNIT changes the average confidence of each of the two types of clusters by only about 1 percentage point, so this overall deviation is largely retained; the novel clusters partitioned by $\pi$ also contain overconfident known-type clusters, and the two offset each other, so that these clusters are underconfident overall by only 1.29.

Table C4: cluECE (%, lower is better) of each calibration method on each generic-image dataset. Bold marks the column-wise minimum within each base method.

| Base method | Calibration | CIFAR-10 | | | CIFAR-100 | | | ImageNet-100 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | All | Old | New | All | Old | New | All | Old | New |
| SimGCD (Wen et al., 2023) | Uncalibrated | 2.2 | 2.1 | 2.2 | 24.1 | 21.6 | 35.9 | 7.6 | **2.8** | 12.3 |
| | TS (Guo et al., 2017) | 1.7 | 1.5 | 1.8 | 6.4 | 6.1 | 7.8 | 7.2 | 2.9 | 11.6 |
| | ETS (Zhang et al., 2020) | 2.3 | 2.0 | 2.5 | 6.5 | 5.6 | 10.7 | 9.3 | 3.0 | 15.6 |
| | Histogram binning (Zadrozny & Elkan, 2001) | 1.9 | 2.0 | 1.9 | 6.9 | **5.4** | 13.9 | 9.4 | **2.8** | 16.1 |
| | Isotonic regression (Zadrozny & Elkan, 2002) | **1.2** | **1.2** | **1.3** | 6.9 | **5.4** | 14.0 | 9.3 | **2.8** | 15.7 |
| | PseudoCal (Hu et al., 2024) | 3.2 | 3.2 | 3.2 | 15.1 | 16.1 | 10.7 | 9.1 | 5.4 | 12.8 |
| | KNIT | 1.7 | 1.5 | 1.8 | **5.6** | **5.4** | **6.7** | **5.8** | 2.9 | **8.7** |
| GET (Wang et al., 2025) | Uncalibrated | 2.0 | 1.9 | 2.1 | 20.7 | 18.8 | 31.6 | 6.5 | 2.7 | 10.3 |
| | TS | 1.7 | **1.6** | 1.7 | 9.2 | 7.6 | 17.9 | 5.3 | 2.6 | 8.0 |
| | ETS | 2.1 | 2.1 | 2.0 | 9.1 | 7.3 | 19.2 | 7.5 | 2.9 | 12.0 |
| | Histogram binning | 1.6 | 1.9 | 1.4 | 9.1 | **7.2** | 19.6 | 8.3 | **2.5** | 14.1 |
| | Isotonic regression | **1.4** | **1.6** | **1.2** | 9.2 | 7.3 | 19.9 | 8.2 | 2.7 | 13.6 |
| | PseudoCal | 3.0 | 2.9 | 3.0 | 17.7 | 17.7 | 17.4 | 6.9 | 4.4 | 9.4 |
| | KNIT | 1.7 | **1.6** | 1.7 | **8.7** | 7.5 | **15.1** | **5.0** | 2.7 | **7.3** |
| SpectralGCD (Caselli et al., 2026) | Uncalibrated | 2.3 | 2.3 | 2.4 | 24.4 | 21.9 | 36.0 | 6.7 | 2.5 | 11.0 |
| | TS | 2.2 | 2.0 | 2.4 | 6.6 | 5.3 | 12.7 | 5.3 | **2.3** | 8.3 |
| | ETS | 2.9 | 2.7 | 3.2 | 6.7 | 5.3 | 13.5 | 6.0 | 2.4 | 9.7 |
| | Histogram binning | 2.1 | 2.0 | 2.1 | 7.1 | 5.5 | 14.5 | 6.5 | 2.4 | 10.7 |
| | Isotonic regression | **1.8** | **1.7** | **1.8** | 7.1 | 5.4 | 14.6 | 6.0 | **2.3** | 9.7 |
| | PseudoCal | 3.0 | 3.0 | 3.0 | 14.7 | 15.4 | 11.7 | 7.3 | 4.0 | 10.5 |
| | KNIT | 2.2 | 2.0 | 2.4 | **5.8** | **4.7** | **11.0** | **4.9** | 2.5 | **7.3** |
| SSR$^2$-GCD (He et al., 2026) | Uncalibrated | 2.0 | 2.7 | 1.2 | 6.7 | 5.2 | 13.5 | 7.3 | 4.6 | 9.9 |
| | TS | 1.6 | 2.2 | 1.0 | 5.5 | **4.1** | 11.3 | 7.5 | 5.1 | 9.9 |
| | ETS | 1.4 | 1.6 | 1.2 | 5.5 | 4.2 | 10.9 | 7.5 | 4.5 | 10.5 |
| | Histogram binning | **1.2** | **1.2** | 1.3 | **5.1** | 4.3 | **8.8** | 7.5 | 4.5 | 10.4 |
| | Isotonic regression | **1.2** | **1.2** | 1.2 | **5.1** | 4.3 | **8.7** | 7.4 | 4.4 | 10.3 |
| | PseudoCal | **1.2** | 1.6 | **0.9** | 7.7 | 7.4 | 9.2 | 8.1 | **3.5** | 12.6 |
| | KNIT | 1.6 | 2.2 | 1.0 | 5.5 | 4.3 | 10.9 | **6.1** | 3.9 | **8.1** |

Table C5: cluECE (%, lower is better) of each calibration method on domain-shifted images (Office-Home domains), in the same format as Table C4.

| Base method | Calibration | Art | | | Clipart | | | Product | | | Real-World | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | All | Old | New | All | Old | New | All | Old | New | All | Old | New |
| SimGCD (Wen et al., 2023) | Uncalibrated | 29.0 | **25.6** | 31.8 | 26.6 | **18.5** | 32.2 | 20.0 | 12.4 | 26.1 | 18.6 | **14.0** | 22.3 |
| | TS | 28.1 | 32.3 | 24.7 | 20.4 | 21.5 | 19.6 | 21.2 | 15.8 | 25.4 | 19.1 | 17.5 | 20.4 |
| | (Guo et al., 2017) | | | | | | | | | | | | |
| | ETS | 28.4 | 33.9 | 24.0 | 20.3 | 21.6 | 19.4 | 22.5 | 18.8 | 25.5 | 20.5 | 20.1 | 20.8 |
| | (Zhang et al., 2020) | | | | | | | | | | | | |
| | Histogram binning | 28.5 | 32.9 | 24.9 | 19.3 | 20.3 | **18.5** | 22.9 | 19.0 | 26.0 | 20.5 | 20.8 | 20.3 |
| | (Zadrozny & Elkan, 2001) | | | | | | | | | | | | |
| | Isotonic regression | 28.3 | 33.0 | 24.7 | 19.3 | 20.1 | 18.7 | 22.8 | 18.6 | 26.2 | 20.3 | 20.3 | 20.3 |
| | (Zadrozny & Elkan, 2002) | | | | | | | | | | | | |
| | PseudoCal | 42.2 | 26.4 | 54.9 | 43.8 | 28.0 | 54.7 | 23.7 | **11.1** | 33.7 | 27.0 | 14.5 | 36.7 |
| | (Hu et al., 2024) | | | | | | | | | | | | |
| | KNIT | **26.6** | 30.7 | **23.3** | **18.6** | 18.7 | 18.6 | **18.2** | 14.1 | **21.5** | **17.6** | 16.5 | **18.5** |
| GET (Wang et al., 2025) | Uncalibrated | 30.1 | **27.1** | 32.3 | 26.7 | **18.0** | 32.2 | 20.6 | 11.4 | 27.5 | 16.5 | 13.4 | 19.1 |
| | TS | 29.7 | 30.8 | 28.8 | 21.3 | 20.3 | 22.0 | 21.1 | 13.3 | 27.0 | 16.4 | 13.3 | 19.2 |
| | ETS | 31.7 | 36.1 | 28.3 | 21.1 | 22.3 | 20.4 | 23.1 | 17.6 | 27.3 | 17.3 | 15.3 | 19.0 |
| | Histogram binning | 31.3 | 35.7 | 27.9 | 20.1 | 20.7 | **19.7** | 23.9 | 18.6 | 27.8 | 17.0 | 15.1 | 18.6 |
| | Isotonic regression | 31.6 | 35.5 | 28.6 | **19.8** | 20.0 | **19.7** | 23.4 | 18.2 | 27.3 | 17.0 | 15.1 | 18.7 |
| | PseudoCal | 45.1 | 31.5 | 55.6 | 44.2 | 26.2 | 55.7 | 23.7 | **7.5** | 35.9 | 22.4 | **11.0** | 32.3 |
| | KNIT | **27.8** | 28.3 | **27.5** | 20.2 | 18.2 | 21.5 | **17.9** | 11.4 | **22.8** | **15.7** | 12.9 | **18.0** |
| SpectralGCD (Caselli et al., 2026) | Uncalibrated | 25.4 | **25.5** | 25.4 | 20.8 | **18.5** | 22.5 | 17.7 | 12.7 | 22.1 | 14.9 | 12.4 | 17.3 |
| | TS | 24.6 | 29.4 | 20.6 | 20.2 | 21.0 | 19.6 | 16.7 | 10.8 | 21.9 | 14.5 | 11.8 | 17.1 |
| | ETS | 24.9 | 30.3 | 20.2 | 20.2 | 21.1 | 19.5 | 17.7 | 12.6 | 22.0 | 15.6 | 13.4 | 17.6 |
| | Histogram binning | 25.5 | 31.5 | 20.4 | 20.3 | 20.9 | 19.7 | 17.4 | 12.9 | 21.4 | 15.7 | 13.7 | 17.5 |
| | Isotonic regression | 25.0 | 30.9 | **20.0** | 20.1 | 20.7 | 19.7 | 17.5 | 12.7 | 21.6 | 15.7 | 13.5 | 17.8 |
| | PseudoCal | 39.7 | 27.9 | 49.7 | 34.1 | 23.8 | 42.2 | 15.7 | **4.0** | 25.7 | 17.6 | **9.9** | 24.7 |
| | KNIT | **23.3** | 27.0 | 20.1 | **18.2** | 19.3 | **17.3** | **13.9** | 9.7 | **17.5** | **13.2** | 11.9 | **14.4** |
| SSR$^2$-GCD (He et al., 2026) | Uncalibrated | 23.4 | 20.7 | 26.4 | 16.9 | **13.7** | 19.9 | 14.2 | 10.1 | 18.4 | **13.5** | 11.2 | **15.7** |
| | TS | 22.4 | 23.6 | **21.1** | 16.8 | 15.8 | 17.7 | 15.2 | 12.5 | 18.0 | 16.9 | 16.5 | 17.2 |
| | ETS | 23.5 | 24.9 | 22.0 | 17.2 | 14.9 | 19.3 | 15.5 | 12.3 | 18.9 | 15.9 | 15.3 | 16.5 |
| | Histogram binning | 24.4 | 25.7 | 23.0 | 18.0 | 14.0 | 21.8 | 16.1 | 12.3 | 20.1 | 16.2 | 14.9 | 17.5 |
| | Isotonic regression | 23.5 | 25.0 | 21.9 | 17.6 | 14.1 | 20.9 | 15.9 | 12.4 | 19.7 | 16.1 | 14.5 | 17.6 |
| | PseudoCal | 29.9 | **19.6** | 41.3 | 23.2 | 14.2 | 31.7 | 14.2 | **7.4** | 21.3 | 15.5 | **8.5** | 22.0 |
| | KNIT | **21.5** | 21.1 | 21.9 | **15.3** | 15.3 | **15.4** | **13.7** | 11.5 | **16.0** | 15.0 | 13.3 | 16.6 |

Table C6: cluECE (%, lower is better) of each calibration method on each fine-grained-image dataset, in the same format as Table C4.

| Base method | Calibration | CUB | | | Flowers-102 | | | FGVC-Aircraft | | | Stanford Cars | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | All | Old | New | All | Old | New | All | Old | New | All | Old | New |
| SimGCD (Wen et al., 2023) | Uncalibrated | 27.3 | 30.2 | 24.9 | 22.2 | 23.3 | 21.6 | 22.7 | 13.7 | 29.0 | 15.0 | 14.6 | 15.2 |
| | TS | 16.7 | 13.4 | 19.5 | 12.8 | 10.9 | 13.9 | 17.7 | 14.0 | 20.2 | 16.0 | 9.3 | 21.2 |
| | (Guo et al., 2017) | | | | | | | | | | | | |
| | ETS | 16.9 | 13.7 | 19.5 | 12.6 | 10.3 | 13.9 | 17.0 | 16.1 | 17.5 | 16.0 | 10.9 | 19.9 |
| | (Zhang et al., 2020) | | | | | | | | | | | | |
| | Histogram binning | 17.8 | 14.7 | 20.4 | 11.9 | 10.0 | 13.0 | 15.5 | 14.6 | **16.0** | 15.4 | 10.8 | 18.9 |
| | (Zadrozny & Elkan, 2001) | | | | | | | | | | | | |
| | Isotonic regression | 17.6 | 14.5 | 20.2 | 11.8 | 9.7 | 13.0 | 15.4 | 14.5 | 16.1 | 15.3 | 10.7 | 18.9 |
| | (Zadrozny & Elkan, 2002) | | | | | | | | | | | | |
| | PseudoCal | 25.8 | 16.5 | 33.7 | 16.1 | **7.2** | 21.3 | 40.7 | 23.7 | 52.5 | 28.5 | 7.0 | 45.1 |
| | (Hu et al., 2024) | | | | | | | | | | | | |
| | KNIT | **14.3** | **12.1** | **16.1** | **10.9** | 10.8 | **10.9** | **14.3** | **11.8** | **16.0** | **11.2** | **6.6** | **14.7** |
| GET (Wang et al., 2025) | Uncalibrated | 30.3 | 31.7 | 29.0 | 23.5 | 24.7 | 22.9 | 19.9 | 13.3 | 24.7 | 17.2 | 13.8 | 20.1 |
| | TS | 12.5 | 9.8 | 14.9 | 13.7 | 13.2 | 13.9 | 17.8 | 14.3 | 20.4 | 13.3 | 7.9 | 18.0 |
| | ETS | 12.2 | 9.2 | 14.8 | 14.2 | 14.1 | 14.3 | 17.8 | 15.6 | 19.4 | 12.9 | 7.3 | 17.8 |
| | Histogram binning | 13.2 | 10.1 | 16.0 | 14.3 | 14.6 | 14.2 | 16.9 | 13.9 | 19.1 | 13.2 | 7.6 | 18.2 |
| | Isotonic regression | 13.2 | 10.0 | 16.1 | 14.2 | 14.5 | 14.0 | 16.9 | 13.9 | 19.0 | 13.3 | 7.5 | 18.3 |
| | PseudoCal | 19.9 | 12.7 | 26.3 | 14.5 | **6.8** | 18.6 | 36.8 | 23.2 | 46.6 | 17.7 | 7.1 | 27.0 |
| | KNIT | **10.7** | **8.6** | **12.5** | **12.4** | 13.0 | **12.0** | **13.9** | **12.6** | **14.8** | **8.4** | **6.6** | **10.0** |
| SpectralGCD (Caselli et al., 2026) | Uncalibrated | 29.1 | 30.0 | 28.4 | 22.9 | 24.4 | 22.1 | 18.3 | 13.3 | 22.3 | 32.0 | 30.9 | 33.0 |
| | TS | 14.0 | 11.5 | 16.1 | 9.9 | 6.8 | 11.7 | 16.6 | 15.3 | 17.6 | 10.4 | 5.8 | 14.7 |
| | ETS | 12.2 | **9.1** | 14.8 | 9.8 | 6.5 | 11.7 | 16.5 | 14.6 | 18.1 | 9.3 | **4.9** | 13.4 |
| | Histogram binning | 12.9 | 9.9 | 15.5 | 9.4 | 5.7 | 11.5 | 15.4 | 13.2 | 17.2 | 10.1 | 5.7 | 14.2 |
| | Isotonic regression | 12.9 | 9.7 | 15.5 | **9.3** | 5.6 | 11.5 | 15.4 | **13.1** | 17.2 | 10.1 | 5.7 | 14.1 |
| | PseudoCal | 17.0 | 11.1 | 22.1 | 12.3 | **3.7** | 17.3 | 39.6 | 31.5 | 46.1 | 12.9 | 6.8 | 18.5 |
| | KNIT | **11.8** | 10.2 | **13.2** | 9.5 | 6.8 | **11.1** | **15.2** | **13.1** | **16.8** | **5.8** | 5.5 | **6.0** |
| SSR$^2$-GCD (He et al., 2026) | Uncalibrated | 13.2 | 12.2 | 14.1 | 14.2 | 13.2 | 14.7 | 16.5 | 15.5 | 17.5 | 17.2 | 15.3 | 19.0 |
| | TS | 13.2 | 12.2 | 14.2 | 8.0 | 6.3 | 9.1 | 12.6 | 13.4 | 11.8 | 9.4 | 6.9 | 11.8 |
| | ETS | 9.8 | 8.3 | 11.2 | 8.0 | 6.2 | 9.1 | 12.6 | 13.1 | 12.0 | **8.1** | 4.8 | 11.3 |
| | Histogram binning | **9.4** | 8.0 | **10.6** | 8.4 | 6.2 | 9.8 | 12.4 | 12.8 | 11.9 | 8.9 | 5.3 | 12.3 |
| | Isotonic regression | **9.3** | **8.0** | 10.7 | 8.2 | 6.1 | 9.4 | 12.3 | 12.6 | 12.0 | 8.7 | 5.3 | 12.0 |
| | PseudoCal | 15.5 | 13.8 | 17.0 | **7.3** | **4.1** | 9.3 | 36.3 | 33.3 | 39.2 | 8.3 | **4.0** | 12.5 |
| | KNIT | 12.1 | 10.2 | 13.9 | 7.9 | 6.2 | **9.0** | **11.2** | **11.5** | **10.9** | 8.2 | 5.7 | **10.7** |

Table C7: Difference in cluECE between KNIT and TS (KNIT − TS, %, negative values indicate KNIT is better) and the number of valid clusters, for each combination. * indicates that the 95% two-stage paired bootstrap interval of this difference does not contain 0. A valid cluster is a known cluster or a novel cluster (Section B) with at least 5 samples in $\mathcal{D}_u$; on CIFAR-10, both types of valid clusters number fewer than 10, and KNIT applies no correction and is identical to TS.

| Dataset | SimGCD | | | GET | | | SpectralGCD | | | SSR$^2$-GCD | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | New | Old | Valid clusters (Old / New) | New | Old | Valid clusters (Old / New) | New | Old | Valid clusters (Old / New) | New | Old | Valid clusters (Old / New) |
| CIFAR-10 | 0.00 | 0.00 | 5 / 5 | 0.00 | 0.00 | 5 / 5 | 0.00 | 0.00 | 5 / 5 | 0.00 | 0.00 | 5 / 5 |
| CIFAR-100 | −1.15 | −0.65 | 80 / 20 | −2.81* | −0.13 | 80 / 18 | −1.64 | −0.56 | 80 / 19 | −0.48 | +0.13 | 80 / 20 |
| ImageNet-100 | −2.89* | +0.03 | 50 / 50 | −0.62 | +0.07 | 50 / 50 | −0.98* | +0.15 | 50 / 50 | −1.73 | −1.20 | 50 / 50 |
| CUB | −3.40* | −1.36* | 99 / 93 | −2.39* | −1.23* | 98 / 97 | −2.88* | −1.30* | 100 / 97 | −0.30 | −1.96* | 100 / 100 |
| Stanford Cars | −6.45* | −2.70* | 97 / 59 | −8.07* | −1.28 | 98 / 85 | −8.68* | −0.31 | 98 / 90 | −1.10 | −1.22* | 98 / 98 |
| FGVC-Aircraft | −4.19* | −2.21* | 50 / 49 | −5.63* | −1.66 | 50 / 45 | −0.80 | −2.23* | 50 / 47 | −0.87 | −1.94* | 50 / 50 |
| Flowers-102 | −3.03* | −0.02 | 25 / 49 | −1.92* | −0.14 | 22 / 50 | −0.63 | +0.05 | 25 / 50 | −0.10 | −0.06 | 26 / 50 |
| Office-Home Art | −1.43 | −1.65 | 21 / 32 | −1.30 | −2.53* | 21 / 32 | −0.57 | −2.34 | 22 / 32 | +0.78 | −2.51* | 33 / 32 |
| Office-Home Clipart | −1.01 | −2.85 | 31 / 32 | −0.44 | −2.09 | 32 / 32 | −2.30 | −1.69 | 33 / 32 | −2.31* | −0.51 | 33 / 32 |
| Office-Home Product | −3.88* | −1.78 | 31 / 32 | −4.25* | −1.93 | 33 / 32 | −4.38* | −1.09 | 33 / 32 | −1.95* | −0.96* | 33 / 32 |
| Office-Home Real-World | −1.93 | −1.00 | 31 / 32 | −1.12 | −0.33 | 32 / 32 | −2.73* | +0.17 | 32 / 32 | −0.60 | −3.25* | 33 / 32 |

Table C8: Per-combination comparison of each calibration method against TS over the 44 combinations (number of combinations). "Better" means that cluECE is smaller than that of TS; "significantly" means that the 95% two-stage paired bootstrap interval of the difference between the two does not contain 0.

| Method | $\text{cluECE}_{\text{new}}$ | | | $\text{cluECE}_{\text{old}}$ | | |
|---|---|---|---|---|---|---|
| | Better than TS | Significantly better | Significantly worse | Better than TS | Significantly better | Significantly worse |
| ETS | 21 | 7 | 10 | 19 | 10 | 18 |
| Histogram binning | 19 | 6 | 9 | 22 | 8 | 10 |
| Isotonic regression | 21 | 8 | 9 | 26 | 10 | 10 |
| PseudoCal | 4 | 0 | 26 | 22 | 12 | 13 |
| KNIT | 39 | 19 | 0 | 34 | 13 | 0 |

Table C9: Average ECE and $\text{cluECE}_{\text{new}}$ (%) of each calibration method for SimGCD and GET on the 11 datasets (22 combinations). ECE is computed on the samples whose true class is a novel class, a sample set different from that of $\text{cluECE}_{\text{new}}$ (Section A).

| Calibration | ECE | $\text{cluECE}_{\text{new}}$ |
|---|---|---|
| Uncalibrated | 13.08 | 22.97 |
| TS (Guo et al., 2017) | 6.34 | 17.18 |
| ETS (Zhang et al., 2020) | 7.25 | 17.45 |
| Histogram binning (Zadrozny & Elkan, 2001) | 8.26 | 17.57 |
| Isotonic regression (Zadrozny & Elkan, 2002) | 7.90 | 17.52 |
| PseudoCal (Hu et al., 2024) | 25.99 | 31.23 |
| KNIT | 6.83 | 14.54 |

Table C10: Accuracy and average confidence of each kind of cluster on long-tailed CIFAR-100 (class-ordering seed 1, $\rho = 20, 50, 100$, 8 combinations) (%, computed over test samples, averaged over the 8 combinations). The type of a cluster is determined by whether the corresponding output unit is supervised by labels during training (Section B), and known clusters and novel clusters are partitioned by $\pi$ (Section 3.1); the accuracy uses novel-class labels and is for diagnosis only. The last column is the number of combinations in which the average confidence after TS is lower than the accuracy.

| Cluster | Accuracy | Average confidence | | Average confidence − accuracy | | Underconfident combinations / 8 |
|---|---|---|---|---|---|---|
| | | TS | KNIT | TS | KNIT | |
| Novel-type clusters | 58.57 | 37.25 | 38.48 | −21.32 | −20.09 | 8 |
| Known-type clusters | 52.74 | 44.10 | 45.09 | −8.64 | −7.65 | 8 |
| Novel clusters ($\pi$) | 41.83 | 40.53 | 39.27 | −1.29 | −2.56 | 5 |
| of which novel-type clusters | 63.57 | 38.64 | 38.40 | −24.93 | −25.17 | 8 |
| of which known-type clusters | 31.75 | 41.38 | 39.53 | +9.63 | +7.77 | 2 |
| Known clusters ($\pi$) | 56.59 | 44.16 | 45.76 | −12.43 | −10.82 | 8 |