# RareGCD: Toward Rare Disease Discovery via Generalized Category Discovery

Yuan Ma[1*], Wei Feng[3*], Junlin Hou[4], Chao Zhang[5], Zongyuan Ge[2], Haoran Xie[1✉], and Lie Ju[2,3✉]

[1] Japan Advanced Institute of Science and Technology
[2] University College London
[3] Monash University
[4] The Hong Kong University of Science and Technology
[5] University of Toyama
[*]Equal contribution
{s2420410,xie}@jaist.ac.jp, lie.ju@ucl.ac.uk

**Abstract.** Generalized Category Discovery (GCD) aims to identify known disease categories while discovering unseen ones in medical images, without relying on exhaustive manual annotations. However, existing GCD methods for medical imaging are typically evaluated under balanced settings, overlooking a critical characteristic of practical scenarios: novel categories often correspond to rare diseases and exhibit severe long-tailed distributions. To address this limitation, we propose RareGCD, a framework tailored for long-tailed medical data. Specifically, RareGCD leverages labeled data from known categories as a strong prior to guide sample association, enabling reliable grouping of unlabeled instances under explicit category constraints. Furthermore, we propose a prototypical contrast learning strategy with density-aware instance-wise temperature scaling, which strengthens contrast supervision for core samples and reduces over-confident updates for unstable samples, improving representation stability under imbalance. To better reflect practical scenarios, we construct long-tailed benchmarks on multiple medical datasets and evaluate RareGCD under both standard and long-tailed settings. Extensive experiments demonstrate that RareGCD consistently outperforms state-of-the-art methods, particularly in recognizing novel disease categories. Code will be made publicly available at https://github.com/myyy777/RareGCD.

**Keywords:** GCD · Rare disease discovery · Medical image analysis.

## 1 Introduction

Deep learning has achieved remarkable success in medical image analysis [1], enabling accurate and efficient pattern recognition for diagnosis and clinical decision support. However, most existing models are developed under a closed-set assumption, where the category space is fixed and identical during training and deployment [20, 17]. In practice, medical image datasets are inherently dynamic

and evolving, with new disease types or imaging patterns emerging over time. In such cases, conventional models tend to misclassify previously unseen categories as known ones, leading to unreliable predictions [19, 15].

Therefore, Generalized Category Discovery (GCD) [21, 6, 8, 7, 10] was introduced to recognize seen categories and cluster novel categories from unlabeled data. Although GCD was initially developed for natural images, recent studies have extended it to medical image analysis [11, 5]. For example, Das et al. [5] proposed MedGCD, which identifies seen categories from labeled data while discovering novel categories from unlabeled data using a dual-view framework with a confidence-aware objective. Existing methods [5, 6, 9] optimize labeled and unlabeled data through separate objectives, resulting in asymmetric supervision strength that limits their ability to reliably discover new disease types. Moreover, these methods are commonly evaluated under relatively balanced settings. However, medical image datasets are inherently long-tailed, where novel categories often correspond to rare diseases with limited samples. Under long-tailed distributions, the supervision gap between labeled and unlabeled data becomes more pronounced. As a result, novel categories with limited samples receive insufficient learning signals, which undermines reliable discovery.

In this paper, we propose RareGCD, a framework that combines representation learning based on prototype-guided sample association with decision-boundary refinement stage, aiming to reliably recognize seen categories and facilitate the discovery of novel rare disease. To mitigate the supervision asymmetry between labeled and unlabeled data, RareGCD first leverages labeled samples as a prior to guide sample association. Specifically, instance relationships are established based on feature similarity, with labeled prototypes serving as prior guidance. All associations are strictly constrained by seen category labels, preventing samples from different seen classes from being merged. With the semantic groups obtained from the association stage, we adopt density-aware prototypical contrast learning to strengthen the representation. By adaptively adjusting instance-wise temperature according to local feature density, the objective strengthens supervision for core samples and mitigates over-confident for unstable samples, improving stability under long-tailed distributions. After an initial training stage with prototype-guided association, we incorporate a parametric classification head for subsequent joint optimization to establish explicit decision boundaries. Extensive experiments on PathMNIST, OrganAMNIST, and BloodMNIST [27] demonstrate that RareGCD achieves superior performance over state-of-the-art methods. By additionally constructing long-tailed data distributions, we further show that RareGCD maintains clear performance advantages under class imbalance, particularly on novel categories.

## 2 Method

### 2.1 Problem Formulation and Framework Overview

We consider the GCD setting for medical image analysis, where a dataset comprises a labeled subset from seen categories and a larger unlabeled subset that

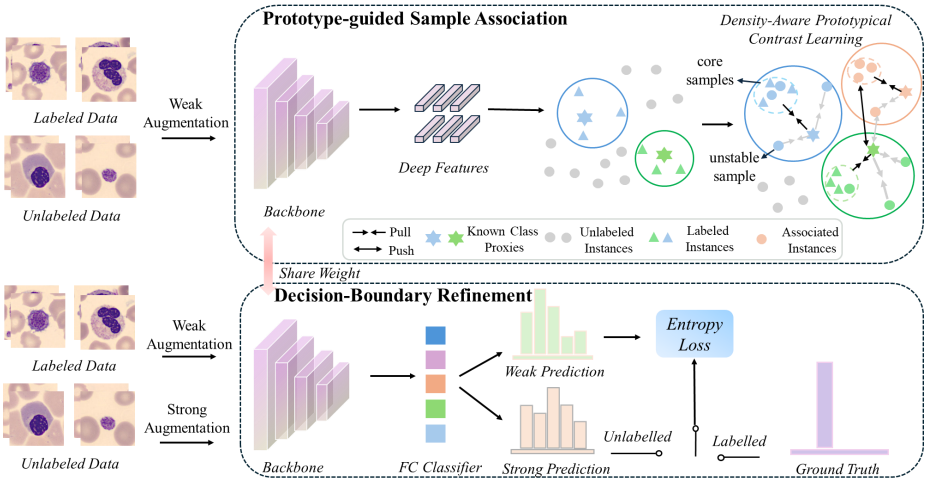


Fig. 1: Overview framework of the proposed RareGCD.

may contain both seen and novel categories. The goal is to jointly recognize seen categories and discover novel categories by clustering the unlabeled data. Specifically, the dataset $\mathcal{D}$ is composed of two parts: $\mathcal{D}_{\mathcal{L}} = \{(x_i, y_i)\}_{i=1}^{N} \subset \mathcal{X} \times \mathcal{Y}_{\mathcal{L}}$, $\mathcal{D}_{\mathcal{U}} = \{x_i\}_{i=1}^{M} \subset \mathcal{X}$, where $\mathcal{D}_{\mathcal{L}}$ denotes the labeled subset of $N$ medical images with known labels $\mathcal{Y}_{\mathcal{L}}$, and $\mathcal{D}_{\mathcal{U}}$ denotes the unlabeled subset of $M$ medical images. The unlabeled data may contain samples from both known and unknown categories, with the label space satisfying $\mathcal{Y}_{\mathcal{L}} \subset \mathcal{Y}_{\mathcal{U}}$. The goal is to correctly recognize known categories and cluster unlabeled samples into semantically meaningful groups.

As shown in Fig. 1, RareGCD consists of three components: prototype-guided sample association, density-aware prototypical contrastive learning, and decision-boundary refinement. The framework first leverages labeled prototypes to guide sample association, forming semantically consistent groups while preserving label consistency among seen categories. Based on these groups, density-aware prototypical contrastive learning refines representations and stabilizes prototype updates under long-tailed distributions. Finally, a parametric classification head is introduced for decision-boundary refinement.

### 2.2 Prototype-guided Sample Association

As shown in Fig. 1, our aim is to encourage semantically similar samples to be grouped together at association stage, while limiting each group to at most one known class. To this end, we first extract deep features for all training samples using a backbone network. For labeled samples, we compute the mean feature of each category and treat it as a known class proxy, which serves as a semantic prior for subsequent association. To capture local relationships in the embedding space, we perform KNN [3] search for each sample over the joint set of prototypes

and instance features. To reduce asymmetric or noisy connections, we retain only reciprocal KNN relations, where two samples are considered neighbors if each lies within the KNN set of the other. To further refine the associations, we compute the Jaccard distance [29] between neighborhood sets to measure structural consistency of two samples and retain only those pairs with high neighborhood agreement. Those kept pairs are regarded as the association candidates. Finally, candidate pairs are processed sequentially, and two instances are merged into the same group only if their association does not introduce a conflict with known category labels. Specifically, associations that would merge samples from different known categories are discarded. This procedure incrementally refines the grouping results while strictly preserving the ground-truth class relations of labeled data.

### 2.3 Density-Aware Prototypical Contrast Learning

After association, we construct a prototype memory, where each prototype represents the feature centroid of a semantic group. Prototypical contrast learning paradigm [25] is commonly employed to learn discriminative representations at the semantic group level. Conventional prototypical contrast learning adopts a shared temperature, implicitly assigning equal contribution to all instances. Under long-tailed distributions, this uniform treatment may amplify the influence of less reliable samples and increase the risk that rare or unstable groups are biased toward dominant categories.

Motivated by this observation, we propose a density-aware prototypical contrast learning that assigns instance-wise temperatures based on cluster-adaptive reliability estimation, enabling robust prototype discrimination under long-tailed medical data distributions. Given an input image $x_i$, we extract a $\ell_2$-normalized feature representation $z_i$ and contrast it against a prototype memory $\mathbf{E} \in R^{C\times d}$. $C$ is the number of the clusters and $c_i \in C$ denotes the cluster assignment of sample $x_i$. The proposed loss is formulated as:

$$\mathcal{L}_{\text{rep}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\big(\mathbf{E}[c_i]^\top z_i/\tau_i\big)}{\sum_{j=1}^{C}\exp(\mathbf{E}[j]^\top z_i/\tau_i)}, \tag{1}$$

where $\tau_i$ denotes an instance-wise temperature that controls the strength of prototype-based discrimination for sample $x_i$. Specifically, at epoch $t$, we compute an initial reliability score $\hat{h}_i^t$ by comparing the similarity mass of the most relevant neighbors to that of all other samples in the same cluster:

$$\hat{h}_i^t := \frac{\sum_{z_j\in\hat{\mathcal{Z}}_i^{\text{topQ}}}\exp\big(z_i^\top z_j\big)}{\sum_{z_j\in\hat{\mathcal{Z}}_i}\exp\big(z_i^\top z_j\big)}, \tag{2}$$

where $\hat{\mathcal{Z}}_i^{\text{top}Q}$ denotes the set of the $Q\%$ most similar representations to $z_i$ in the same cluster. A lower reliability score indicates that the unstable sample

has weaker local support within its cluster and is less representative of the corresponding prototype. The clipped score is then normalized in a cluster-wise manner to the range $[\tau^{\min}, \tau^{\max}]$, yielding an instance-wise temperature $\tau_i$:

$$\tau_i = \tau^{\max} - \frac{\hat{h}_i^t - \min\left(\mathcal{H}_{c_i}^t\right)}{\max\left(\mathcal{H}_{c_i}^t\right) - \min\left(\mathcal{H}_{c_i}^t\right)} \left(\tau^{\max} - \tau^{\min}\right), \tag{3}$$

where $\mathcal{H}_{c_i}^t$ denotes the set of reliability scores for all instances belonging to cluster $c_i$ at epoch $t$.

After each batch forward pass, the proxy memory features are updated in a moving-average manner [26]: $\mathbf{E}[c_i] \leftarrow \mu\,\mathbf{E} + \left(1-\mu\right) z_i$, where $\mu \in [0,1]$ is an updating rate. To encourage a mutually reinforcing interaction between representation learning and data association, we perform these two processes in an alternating manner. As representations become more discriminative, data association becomes more accurate, which in turn provides cleaner supervision for subsequent representation learning.

### 2.4 Decision-Boundary Refinement

To explicitly refine decision boundaries, we adopt a parametric classifier trained via self-distillation [24], where the sharpened prediction from one augmented view supervises the other to enforce prediction consistency. The total loss $\mathcal{L}_{cls}$ includes the cross-entropy loss on the labeled data, self-distillation and entropy maximization on both labeled and unlabeled data. Formally, with $K = |\mathcal{Y}_l \cup \mathcal{Y}_u|$ denoting the total number of categories, where $p_i^{(k)}$ denotes the predicted probability of assigning $x_i$ to category $k$. The soft pseudo-label $q_i$ is obtained from another augmented view of the same sample, using a sharpened prediction to provide a more confident supervisory signal. The classification objectives are then defined as the cross-entropy loss between the predictions and pseudo-labels or ground-truth labels: $\ell(q', p) = -\sum_k q'^{(k)} \log p^{(k)}$. The unsupervised and supervised classification losses are given by:

$$\mathcal{L}_{cls}^u = \frac{1}{|B|} \sum_{i \in B} \ell(q_i', p_i) - \epsilon H(\bar{p}), \quad \mathcal{L}_{cls}^s = \frac{1}{|B^l|} \sum_{i \in B^l} \ell(y_i, p_i), \tag{4}$$

where $y_i$ denotes the one-hot ground-truth label of $x_i$ and $B^l$ corresponds to the labeled subset of $B$. Following [24], we adopt a mean-entropy maximization regularizer for the unsupervised objective. Here $\bar{p} = \frac{1}{2|B|} \sum_{i \in B} (p_i + p_i')$ denotes the mean prediction of a batch, and the entropy is defined as $H(\bar{p}) = -\sum_k \bar{p}^{(k)} \log \bar{p}^{(k)}$. The overall classification objective is defined as $\mathcal{L}_{cls} = (1 - \lambda)\mathcal{L}_{cls}^u + \lambda \mathcal{L}_{cls}^s$, $\lambda$ is set to 0.35.

In the association stage, we only train the backbone network with density-aware prototypical contrast learning without knowing the number of class. The training objective in the decision-boundary refinement stage [23] is defined as the weighted sum of the classification losses and the representation loss, $\mathcal{L} = \mathcal{L}_{\text{cls}} + \beta\,\mathcal{L}_{\text{rep}}$, where $\beta = 0.1$.

Table 1: Comparison of RareGCD with SOTA methods.

| Method | PathMNIST | | | OrganAMNIST | | | BloodMNIST | | |
|---|---|---|---|---|---|---|---|---|---|
| | All | Seen | Novel | All | Seen | Novel | All | Seen | Novel |
| CMS | 73.14 | 87.71 | 61.51 | 75.96 | 93.67 | 51.57 | 69.33 | 76.12 | 61.98 |
| GCD | 76.42 | 91.34 | 64.51 | 77.65 | 85.68 | 66.59 | 75.25 | 78.66 | 71.56 |
| GPC | 71.40 | 84.57 | 60.88 | 60.64 | 71.45 | 45.69 | 67.73 | 71.25 | 63.92 |
| SimGCD | 77.84 | 92.13 | 66.44 | 75.20 | 88.97 | 56.24 | 70.41 | 78.92 | 61.19 |
| SPTNet | 83.15 | **95.91** | 72.96 | 80.75 | 88.95 | 69.46 | 77.02 | 83.26 | 70.27 |
| CIPR | 82.88 | 94.31 | 73.75 | 82.29 | 94.20 | 65.89 | 76.85 | 82.67 | 71.63 |
| MedGCD | 79.09 | 93.33 | 67.72 | 78.53 | 90.44 | 62.13 | 73.49 | 80.16 | 66.27 |
| RareGCD | **84.68** | 95.35 | **75.61** | **84.52** | **91.45** | **74.97** | **79.57** | **84.31** | **74.44** |

## 3 Experiments and Results

**Dataset:** Following [5], we evaluate the performance of RareGCD on three benchmark datasets: PathMNIST, OrganAMNIST, and BloodMNIST [27], which contain 9, 11, and 8 categories, respectively. For PathMNIST and OrganAMNIST, 5 categories are considered seen, whereas for BloodMNIST, 4 categories are considered seen, with the remaining categories regarded as novel. To simulate practical scenarios, we designate the categories with the smallest sample sizes as novel categories. Furthermore, within each seen category, we randomly select 50% of the samples as labeled and treat the remaining 50% as unlabeled.

**Implementation Details:** The proposed method was implemented in PyTorch on an NVIDIA RTX PRO 6000 GPU. Following [5], we use ResNet-18 as the backbone [13]. For both training stages, SGD optimizer is utilized, and the default epoch number is 200, with a batch size of 512. Learning rate is 0.01 for the representation learning stage decayed with a cosine annealing schedule, and 0.1 for the decision-boundary refinement stage of joint training. RareGCD employed random horizontal flipping for weak augmentation and randaugment [4] for strong augmentation. We compute classification accuracy for seen and clustering accuracy for novel categories. Clustering accuracy is determined using the Hungarian algorithm [16]. Joint accuracy across all categories is reported using clustering accuracy for both seen and novel categories. Following [5, 21], we assume that the number of classes is known in all experiments to ensure a fair comparison. We reimplemented the MedGCD based on the original paper [5].

### 3.1 Comparison with State-of-the-Art Methods

We compare the performance of proposed RareGCD with SOTA methods, including CMS [2], GCD [21], GPC [28], SimGCD [24], CIPR [12], SPTNet [22] and MedGCD [5]. Table 1 shows the overall accuracy , seen accuracy and novel accuracy of RareGCD compared to baseline methods. We observe that RareGCD consistently outperforms all baseline methods, with particularly strong improvements in the accuracy of novel categories. Specifically, RareGCD improves novel accuracy ranging from 2.36% to 14.73%, 5.51% to 29.28%, and 2.81% to 13.25% on PathMNIST, OrganAMNIST, and BloodMNIST, respectively, compared to

Table 2: Comparison of RareGCD with SOTA methods under different imbalance ratios.

| $r = 50$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | PathMNIST | | | OrganAMNIST | | | BloodMNIST | | |
| | All | Seen | Novel | All | Seen | Novel | All | Seen | Novel |
| CMS | 53.29 | 54.61 | 44.19 | 62.22 | 65.57 | 43.97 | 66.59 | 70.43 | 48.58 |
| GCD | 59.24 | 62.13 | 42.11 | 64.09 | 66.82 | 49.21 | 68.23 | 73.19 | 44.97 |
| GPC | 52.61 | 54.89 | 38.61 | 57.23 | 62.15 | 30.42 | 64.35 | 68.74 | 43.76 |
| SimGCD | 59.97 | 61.02 | 53.57 | 62.21 | 65.02 | 46.90 | 67.90 | 72.21 | 47.68 |
| SPTNeT | 67.92 | 71.82 | 43.98 | 62.03 | 71.19 | 38.75 | 70.21 | 80.91 | 20.36 |
| CIPR | 68.75 | 71.42 | 52.36 | 65.98 | 67.17 | 59.49 | 71.85 | 75.93 | 52.71 |
| MedGCD | 61.92 | 63.16 | 49.40 | 64.18 | 67.79 | 44.51 | 68.19 | 72.45 | 48.21 |
| RareGCD | **70.58** | **72.03** | **61.68** | **71.63** | **73.22** | **62.97** | **74.36** | **78.61** | **54.43** |
| $r = 100$ | | | | | | | | | |
| Method | PathMNIST | | | OrganAMNIST | | | BloodMNIST | | |
| | All | Seen | Novel | All | Seen | Novel | All | Seen | Novel |
| CMS | 46.31 | 48.89 | 22.24 | 51.66 | 53.24 | 38.42 | 61.42 | 66.48 | 23.96 |
| GCD | 51.92 | 53.48 | 37.35 | 55.63 | 57.29 | 41.89 | 64.60 | 70.78 | 21.30 |
| GPC | 47.58 | 49.64 | 28.34 | 44.57 | 46.95 | 24.87 | 46.52 | 50.89 | 15.90 |
| SimGCD | 53.51 | 55.36 | 36.23 | 53.54 | 55.08 | 40.74 | 64.44 | 70.61 | 21.25 |
| SPTNeT | 55.65 | 58.61 | 28.02 | 58.37 | 66.76 | 31.43 | 47.56 | 51.77 | 18.06 |
| CIPR | 54.63 | 56.17 | 39.25 | 63.74 | 65.73 | 47.26 | 65.79 | 70.74 | 31.10 |
| MedGCD | 54.29 | 55.93 | 38.98 | 55.28 | 56.39 | 46.09 | 64.86 | 70.63 | 24.43 |
| RareGCD | **57.43** | **58.97** | **43.05** | **66.51** | **67.86** | **55.34** | **69.06** | **73.18** | **40.19** |

other methods. RareGCD consistently outperforms MedGCD across all datasets. The results demonstrate the effectiveness of RareGCD in learning superior representations from unlabeled data, enabling accurate classification of both seen and novel categories.

### 3.2 Results under Long-Tailed Settings

Novel categories often correspond to rare diseases and thus exhibit long-tailed distributions. To reflect this practical challenge, following the setting in [18, 14], we construct a long-tailed distribution by downsampling each class according to an exponential decay rule on all datasets. Given $k$ classes and an imbalance ratio $r = \frac{N_0}{N_{k-1}}$, where $N_0$ denotes the number of samples in the most frequent (head) class, the number of samples for class $j \in [0, k)$ is computed as: $N_j = N_0 \cdot r^{-\frac{j}{k-1}}$. In contrast to settings that predefine the head class by label index, we identify the head class as the class with the largest number of samples and use its sample count as $N_0$. For each class, we randomly sample $N_j$ instances to form the final training set. All other settings follow the experimental setup described above.

Table 2 reports the comparison results under long-tailed settings with imbalance ratios $r = 50$ and $r = 100$. Overall, RareGCD consistently outperforms all baseline methods across both datasets, demonstrating strong robustness to class imbalance. Under the setting of $r = 50$, RareGCD achieves performance gains of 8.11%, 3.48% and 1.72% over the strongest baseline on PathMNIST,

Table 3: Ablation results under long-tailed settings ($r = 50$) on PathMNIST and OrganAMNIST.

| Methods | PathMNIST ($r = 50$) | | | OrganAMNIST ($r = 50$) | | |
|---|---|---|---|---|---|---|
| | All | Seen | Novel | All | Seen | Novel |
| RareGCD (Kmeans) | 51.79 | 55.36 | 29.88 | 53.09 | 57.15 | 30.97 |
| RareGCD (Semi-Kmeans) | 60.93 | 62.54 | 51.05 | 61.02 | 62.39 | 53.55 |
| w/o $\mathcal{L}_{\text{rep}}$ | 66.47 | 69.43 | 48.31 | 67.24 | 70.19 | 51.16 |
| w/o Param. Cls | 68.86 | 70.32 | 60.01 | 68.34 | 69.71 | 60.87 |
| **RareGCD** | **70.58** | **72.03** | **61.68** | **71.63** | **73.22** | **62.97** |

OrganAMNIST and BloodMNIST for novel class, respectively. When the imbalance becomes more severe ($r = 100$), RareGCD still surpasses the best competing method across all datasets.

### 3.3 Ablation Studies

To investigate how each component affects the model performance, we perform ablation experiments and present the results in Table 3. We first replace the proposed sample association module with conventional clustering methods, including K-means and Semi-Kmeans [21]. This replacement leads to clear performance degradation in overall accuracy, with particularly severe drops on novel categories. Specifically, the novel-class accuracy decreases by 31% and 10% on PathMNIST when using Kmeans and Semi-Kmeans, respectively. These results indicate that the proposed prototype-guided sample association performs more effectively on long-tailed datasets. We further replace the density-aware prototypical contrast learning $\mathcal{L}_{\text{rep}}$ with a standard contrast learning. While this substitution results in only a slight performance drop on seen classes, it causes substantial degradation on novel classes, with accuracy reductions of 13% and 11% on PathMNIST and OrganAMNIST, respectively. In addition, w/o Param. Cls denotes that removes the decision-boundary refinement stage during training. The results show a moderate but consistent performance decline, indicating that parametric classification provides complementary supervision that stabilizes representation learning, especially for seen categories.

## 4 Conclusion

In this work, we propose RareGCD to address the challenge of generalized category discovery in practical scenarios, particularly under long-tailed distributions where novel categories correspond to rare diseases. RareGCD addresses this challenge by integrating prototype-guided sample association with density-aware prototypical contrast learning tailored to long-tailed data, alongside a parametric classification branch for stable category-level supervision. Extensive experiments under both standard and realistic long-tailed settings demonstrate that RareGCD consistently outperforms state-of-the-art methods.

**Acknowledgments.** This work is supported by JST SPRING JPMJSP2102.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References


1. Chen, X., Wang, X., Zhang, K., Fung, K.M., Thai, T.C., Moore, K., Mannel, R.S., Liu, H., Zheng, B., Qiu, Y.: Recent advances and clinical applications of deep learning in medical image analysis. Medical image analysis **79**, 102444 (2022)
2. Choi, S., Kang, D., Cho, M.: Contrastive mean-shift learning for generalized category discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23094–23104 (2024)
3. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE transactions on information theory **13**(1), 21–27 (1967)
4. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 702–703 (2020)
5. Das, A., Gautam, C., Agrawal, P., Yang, F., Liu, Y., Savitha, R.: Medgcd: Generalized category discovery in medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 427–437. Springer (2025)
6. Feng, W., Ge, Z.: Generalized category discovery under domain shift: A frequency domain perspective. Advances in Neural Information Processing Systems **38**, 111721–111749 (2026)
7. Feng, W., Jiang, Y., Zhou, S., Ge, Z.: Beyond the static world: Continual category discovery under visual drift. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25032–25042 (2026)
8. Feng, W., Jiang, Y., Zhou, S., Qi, Z., Xu, Z., Wang, Z., Tang, F., Ge, Z.: Seeing through the shift: Causality-inspired robust generalized category discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17766–17775 (2026)
9. Feng, W., Wang, B., Wang, Z., Zhou, S., Ge, Z.: Leveraging image-text pairs for generalized category discovery in medical image classification. IEEE Transactions on Medical Imaging (2026)
10. Feng, W., Zhou, S., Jiang, Y., Ge, Z.: Prism: Progressive robust learning for open-world continual category discovery. In: The Fourteenth International Conference on Learning Representations (2026)
11. Feng, W., Zhou, S., Jiang, Y., Tang, F., Ge, Z.: Neighbor-guided unbiased framework for generalized category discovery in medical image classification. IEEE Journal of Biomedical and Health Informatics (2025)
12. Hao, S., Han, K., Wong, K.Y.K.: Cipr: An efficient framework with cross-instance positive relations for generalized category discovery. arXiv preprint arXiv:2304.06928 (2023)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Ju, L., Wu, Y., Wang, L., Yu, Z., Zhao, X., Wang, X., Bonnington, P., Ge, Z.: Flexible sampling for long-tailed skin lesion classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 462–471. Springer (2022)

15. Ju, L., Zhou, S., Zhou, Y., Lu, H., Zhu, Z., Keane, P.A., Ge, Z.: Delving into out-of-distribution detection with medical vision-language models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 133–143. Springer (2025)
16. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)
17. Liu, Q., Yu, L., Luo, L., Dou, Q., Heng, P.A.: Semi-supervised medical image classification with relation-driven self-ensembling model. IEEE transactions on medical imaging **39**(11), 3429–3440 (2020)
18. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2537–2546 (2019)
19. Ma, Y., Hou, J., Zhang, C., Zhou, Y., Ge, Z., Xie, H., Ju, L.: Benchmarking real-world medical image classification with noisy labels: Challenges, practice, and outlook. Pattern Recognition p. 113647 (2026)
20. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems **33**, 596–608 (2020)
21. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Generalized category discovery. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7492–7501 (2022)
22. Wang, H., Vaze, S., Han, K.: Sptnet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. arXiv preprint arXiv:2403.13684 (2024)
23. Wang, M., Zhong, Z., Gong, X.: Prior-constrained association learning for fine-grained generalized category discovery. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 21162–21170 (2025)
24. Wen, X., Zhao, B., Qi, X.: Parametric classification for generalized category discovery: A baseline study. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16590–16600 (2023)
25. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3733–3742 (2018)
26. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3415–3424 (2017)
27. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1), 41 (2023)
28. Zhao, B., Wen, X., Han, K.: Learning semi-supervised gaussian mixture models for generalized category discovery. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16623–16633 (2023)
29. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1318–1327 (2017)