Federated Generalized Category Discovery

Nan Pu[1] Wenjing Li[2,3*] Xingyuan Ji[5,6] Yalan Qin[7] Nicu Sebe[1] Zhun Zhong[2,4*]
[1]University of Trento, Trento, Italy [2]Hefei University of Technology, Hefei, China
[3]University of Leeds, Leeds, UK [4]University of Nottingham, Nottingham, UK
[5]Xi'an Jiaotong University, Xi'an, China [6]Leiden University, Leiden, the Netherlands
[7]Shanghai University, Shanghai, China

Abstract

Generalized category discovery (GCD) aims at grouping unlabeled samples from known and unknown classes, given labeled data of known classes. To meet the recent decentralization trend in the community, we introduce a practical yet challenging task, Federated GCD (Fed-GCD), where the training data are distributed among local clients and cannot be shared among clients. Fed-GCD aims to train a generic GCD model by client collaboration under the privacy-protected constraint. The Fed-GCD leads to two challenges: 1) representation degradation caused by training each client model with fewer data than centralized GCD learning, and 2) highly heterogeneous label spaces across different clients. To this end, we propose a novel Associated Gaussian Contrastive Learning (AGCL) framework based on learnable GMMs, which consists of a Client Semantics Association (CSA) and a global-local GMM Contrastive Learning (GCL). On the server, CSA aggregates the heterogeneous categories of local-client GMMs to generate a global GMM containing more comprehensive category knowledge. On each client, GCL builds class-level contrastive learning with both local and global GMMs. The local GCL learns robust representation with limited local data. The global GCL encourages the model to produce more discriminative representation with the comprehensive category relationships that may not exist in local data. We build a benchmark based on six visual datasets to facilitate the study of Fed-GCD. Extensive experiments show that our AGCL outperforms multiple baselines on all datasets. Code is available at https://github.com/TPCD/FedGCD.

1. Introduction

Generalized category discovery (GCD) seeks to categorize unlabeled samples from known and unknown classes by leveraging labeled data of known classes. As a more practical extension of novel category discovery (NCD) [4, 10–12, 19, 36, 42, 45, 47, 48], GCD has attracted increasing attention. While existing GCD methods [9, 34, 38, 39, 41] have achieved promising performance, they always require centralized training, where the training data need to be accessed at once. However, this strategy violates many practical application scenarios: the GCD data are distributively collected by different local clients and the data in each client cannot be shared with others due to the privacy concerns. For instance, as shown in Fig. 1, a global species research center plans to discover the new species of global birds through the collaboration of local stations located around the world. Each local station is responsible for capturing and partially annotating bird images. Due to the difference in local policies and laws, it is hard to make an agreement to share the local data between stations. Thus, a decentralized

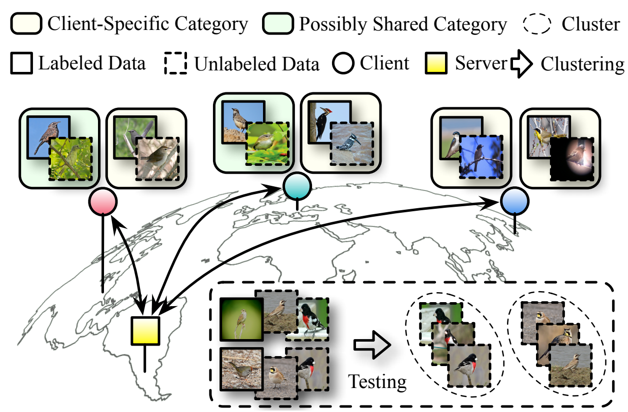

Figure 1. Illustration of the proposed Fed-GCD with the global bird species discovery case. In Fed-GCD, the data are distributively collected from the different local stations (clients) over the world, which are partially annotated. Each client includes client-specific categories and may share some common categories with the other clients. Moreover, due to data privacy, the raw data stored in local clients are not allowed to share with the central server or other clients. The goal of Fed-GCD is to collaboratively train a generic GCD model under the privacy constraint, and then utilize it to discover novel categories in the unlabeled data on the server.

*Corresponding Author.

system is required to handle this pragmatic GCD scenario.

To meet this requirement, we propose a practical yet challenging task, namely Federated GCD (Fed-GCD), in which the GCD data are individually collected and partially annotated by local clients as well as cannot be shared with other clients. The objective of Fed-GCD is to train a generic GCD model via the collaboration across local clients without sharing local samples, which can recognize both known and unknown categories in the unlabeled data. Compared with the conventional federated learning (FL) setups [31, 32, 43], in Fed-GCD, local data are partially-labeled and unlabeled data may belong to unknown categories that disappear in labeled data. In addition, clients may share some common categories since some species of birds could live in different continents as shown in Fig. 1, and the different clients may have distinct client-specific categories. Attributed to such a complicated yet real situation, Fed-GCD suffers from 1) additional difficulties caused by open-set learning on limited local data, and 2) more severe data heterogeneity problems due to the inconsistent label space between clients.

To tackle the challenges, we propose an Associated Gaussian Contrastive Learning (AGCL) framework, which unifies the discriminative representation learning on the limited local data and the heterogeneous category aggregation on the central server, benefiting from learnable GMMs with a maximum likelihood regularization. Specifically, we propose to represent the potential classes by a learnable Gaussian mixture model (GMM), which brings two advantages. First, the learnable mechanism enables us to perform class-aware contrastive learning with dynamic Mahalanobis distance, which can reduce the side effects of inaccurate clustering. Second, modeling the classes as GMMs is favorable for generating informative feature-level samples of each category on server, without assessing the raw data.

To this end, we propose a client semantics association (CSA) on the central server and a global-local GMM Contrastive Learning (GCL) on local clients. CSA builds a new feature set by sampling from each category of the uploaded local GMMs generated by clustering local data. Then, CSA aggregates the category knowledge by clustering on the feature set, which yields a global GMM. This process not only implicitly aligns the shared classes across local clients, but also aggregates client-specific category information. As a result, the global GMM can enrich both the intra- and inter-class relationships for local training. GCL targets at performing robust contrastive representation learning by jointly using global and local GMMs. On the one hand, the GMM-based contrastive learning is insensitive to wrongly pseudo-labeled samples, which can help the model to learn robust representation. On the other hand, the association of global-local GMMs enforces the model to learn more generalized representation in a complementary way.

Table 1. Comparison between different federated learning (FL) setups. "FS", "SS" and "SE" denote fully-supervised, self-supervised and semi-supervised, respectively.

FL Setup	Out of Categories	Annotation on Client
FS [32]	✗	Fully Labeled
SS [43]	✗	Unlabeled
SE [31]	✗	Partially Labeled
Fed-GCD	✓	Partially Labeled

We summarize the contributions of this work as follows:

- **Task contribution.** We explore a new yet practical GCD task, namely Fed-GCD, which investigates GCD problems under a federated learning scenario.
- **Technical contribution.** We propose a new AGCL framework for Fed-GCD. AGCL fully takes the advantage of the local and the global GMMs to learn generalized representation in a comprehensive manner.
- **Empirical contribution.** We build a Fed-GCD benchmark with two heterogeneity degrees based on six datasets to simulate possible conditions in real-world GCD applications. Experiments demonstrate that our AGCL can improve performance across all settings.

2. Related Work

Generalized Category Discovery (GCD) aims to categorize all images in an unlabelled set by using the knowledge learned from a set of labeled categories. Unlike earlier related tasks such as Novel Category Discovery [10, 11] (NCD) and generalized transfer learning [15, 16], GCD assumes that the unlabeled data comes from both known and unknown categories. Therefore, GCD is a practical and challenging task that has attracted increasing attention [4, 9, 12, 19, 19, 36, 38, 39, 41, 42, 45, 47]. For example, GCD [39] has indicated that the combination of self-supervised and supervised representation learning is helpful for improving clustering discovery. XCon [9] has proposed learning with multiple experts for fine-grained category discovery. OpenCon [38] and DCCL [34] have demonstrated the significant superiority of jointly considering prototypical contrastive learning and pseudo-label assignment. Recently, GPC [46] has introduced GMM to address GCD tasks. However, the semi-supervised GMMs proposed in GPC [46] are only used for estimating the number of unknown categories instead of end-to-end representation learning, which is different from ours. *Although these methods show promising performance under GCD assumptions, they neglect the increasingly important issue of data privacy. To investigate this overlooked issue and address additional technical bottlenecks, we design a Fed-GCD task and introduce a new AGCL framework accordingly.*

Federated Learning (FL) is a promising solution for privacy-preserving decentralized training. In the typical FL algorithm, FedAvg [32], the goal is to learn a global model by averaging weight parameters across local models trained on private client datasets. Most existing FL works [1, 6, 20, 27, 29, 30] focus on supervised learning settings, where the local private data are fully labeled. However, the assumption that all of the data examples are fully annotated is not realistic for real-world applications like GCD. Thus, one early work [43] has attempted to introduce self-supervised learning into the FL framework. Later, since there is often partially-labeled data in real-world scenarios, some semi-supervised FL approaches [22, 31] are proposed to exploit the partial supervision and learn better representations with few annotation costs. As summarized in Tab. 1, these works assume local clients share a common label space that is infeasible for GCD tasks. *In contrast, our Fed-GCD is challenged by more severe issues of data heterogeneity, because the label spaces are even non-overlapping among clients.*

Contrastive learning (CL) has been demonstrated to be highly effective for representation learning in a self-supervised setting [3]. Inspired by the powerful CL approaches [8, 13, 17, 21], GCD [39] has introduced a combination of the self-supervised and the semi-supervised learning to enhance GCD representation. Moreover, prototypical contrastive learning [28] (PCL) further considers class-level supervision by contrasting instance features with a set of prototypes. However, PCL needs an instance-level memory buffer to produce the prototype set, which is computationally and memory-intensive. *In contrast to the PCL that focuses on the learning of prototypes, our GCL considers additional class-aware variances to comprehensively model data distributions without instance buffer, by incorporating the classical GMM and contrastive learning in a unified framework. This allows models to be insensitive to outliers, especially for unreliable clusters.*

3. Federated Generalized Category Discovery

3.1. Problem Definition and Formulation

Given the practical requirements of generalized category discovery (GCD) applications (*e.g.*, species distribution and data privacy), it is necessary to build a generic GCD model via collaborative decentralized training across clients without sharing their local data. To meet these requirements, we propose a federated generalized category discovery (Fed-GCD) task. In Fed-GCD task, the local training data collected by each client are partially labeled, where the labeled data belong to known categories, and the unlabeled data may come from known or unknown novel categories. Additionally, each client learns on its distinct label set, which contains client-specific categories and may include some shared common categories. Compared to the semi-supervised federated learning [18] (semi-FL) setting that assumes both labeled and unlabeled data belong to known categories and share a common label space, Fed-GCD is more challenging due to highly-heterogeneous data issues attributed to inconsistent label spaces between clients and additional difficulties caused by open-set learning on local data. In light of this, Fed-GCD aims to 1) improve the local GCD model's representation learning ability on limited local data in open-set learning scenarios, and 2) associate the heterogeneous local label spaces to provide comprehensive category knowledge for local training. To the best of our knowledge, we are the first to explore the FL setup in GCD.

Formally, in the Fed-GCD task, there are N^L local client models $\{\Theta_n^L\}_{n=1}^{N^L}$ and one central server with the GCD model Θ^G. In the beginning, the global model Θ_0^G is initialized with the weights pre-trained on a publicly available large dataset (*e.g.*, ImageNet [5]) and distributed to each client. Given the local dataset on n-th client $\mathcal{D}_n^L = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N_n^L} \in \mathcal{X}_n^L \times \mathcal{Y}_n^L$ with the corresponding image set $\mathcal{X}_n^L$ and label set $\mathcal{Y}_n^L$, the n-th client is required to train its local model Θ_n^L based on the distributed global model Θ_0^G by leveraging its local dataset $\mathcal{D}_n^L$. In our Fed-GCD setup, we assume that for i-th and j-th client, $i \neq j$, $\mathcal{Y}_i^L$ and $\mathcal{Y}_j^L$ might be partially overlapping or completely non-overlapping, but their label space cannot be same (*i.e.*, $\mathcal{Y}_i^L \bigcup \mathcal{Y}_j^L \neq \mathcal{Y}_i^L or \mathcal{Y}_j^L$). To simulate such data distribution that often exists in real-world GCD applications, we adopt the parametric Dirichlet distribution [14] to control the degree of data heterogeneity.

3.2. Baseline

We employ the commonly-used FedAvg [32] algorithm, as our basic framework. Due to the inconsistent label spaces between local clients, we follow the previous FL work [26] that only sends the feature extractor to the server. Given a feature extractor f parameterized by Θ, the extracted representation is defined as $\boldsymbol{v} = f(\boldsymbol{x})$. As illustrated in Fig. 2 (a), (d) and (e), the steps of the baseline for the collaborative training are as follows.

Step I. In the t-th communication round, the server first aggregates the client models Θ_i^L uploaded from the last communication round, by taking a weighted average of them:

$$\Theta_{t+1}^G = \sum_{n=1}^{N^L} \frac{N_n^L}{N} \cdot \Theta_n^L, \quad N = \sum_{i=1}^{N^L} N_i^L. \tag{1}$$

Then, the averaged model is distributed to each client.

Step II. Based on the received global model, the i-th client trains its model by using local data $\mathcal{D}_i^L$ with the instance contrastive learning loss $\mathcal{L}^I$ proposed in [39] (in Fig. 2 (d)). Specifically, we define that x_i and $\hat{x}_i$ are two views of random augmentations for the same image in a mini-batch $\mathcal{B} = \mathcal{B}^L \cup \mathcal{B}^U$, consisting of the labeled subset $\mathcal{B}^L$ and unlabeled subset $\mathcal{B}^U$. The extracted representation $\mathbf{v}_i$

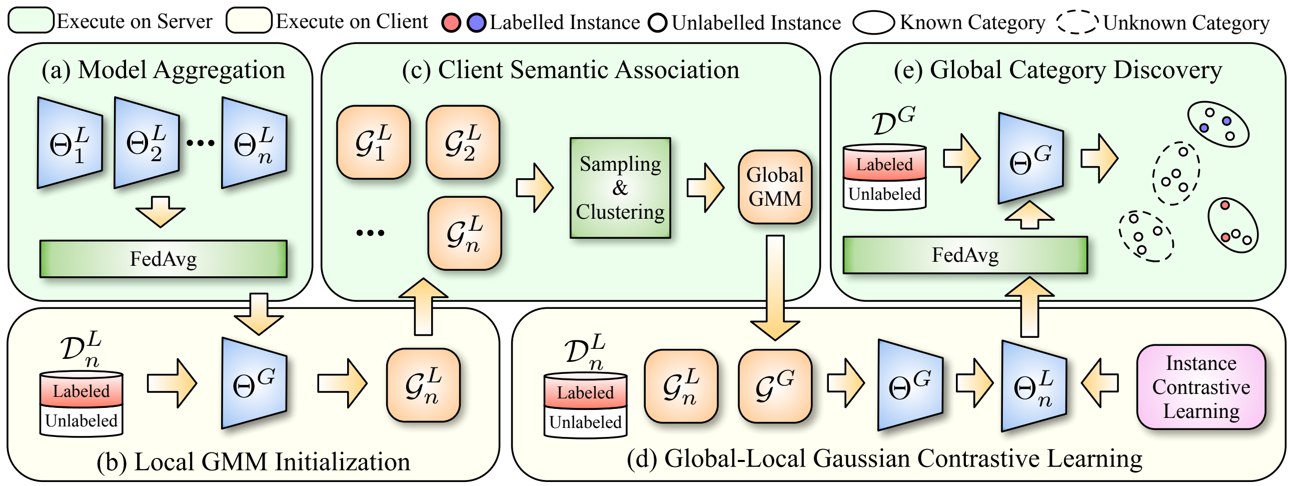

Figure 2. Diagram of our Associated Gaussian contrastive learning (AGCL) framework. We first apply FedAvg [32] to aggregate the uploaded local models, resulting in a global model that will be distributed to all clients. Then, after leveraging the distributed models to extract image features, local clients are required to cluster these features and initialize local GMMs. Next, the local GMMs are uploaded to the central server, and aggregated by the proposed CSA, to generate a global GMM before local training. Later, the server distributes the global GMM to each client. Based on the global-local GMMs, client models are collaboratively optimized by the proposed GCL. Finally, a generic model is trained for global category discovery.

is further projected by a MLP projection head h to high-dimensional embedding space for instance-level contrastive learning. The loss function is formulated as:

$$\begin{aligned}\mathcal{L}^n_{ins} &= (\lambda-1)\sum_{i\in\mathcal{B}}\log\frac{\mathcal{S}_{ins}(\boldsymbol{v}_i,\hat{\boldsymbol{v}}_i,\tau^S)}{\sum_{j\in\mathcal{B},j\neq i}\mathcal{S}_{ins}(\boldsymbol{v}_i,\boldsymbol{v}_j,\tau^S)}\\ &+\sum_{i\in\mathcal{B}^L}\frac{-\lambda}{|\mathcal{P}(i)|}\sum_{p\in\mathcal{P}(i)}\log\frac{\mathcal{S}_{ins}(\boldsymbol{v}_i,\boldsymbol{v}_p,\tau^L)}{\sum_{j\in\mathcal{N}(i)}\mathcal{S}_{ins}(\boldsymbol{v}_i,\boldsymbol{v}_j,\tau^L)},\end{aligned} \tag{2}$$

$$\mathcal{S}_{ins}(\boldsymbol{v},\hat{\boldsymbol{v}},\tau)=\exp\left(h\left(\boldsymbol{v}\right)\cdot h\left(\hat{\boldsymbol{v}}\right)/\tau\right), \tag{3}$$

where $\mathcal{P}(i)$ and $\mathcal{N}(i)$ are the positive and the negative index set for the anchor image $i\in\mathcal{B}^L$. λ is a trade-off factor to balance self-supervised and supervised learning.

Step III. The updated global model will be transmitted to each client. **Step I** and **II** are repeated until convergence. Ultimately, we use the final global model to discover new categories in the unlabeled data on server (in Fig. 2 (e)).

3.3. Limitations and Motivations

Although the baseline approach works on our Fed-GCD benchmark, it shows unsatisfactory performance compared with centralized training, especially on fine-grained GCD datasets (see Tab. 3). We argue that the main reasons are attributed to two aspects: 1) the GCD [39] applied in local client training mainly focuses on instance-level contrastive learning while it neglects class-level contrastive learning, especially on unlabeled data. Since class-level or prototypical supervision plays an important role in open-set learning [38], the Fed-GCD fails to collaboratively train a robust global GCD model without discriminative local models; 2) sharing only the backbone network is inefficient to leverage the comprehensive category relationship that may not be observed in local clients. Moreover, although the label space of each client in Fed-GCD might potentially share some common semantic information (*e.g.*, a specie of bird distributed on different continents), the server has no explicit knowledge to align or leverage such class-level relationships under privacy protection constraints.

To overcome these limitations, we consider representing the class-level knowledge by a learnable Gaussian mixture model (GMM), which is initialized by a parameter-free clustering approach. Each component of the GMM models a potential class/cluster with class-specific mean and variance, which naturally results in a concentration-based distance metric for robust contrastive learning. This idea enables models to 1) mitigate the negative effects caused by inaccurate clustering and enforce class-level supervision into local training, and 2) generate informative feature-level samples of each category for knowledge aggregation on the server without leaking original data.

4. Federated Gaussian Contrastive Learning

Based on the above analyses, we propose a novel Associated Gaussian contrastive learning (AGCL) framework to accomplish efficient Fed-GCD. AGCL consists of a global-local GMM Contrastive Learning (GCL) on local clients and a client semantics association (CSA) on the central server. The former enforces a class-level contrastive learning in local training by jointly using a global GMM and a local one, where the local GMM is created by clustering on local data and the global GMM is distributed from the central server. The latter serves to aggregate heterogeneous category knowledge contained in the local GMMs following a client-agnostic manner, and generates the global GMM to provide comprehensive category relationship for local training. The goal of AGCL is to improve representation learning by enforcing class-aware GCL and associating related semantic knowledge scattered across clients.

4.1. Gaussian Contrastive Learning

As empirically demonstrated in [9, 28, 44], class-level or prototypical contrastive learning is efficient for learning a clustering-friendly representation. Recently, open-set contrastive learning [38] further indicates that such representation learning can significantly improve the GCD model's abilities to discover both known and unknown categories. However, these methods represent a class by using only the center or the mean of the class, which is insufficient and vulnerable to wrong pseudo-labeling caused by inaccurate clustering. To address this issue, we propose to employ a classical Gaussian mixture model (GMM) to model potential cluster distributions, and then perform class-level contrastive learning across the components of the GMM.

Formulating learnable GMMs in Fed-GCD setup. We assume that the n-th client generates a GMM $\mathcal{G}_n^L = \{\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i)\}_{i=1}^{M_n^L}$ with M_n^L components, where the $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ are the mean and variance of the i-th component. We use a component to model a potential class/category. For simplicity, we assume that the covariance matrix $\boldsymbol{\sigma}$ is diagonal and each cluster has the equal prior probability. By maximizing the posterior of $\boldsymbol{v}_i$ belonging to the y_i-th cluster, the margin-based GMM loss on the n-th client is derived as:

$$\mathcal{L}_{gmm}(\mathcal{G}_n^L, \boldsymbol{v}_i, y_i) = -\log \frac{|\boldsymbol{\sigma}_i|^{-\frac{1}{2}}\, \mathcal{S}_{gcl}(\boldsymbol{v}_i, y_i)}{\sum_{j=1, j\neq y_i}^{M_n^L} |\boldsymbol{\sigma}_j|^{-\frac{1}{2}}\, \mathcal{S}_{gcl}(\boldsymbol{v}_j, y_j)}, \tag{4}$$

$$\mathcal{S}_{gcl}(\boldsymbol{v}_i, y_i) = \exp\left(-\frac{1}{2}\left\|\frac{\boldsymbol{v}_i - \boldsymbol{\mu}_{y_i}}{\boldsymbol{\sigma}_{y_i}}\right\|^2 \cdot (1+m)\right), \tag{5}$$

where $\mathcal{S}_{gcl}$ is the similarity metric based on squared Mahalanobis distance and the m is a non-negative margin factor to increase the inter-class dispersion.

Discussion. Prototypical contrastive learning (PCL) [28] is a pioneering method to introduce class-level supervision into unsupervised contrastive learning. PCL estimates a scalar concentration as the temperature parameter to scale the similarity between a feature and its prototype. Although it is efficient for learning discriminative representation, it fails to model a precise representation distribution that is supposed to generate reliable representations for the downstream clustering. Here, we discuss the differences between PCL and GCL. The similarly metric of the PCL is:

$$\mathcal{S}_{pcl}(\boldsymbol{v}_i, y_i) = \exp\left(\boldsymbol{v}_i \cdot \boldsymbol{\mu}_{y_i} / \phi_{y_i}\right), \tag{6}$$

where ϕ_i is the estimated temperature parameter for the i-th cluster. Comparing Eqs. (5) and (6), different from PCL, we model the clusters via the GMM with additional covariance matrices, and naturally derive the squared Mahalanobis distance as distance metric for contrastive learning. *This allows models to dynamically control the temperatures in a dimension-wise way and to learn more reliable distributions of representations for the subsequent sampling.*

Furthermore, we introduce a maximum likelihood regularization term to explicitly compact clusters and constrain covariance, to avoid trivial solutions. For example, GMM generates a high classification accuracy, but the sample embedding is far away from the center of the cluster due to the large class-specific variance. Using the regularization loss can constrain the distance between the sample embedding and its corresponding cluster center as well as reduce the overlarge variances. The regularization loss is:

$$\mathcal{L}_{reg}(\mathcal{G}_n^L, \boldsymbol{v}_i, y_i) = -\log(\mathcal{S}_{gcl}(\boldsymbol{v}_i, y_i)) + \frac{1}{2}\log|\boldsymbol{\sigma}_{y_i}|. \tag{7}$$

Taking Eqs. (4), (5) and (7), the overall GCL loss is:

$$\mathcal{L}_{gcl}^n(\mathcal{G}_n^L) = \sum_{i=1}^{N_n^L} \mathcal{L}_{gmm}(\mathcal{G}_n^L, \boldsymbol{v}_i, y_i) + \alpha \mathcal{L}_{reg}(\mathcal{G}_n^L, \boldsymbol{v}_i, y_i), \tag{8}$$

where α is a weighting factor. Optimized by this objective, the cluster-specific mean and variance can be learned.

Remark. Different from the GMMs employed in [46], which are frozen during the whole training process, our GMMs are dynamically learning in an end-to-end fashion.

Semi-FINCH for Local GMM Initialization. While the number of GMM components can be generally pre-defined by the ground-truth category number, the number of categories to be discovered is unknown in GCD assumptions. In order to stabilize the GMM training and determine the number of its components (i.e., the number of potential categories), we propose to initialize the learnable GMMs based on improved parameter-free clustering assignment.

To leverage the category knowledge contained in labeled data, we first propose to improve an hierarchical clustering approach, FINCH [37], to a semi-supervised extension. Then, we use it to estimate the class number and assign local pseudo labels for client data. Specifically, we 1) extract features and calculate pair-wise similarities, 2) search the 1^{st}-neighbor of each sample based on the similarities, 3) force one random labeled sample that belongs to the same category as its 1^{st}-neighbor and then apply FINCH algorithm (detailed in supplementary materials) to yield multi-level clustering results, and 4) leverage the clustering accuracy of labeled samples as the index to select the clustering level. Finally, we choose the level that achieve highest clustering accuracy as the estimated results to calculate the cluster-specific mean and covariance to initialize the learnable GMM, as shown in Fig. 2 (b). In short, the proposed semi-FINCH can automatically capture the potential semantic relationships among both labeled and unlabeled samples with the guidance of labeled data, providing a reasonable initialization for local GMMs.

4.2. Client Semantic Association

In Fed-GCD task, the central server is unreasonable to get prior knowledge to align or identify local categories, due to privacy constraints. To overcome these bottlenecks, we propose a sample yet efficient client semantic association (CSA) by category-agnostic sampling and re-clustering, which aims to associate common semantic knowledge contained in local GMMs, and aggregate diverse local category information for enriching global category knowledge.

Category-Agnostic Knowledge Association. Let a set of the uploaded local GMMs as $\mathcal{G}^L = \{\mathcal{G}_n^L\}_{n=1}^{N^L}$, where $\mathcal{G}_n^L = \{\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i)\}_{i=1}^{M_n^L}$. We sample N^S instances from each component and attach its mean, which results in a new feature set $\mathcal{F}$ with size of $(\sum_i M_i^L) \times (N^S + 1) \times N^L$. Similarly, by applying parameter-free FINCH clustering on $\mathcal{F}$, the central server generates a new global GMM, as illustrated in Fig. 2 (c). The global GMM will be sent to each client for the subsequent local training.

Discussion. Intuitively, the samples in $\mathcal{F}$ with similar semantics will be re-clustered into a super-cluster that contains more information with a large variance. This process implicitly associates common classes scattered in clients, thereby further enriching intra-class information. Meanwhile, the clusters with relatively independent semantics will be preserved. In this way, CSA incorporates diverse knowledge from different clients into the global GMM, which establishes a bridge across clients. This allows the isolated knowledge to mutually transfer among clients, which provides complementary supervision (e.g., more negative classes and rich positive samples) for local GCL.

4.3. Associated Gaussian Contrastive Learning

As illustrated in Fig. 2 (d), taking the n-th client as an example, AGCL first uses nearest neighbor search to assign global pseudo labels for $\mathcal{D}_n^L$ based on the distance induced by the global GMM. Then, we consider both the global and the local GMM to guide the optimization of the local model. AGCL uses a convex combination of them to achieve an optimal balance between the local and the global knowledge learning. The objective of AGCL on the n-th client is:

$$\mathcal{L}_{agcl}^n = \mathcal{L}_{ins}^n + (1-\gamma)\mathcal{L}_{gcl}^n(\mathcal{G}^G) + \gamma\mathcal{L}_{gcl}^n(\mathcal{G}_n^L), \quad (9)$$

where the γ is a trade-off factor to control the strength of learning on global-local GMMs. When γ is equal to 1, AGCL leverages only the local class-level supervision for representation learning. On the contrary, AGCL relies on only the aggregated global category information.

5. Experiment

5.1. Experimental Setup

Dataset. To facilitate the study of Fed-GCD task, we reorganize three commonly-used generic image classification datasets (*i.e.*, CIFAR-10 [24], CIFAR-100 [24] and ImageNet-100 [39]) and three more challenging fine-grained image classification datasets (*i.e.*, CUB-200 [40], Stanford Cars [23], and Oxford-IIIT Pet [33]) to construct a new Fed-GCD benchmark. For each dataset, first, we sample a subset of half the classes as “Old” categories in the original training set, and 50% of instances of each labeled class are drawn to form the labeled set, and all the remaining data form the unlabeled set. With the same rate of labeled-unlabeled splitting, we split the original testing set into labeled and unlabeled subsets for class number estimation and GCD testing on server. Then, we further leverage the β-Dirichlet distribution [14] to split the training set into N^L subsets, where the N^L subsets are regarded as local datasets individually stored in each client. We set N^L=5 in all experiments. The statistics are summarized in Tab. 6 (see Appendix).

Evaluation Protocols. Due to the varying data distribution in different Fed-GCD applications, we present two evaluation protocols to separately simulate the normally heterogeneous (NH) and extremely heterogeneous (EH) scenarios by adjusting β in Dirichlet distribution [14]. Specifically, we set $\beta = 0.2$ and $\beta = 0.05$ for NH and EH, respectively. The NH setting exists few common classes while there is no labeled categories shared across all clients in the EH setting (Tab. 6 in Appendix). For each dataset, we learn a global model in a decentralized training fashion. Following [39], during testing, we first estimate the number of the potential categories (*i.e.*, k) in the non-overlapping test set by using the labeled data stored on server. Then we calculate the maximum of clustering accuracy between the ground truth labels and the label assignment with the estimated k over the set of permutations via Hungarian algorithm [25]. Last, we measure the clustering accuracy for “All”, “Old” and “New” categories.

5.2. Implementation Details

On each client, we adopt the same backbone network, a ViT [7] pre-trained by DINO [2], and use its [CLS] to-

Table 2. Results on generic datasets with two different degrees of data heterogeneity.

Methods	NH setting ($\beta = 0.2$)									EH setting ($\beta = 0.05$)								
	CIFAR10			CIFAR100			ImageNet-100			CIFAR10			CIFAR100			ImageNet-100		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
Centralized-GCD	83.6	85.8	82.0	54.9	56.1	53.7	72.1	80.7	67.5	83.6	85.8	82.0	54.9	56.1	53.7	72.1	80.7	67.5
Centralized-PCL	87.1	87.5	86.9	57.9	57.0	58.0	74.9	81.5	67.9	87.1	87.5	86.9	57.9	57.0	58.0	74.9	81.5	67.9
Centralized-GCL	86.7	86.7	86.7	58.5	57.2	58.1	76.1	83.7	68.4	86.7	86.7	86.7	58.5	57.2	58.1	76.1	83.7	68.4
FedAvg + GCD	80.7	82.3	80.3	49.6	52.1	49.3	69.8	77.1	65.7	78.7	80.1	78.3	47.3	49.2	45.9	66.4	74.8	62.1
FedAvg + PCL	81.6	82.7	80.9	53.2	54.1	51.7	72.4	79.5	66.0	80.0	80.7	79.4	50.4	51.6	49.0	70.1	77.0	63.3
FedAvg + GPC	81.3	81.7	80.5	52.8	53.5	51.4	72.1	78.2	65.7	80.1	80.4	78.4	50.0	51.3	48.9	69.8	76.8	63.1
FedAvg + GCL	83.2	84.9	82.8	54.1	55.7	54.0	74.1	**81.8**	67.3	82.2	82.4	81.9	52.1	53.2	51.9	72.5	79.8	65.3
FedAvg + AGCL	84.7	85.5	84.6	**56.1**	**56.8**	**55.3**	**74.8**	80.2	**69.8**	82.5	83.4	82.2	54.2	54.6	54.0	73.1	78.1	67.0
FedProx + AGCL	**84.8**	**85.8**	**84.7**	55.9	56.5	54.9	74.7	80.3	69.5	**83.0**	**84.1**	**82.8**	**54.7**	**55.1**	**54.2**	**74.2**	**78.8**	**67.7**

Table 3. Results on fine-grained datasets with two different degrees of data heterogeneity.

Methods	NH setting ($\beta = 0.2$)									EH setting ($\beta = 0.05$)								
	CUB-200			Stanford-Cars			Oxford-Pet			CUB-200			Stanford-Cars			Oxford-Pet		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
Centralized-GCD	51.3	57.3	45.4	39.7	58.0	31.2	80.2	85.1	77.6	51.3	57.3	45.4	39.7	58.0	31.2	80.2	85.1	77.6
Centralized-PCL	58.0	55.7	60.0	41.4	55.3	38.0	86.2	85.7	86.4	58.0	55.7	60.0	41.4	55.3	38.0	86.2	85.7	86.4
Centralized-GCL	58.1	55.9	60.3	41.7	55.5	38.1	85.5	85.8	85.2	58.1	55.9	60.3	41.7	55.5	38.1	85.5	85.8	85.2
FedAvg + GCD	46.3	54.7	40.1	32.4	49.8	28.3	76.2	77.8	75.2	43.3	52.8	38.9	30.4	45.1	26.5	72.1	76.4	71.5
FedAvg + PCL	51.3	53.5	50.8	35.3	47.7	33.4	79.4	80.3	79.1	47.5	53.0	46.3	32.6	45.5	29.2	76.6	77.9	74.7
FedAvg + GPC	49.1	51.3	46.8	34.1	45.5	32.6	78.8	78.5	79.1	45.3	51.2	44.7	30.9	45.3	27.8	73.1	77.3	73.5
FedAvg + GCL	53.7	**54.9**	53.2	36.0	48.1	33.7	80.7	81.3	80.2	52.2	**53.3**	51.9	35.3	**45.7**	31.5	79.5	81.5	78.6
FedAvg + AGCL	55.2	52.5	56.7	38.2	**50.8**	36.0	**82.7**	**83.9**	**82.3**	53.1	52.9	54.2	36.4	44.9	32.8	81.4	82.0	80.7
FedProx + AGCL	**55.4**	52.7	**56.8**	**38.5**	50.7	**36.4**	82.5	83.6	82.2	**53.6**	53.2	**54.5**	**36.9**	45.2	**33.0**	**81.5**	**82.1**	**80.8**

ken for GCL learning and new category discovery. Following GCD [39], the instance contrastive learning is implemented by a projection head with 65,536 dimensions and two randomly-augmented views of an image, and λ, τ^S and τ^L are 0.35, 0.07 and 0.05, respectively. We fine-tune only the last block of the ViT [7] with an initial learning rate of 0.1 and upload it to the central server in each communication. The parameters in global GMMs are frozen during AGCL since the local data that include insufficient category information are not able to optimize the global GMM. The local GMMs and projection head are trained with an initial learning rate of 0.001 and 0.01, respectively. All the fine-tuned parameters are optimized by SGD [35] for 200 epochs with a cosine annealing schedule. The size of the mini-batch is set to 128. The hyper-parameters α, γ, m and N^S are set to 0.01, 0.9, 0.3 and 1 in all experiments.

5.3. Performance Evaluation

Since this work is the first to explore GCD tasks under a federated learning challenge, there is no Fed-GCD-specific method used for comparison. Thus, we first adapt the current GCD methods without parametric classifier (GCD [39], PCL [28], GPC [46]) into Fed-GCD as the strong baseline ("FedAvg + GCD", "FedAvg + PCL" and "FedAvg + GPC"). Note that we use our pseudo-label generation for "FedAvg + PCL", while keeping GPC's original approach in [46] for "FedAvg + GPC". Then, we implement the AGCL without global GCL ("FedAvg + GCL") and the full AGCL ("FedAvg + AGCL") to investigate the effects of our global-local GCL. Moreover, to provide a reference performance, we evaluate the centralized training performance of GCD ("Centralized-GCD"), PCL ("Centralized-GCD") and GCL ("Centralized-GCL"). Finally, we adapt AGCL in the advanced heterogeneous federated learning framework [29] ("FedProx + AGCL"), for a comprehensive comparison.

Table 4. The ablation study on the EH setting (β=0.05).

Setup	Component				CUB-200			Oxford-Pet		
	$\mathcal{L}_{ins}$	$\mathcal{L}_{gmm}^{L}$	$\mathcal{L}_{reg}$	$\mathcal{L}_{gmm}^{G}$	All	Old	New	All	Old	New
a)	✓				43.3	52.8	38.9	72.1	76.4	71.5
b)		✓			48.9	50.5	48.5	76.8	78.5	75.1
c)		✓	✓		50.6	51.8	49.8	78.0	80.7	77.4
d)	✓	✓	✓		52.2	53.1	52.0	79.5	81.5	78.6
e)	✓	✓	✓	✓	53.1	52.9	54.2	81.4	82.0	80.7

Summary. The experimental results demonstrate that 1) the proposed Fed-GCD task is challenging due to the severe data heterogeneity, which results in a large accuracy degradation between the centralized and decentralized training; 2) AGCL achieves consistent improvement in all settings. Benefiting from aggregating different categories scattered on clients, AGCL achieves better performance, especially on fine-grained tasks in the EH setting. 3) we verify the superiority of the end-to-end GMM design, which is more suitable than GPC for heterogeneous data. The GMM in GPC [46] are fixed, which might lead to inconsistent prototype updates, thereby degrading performance.

5.4. Effectiveness of Each Component of AGCL

We conduct five group ablation studies on the CUB-200 and the Pet datasets. The results are shown in Tab. 4. The setup (a) is the baseline method, *i.e.*, "FedAvg + GCD".

Effectiveness of local GCL. The setup a) and c) indicate that GCL outperforms the baseline by a large margin,

Table 5. Experiments with 10 Clients (*i.e.*, $N^L = 10$).

Setup	**NH** setting ($\beta = 0.05$)								
	CIFAR10			CIFAR100			ImageNet-100		
	All	Old	New	All	Old	New	All	Old	New
Centralized-GCD	83.6	85.8	82.0	54.9	56.1	53.7	72.1	80.7	67.5
Centralized-GCL	86.7	86.7	86.7	58.5	57.2	58.1	76.1	83.7	68.4
FedAvg + GCD	63.4	60.0	66.7	47.3	48.3	45.6	62.3	70.8	60.1
FedAvg + GCL	**68.2**	**64.2**	70.1	**52.5**	**53.9**	51.0	67.3	74.5	60.8
FedAvg + AGCL	68.1	63.8	**70.3**	52.2	53.6	**52.4**	**67.5**	**74.8**	**61.1**

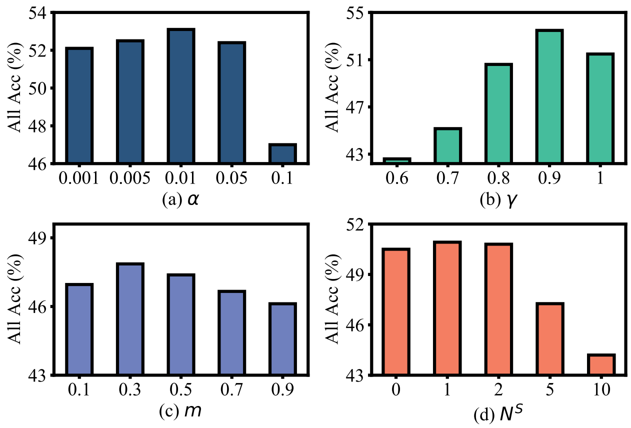

Figure 3. Illustration of impacts of hyper-parameters.

demonstrating the importance of class-level supervision in GCD. Especially for the accuracy of new classes, GCL outperforms the baseline by 10.9% on the CUB-200 dataset.

Effectiveness of regularization in Eq. (7). From the setup b) and c), we find that enforcing the regularizing loss can achieve consistent improvement, as the regularization loss encourages models to avoid trivial sub-optimal solutions.

Effectiveness of AGCL. The d) and e) demonstrate that CSA can associate heterogeneous category knowledge even without commonly-shared categories in EH setting. The associated knowledge contained in the global GMM complements representation learning based on local GMMs, thereby enhancing global new category discovery. Through analyzing the visualization results in Fig. 4 , we find that 1) most global cluster centers are located at the ground-truth centers without access to raw data, which demonstrates the effectiveness of category knowledge aggregation; 2) as for the blue-green cluster in the purple dashed line, the global cluster may serve as a super-class to provide distinct semantics for improving representation discriminability.

5.5. Hyper-Parameter Analyses

Impact of regularization weight in Eq. (8) is illustrated in Fig. 3 (a), which indicates that α is not sensitive from 0.001 to 0.05. α achieves optimal performance at 0.01.

Impact of trade-off factor in Eq. (9) is illustrated in Fig. 3 (b). We find that local GMMs dominate the AGCL while global GMM plays an assistant role in providing a few complementary information. This is because some aggregated category knowledge does not exist in the current

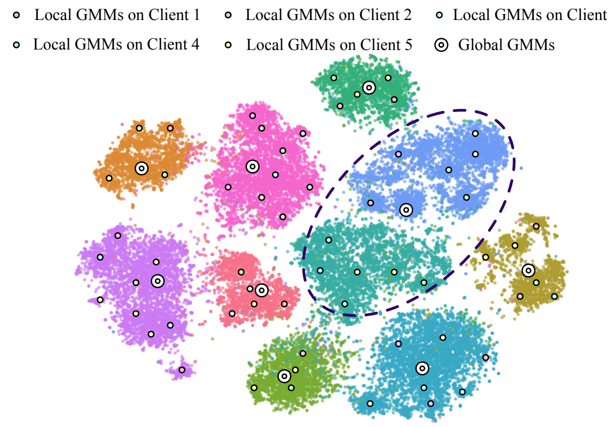

Figure 4. Visualization of learned GMMs on CIFAR10.

client. Thus, the global GMM cannot provide efficient supervision compared to the local GMM. Thus, when AGCL mainly relies on the global GMM, the performance of local training will be largely degraded (e.g.,$\gamma = 0.6$).

Impact of margin parameter in Eq. (5) is illustrated in Fig. 3 (c). We find that the margin parameter is not sensitive. We choose the optimal m=0.3 in all experiments.

Impact of sampling parameter in CSA is illustrated in Fig. 3 (d). We find that 1) clustering only means of all components (*i.e.*, $N^S = 0$) can improve the performance as well, and 2) sampling more than 2 samples leads to worse performance. Thus, N^S is set to 1 in all experiments.

Impact of number of clients. From Tab. 5, when $N^L = 10$, GCD suffers from a larger performance drop than our GCL, compared with the centralized training and FedAvg baseline. GCL achieves a relatively robust performance.

6. Conclusion

In this work, we propose a new Fed-GCD task, based on the practical requirement of decentralized training trends. To handle this task, we propose a novel Associated Gaussian Contrastive Learning (AGCL) framework specifically designed to overcome the unique challenges posed by Fed-GCD. Moreover, we build a benchmark based on six visual datasets to facilitate the study of Fed-GCD. Extensive experiments show that AGCL outperforms the FedAvg-based baseline on all datasets. In future, we attempt to relieve the requirement of storing labeled data in the central server to meet more realistic scenarios for Fed-GCD.

Acknowledgement. We acknowledge the support of the MUR PNRR project **iNEST**-Interconnected Nord-Est Innovation Ecosystem (ECS00000043) funded by the NextGenerationEU. Also, this work was partially supported by the EU Horizon projects ELIAS (No. 101120237) and AI4Trust (No. 101070190).

References

[1] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N. Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *ICLR*, 2021. 3

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 6

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3

[4] Haoang Chi, Feng Liu, Wenjing Yang, Long Lan, Tongliang Liu, Bo Han, Gang Niu, Mingyuan Zhou, and Masashi Sugiyama. Meta discovery: Learning to discover novel classes given very limited data. In *ICLR*, 2022. 1, 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3

[6] Enmao Diao, Jie Ding, and Vahid Tarokh. Heterofl: Computation and communication efficient federated learning for heterogeneous clients. In *ICLR*, 2021. 3

[7] 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 *ICLR*, 2021. 6, 7

[8] Wei Fan, Kunpeng Liu, Hao Liu, Yong Ge, Hui Xiong, and Yanjie Fu. Interactive reinforcement learning for feature selection with decision tree in the loop. *TKDE*, 2021. 3

[9] Yixin Fei, Zhongkai Zhao, Siwei Yang, and Bingchen Zhao. Xcon: Learning with experts for fine-grained category discovery. In *BMVC*, 2022. 1, 2, 5

[10] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *ICCV*, 2021. 1, 2

[11] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *CVPR*, 2019. 2

[12] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel: Automatically discovering and learning novel visual categories. *IEEE TPAMI*, 2021. 1, 2

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3

[14] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Federated visual classification with real-world data distribution. In *ECCV*, 2020. 3, 6

[15] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In *ICLR*, 2018. 2

[16] Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. In *ICLR*, 2018. 2

[17] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 2020. 3

[18] Wonyong Jeong, Jaehong Yoon, Eunho Yang, and Sung Ju Hwang. Federated semi-supervised learning with inter-client consistency & disjoint learning. In *ICLR*, 2021. 3

[19] KJ Joseph, Sujoy Paul, Gaurav Aggarwal, Soma Biswas, Piyush Rai, Kai Han, and Vineeth N Balasubramanian. Novel class discovery without forgetting. In *ECCV*, 2022. 1, 2

[20] Mikhail Khodak, Maria-Florina Balcan, and Ameet Talwalkar. Adaptive gradient-based meta-learning methods. In *NeurIPS*, 2019. 3

[21] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *NeurIPS*, 2020. 3

[22] Woojung Kim, Keondo Park, Kihyuk Sohn, Raphael Shu, and Hyung-Sin Kim. Federated semi-supervised learning with prototypical networks. *arXiv preprint arXiv:2205.13921*, 2022. 3

[23] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshop*, 2013. 6

[24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[25] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 1955. 6

[26] Chenglin Li, Di Niu, Bei Jiang, Xiao Zuo, and Jianming Yang. Meta-har: Federated representation learning for human activity recognition. In *WWW*, 2021. 3

[27] Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*, 2019. 3

[28] Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations. In *ICLR*, 2021. 3, 5, 7

[29] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *MLSys*, 2020. 3, 7

[30] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. In *ICLR*, 2021. 3

[31] Haowen Lin, Jian Lou, Li Xiong, and Cyrus Shahabi. Semifed: Semi-supervised federated learning with consistency and pseudo-labeling. *arXiv preprint arXiv:2108.09412*, 2021. 2, 3

[32] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 2017. 2, 3, 4

[33] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 6

[34] Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In *CVPR*, 2023. 1, 2

[35] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 1999. 7

[36] Subhankar Roy, Mingxuan Liu, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Class-incremental novel class discovery. In *ECCV*, 2022. 1, 2

[37] Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. Efficient parameter-free clustering using first neighbor relations. In *CVPR*, 2019. 5

[38] Yiyou Sun and Yixuan Li. Opencon: Open-world contrastive learning. In *TMLR*, 2022. 1, 2, 4, 5

[39] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *CVPR*, 2022. 1, 2, 3, 4, 6, 7

[40] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 6

[41] Muli Yang, Yuehua Zhu, Jiaping Yu, Aming Wu, and Cheng Deng. Divide and conquer: Compositional experts for generalized novel class discovery. In *CVPR*, 2022. 1, 2

[42] Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Self-labeling framework for novel category discovery over domains. In *AAAI*, 2022. 1, 2

[43] Fengda Zhang, Kun Kuang, Long Chen, Zhaoyang You, Tao Shen, Jun Xiao, Yin Zhang, Chao Wu, Fei Wu, Yueting Zhuang, et al. Federated unsupervised representation learning. *Frontiers of Information Technology & Electronic Engineering*, 2023. 2, 3

[44] Lu Zhang, Lu Qi, Xu Yang, Hong Qiao, Ming-Hsuan Yang, and Zhiyong Liu. Automatically discovering novel visual categories with self-supervised prototype learning. *arXiv preprint arXiv:2208.00979*, 2022. 5

[45] Xinwei Zhang, Jianwen Jiang, Yutong Feng, Zhi-Fan Wu, Xibin Zhao, Hai Wan, Mingqian Tang, Rong Jin, and Yue Gao. Grow and merge: A unified framework for continuous categories discovery. *NeurIPS*, 2022. 1, 2

[46] Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery. *ICCV*, 2023. 2, 5, 7

[47] Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *CVPR*, 2021. 1, 2

[48] Zhun Zhong, Linchao Zhu, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In *CVPR*, 2021. 1