OpenIMR: Open-World Semi-Supervised Learning on Graphs

Anonymous Author(s)
Affiliation
Address
email

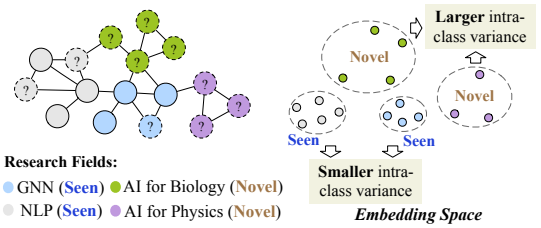
Abstract

Open-world semi-supervised learning (open-world SSL) that aims to classify instances into seen classes or newly discovered classes (referred to as novel classes) is an emergent task in deep learning community. Due to the supervised losses on labeled data, seen classes are learned more effectively than unlabeled novel classes, resulting in smaller intra-class variances within the embedding space. In contrast to vision and text data, graph data typically requires to be learned from scratch due to the lack of generic pre-trained graph encoders. Without high-quality initial representations for instances from novel classes, the imbalance of intra-class variances between seen and novel classes could be more noticeable on graph data. Through preliminary experiments and theoretical analysis, we identify that the imbalance phenomenon can inhibit the accurate discovery of novel classes. However, we also find that when the seen classes are sufficiently learned, this negative impact can be mitigated. Motivated by the findings, we propose an approach called **IM**balance-**R**educed **Open**-world SSL (OpenIMR), which enhances the learning of novel classes, narrowing the gap between intra-class variances. Meanwhile, OpenIMR learns compact representations for seen classes, which has the potential to distinguish them from novel classes and thus alleviates the negative impact caused by the imbalance issue. Experiments on seven widely-used graph benchmarks demonstrate the effectiveness of OpenIMR. The code is now available at https://github.com/anonymousforConf/OpenIMR.

1 Introduction

Despite the great successes of deep learning (DL) [16, 37, 5, 22], most DL models are developed under the closed-world setting, assuming that labeled and unlabeled data originate from the same set of classes. However, this assumption rarely holds in real-world applications, as it is impractical for humans to recognize and label all possible classes in the world. This problem could be solved by zero-shot learning (ZSL) [45, 26, 4, 25], while prior knowledge such as semantic attributes of seen and/or novel classes is required by ZSL. Considering that it could be difficult to accurately describe the prior knowledge, there has been a recent focus on open-world semi-supervised learning (open-world SSL) [2, 28, 35] — also referred to as generalized category discovery (GCD) [38, 43, 49, 7] in the field of computer vision. The goal is to classify unlabeled instances into previously seen classes or newly discovered classes. In this work, we aim to investigate open-world SSL on graph data, which is a practical yet relatively under-explored problem. Figure 1a shows an example of open-world SSL in a coauthor network, where nodes stand in for authors, edges depict coauthor relationships, and the colors of the nodes denote the authors' primary research fields. Since new fields emerge over time but lack labeled data, it is desirable to develop a model capable of classifying unlabeled authors into previously known fields or new fields.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.



(a) An example of open-world SSL in a coauthor network.

	Imbalance Rate	Separation Rate	Accuracy (%)	
			Seen	Novel
InfoNCE	1.002	1.239	72.8	72.7
+SupCon	1.071	1.271	75.1↑	71.0 ↓
+SupCon+CE	1.089	1.275	77.1↑	73.0 ↑
OpenIMR	1.048	1.430	78.3↑	75.9 ↑

(b) Effects of imbalance issue on Coauthor CS.

Figure 1: Motivations. (a) The (dashed) solid circles represent (unlabeled) labeled nodes. (b) The results are averaged over ten runs. Results on additional datasets are reported in Section 5. Calculation of the imbalance/separation rate is introduced in Appendix A.

Open-world SSL is akin to a semi-supervised clustering problem, where instances from seen classes are partially labeled. Due to the supervised losses, seen classes are often learned better than novel classes, so that they tend to have smaller intra-class variances in the embedding space [2]. Consequently, there exists an imbalance of intra-class variances between seen and novel classes, illustrated on the right side of Figure 1a. In the field of computer vision, pre-trained feature encoders such as pre-trained CNN [16, 5] and pre-trained Vision Transformer [8, 3] enable high-quality initial representations for instances from novel classes. This implies that the imbalance can be mitigated with powerful pre-trained vision encoders. In the field of graph learning, pre-trained graph neural networks (pre-trained GNNs) such as GCC [27] and Graphormer [47] typically pre-train and infer on the same/similar type(s) of graphs, as data distributions can vary significantly across different graph types (e.g., a social network and a biological network have very different data properties). This restricts the applicability of current pre-trained GNNs. Considering the absence of generic pre-trained encoders for graph learning, node representations often need to be learned from scratch. Therefore, it is necessary to discuss the imbalance phenomenon on graph datasets.

We start with preliminary experiments to explore whether the imbalance of intra-class variances poses a significant issue for open-world SSL on graph data. Taking the well-known Coauthor CS graph benchmark [31] as an example, we employ the unsupervised contrastive loss InfoNCE [36] to learn both labeled and unlabeled nodes, which enables unbiased learning of seen and novel classes. Then we successively add supervised losses, namely SupCon [21] and cross-entropy (CE), to learn the labeled nodes. By doing this, the intra-class variances of seen classes gradually decrease, leading to increasing imbalance rates as reported in the second column of Table 1b. Since the seen classes are learned more effectively, they tend to be more distinct from the novel classes in the embedding space, resulting in increased separation rates as reported in the third column of Table 1b. Under different imbalance/separation rates, we run the classic K-Means++ algorithm [1] to cluster nodes and report the accuracy on both seen and novel classes in Table 1b. We observe that (1) a higher imbalance rate can suppress the accuracy on novel classes (Cf. InfoNCE vs. InfoNCE+SupCon) as the dispersed representations of novel classes are prone to be incorrectly clustered, while (2) if the seen classes are thoroughly learned and become distinct from the novel classes, increasing the imbalance rates does not consistently limit the accuracy on novel classes (Cf. InfoNCE+SupCon vs. InfoNCE+SupCon+CE). Theoretical analysis in Section 4.1 supports the above findings. Motivated by these insights, we propose a novel approach called **IM**balance-**R**educed **Open**-world SSL (OpenIMR), which aims to effectively learn the seen classes and also reduce the imbalance of intra-class variances between seen and novel classes. OpenIMR achieves the goal as demonstrated in Table 1b.

The core design in OpenIMR is a contrastive learning (CL) framework called **P**seudo **L**abel-enhanced **C**ontrastive **L**earning (PLCL). The framework unifies CL for both labeled and unlabeled nodes, enabling the balance of learning strengths across different classes to mitigate the imbalance of intra-class variances. In terms of pseudo-labeling, a common technique is to train a classification head with a supervised loss [35, 2, 28]. However, predictions made by the head can easily over-fit the labeled classes, requiring well-designed regularizations to alleviate the issue. To address this, we convert to unsupervised K-Means clustering and select reliable cluster labels for pseudo-labeling. To deal with the unordered cluster IDs, we perform alignment between the cluster IDs and the ordered class IDs. It is important to note that the cluster IDs corresponding to novel classes can not be aligned, indicating that the class IDs for novel classes are unordered. However, since PLCL captures semantic

information by learning whether two samples share the same class ID rather than assigning samples to specific classes (as done in cross-entropy), the unordered class IDs of novel classes do not hinder the effectiveness of PLCL. Overall, the work makes the following contributions:

- We conduct preliminary experiments and theoretical analysis to identify that the imbalance of intra-class variances is a critical factor that impacts open-world SSL on graph data.
- We develop OpenIMR for open-world SSL on a graph, which emphasizes well learning the seen classes and alleviates the imbalance issue by enhancing the learning of novel classes.
- We conduct comprehensive experiments on seven real-world graphs, comparing OpenIMR with representative baselines to demonstrate its effectiveness. Especially on Coauthor Physics and ogbn-Products, OpenIMR respectively obtains gains of 11.8% and 12.5% in overall accuracy.

2 Related Work

2.1 Open-World Learning

Open-world SSL is related to but distinct from ***open-set semi-supervised learning (OSSL)*** [17, 6, 12], ***novel class discovery (NCD)*** [18, 53, 13, 44], ***zero-shot learning (ZSL)*** [45, 26, 4], and ***open-world graph learning (OGL)*** [44, 33]. In specific, OSSL regards samples of novel classes as outliers and only classifies samples of seen classes. NCD classifies the outliers into different novel classes, while it assumes that the unlabeled data consists of only novel classes. In contrast, open-world SSL not only classifies samples of seen classes but also classifies outliers into novel classes. ZSL relies on prior knowledge like semantic attributes of classes or knowledge stored in pre-trained models. It learns mappings between samples and the prior semantic information for prediction. Compared to ZSL, open-world SSL is more flexible as it does not make any assumptions about prior knowledge. A recent approach called Zero-Knowledge ZSL (ZK-ZSL) [25] has been proposed, which doesn't require semantic attributes for novel classes. However, ZK-ZSL still relies on prior knowledge of seen classes and vision features obtained from a pre-trained ResNet. In the graph domain, previous work like OpenWGL [44] and OODGAT [33] primarily study the OSSL problem on graph data, which can not deal with the open-world SSL setting. Current studies [38, 2, 28] extend methods of OSSL and NCD to solve the open-world SSL problem. However, these extensions perform poorly. That means it is necessary to design specific methods for open-world SSL.

Open-World Semi-Supervised Learning (Open-World SSL). Existing efforts in open-world SSL can be categorized into two main approaches: end-to-end methods [28, 2, 35, 7] and two-stage methods [38, 49]. The former jointly optimizes a feature encoder and a classification head in an end-to-end manner, where the cross-entropy loss as well as an unsupervised clustering objective [2, 28, 13] are widely used for model optimization. However, due to supervised losses, end-to-end methods are likely to over-fit seen classes, thereby needing well-designed regularizations to alleviate the over-fitting issue [2, 28, 35, 7]. Conversely, two-stage methods [38, 49] decouple representation learning and prediction. They employ non-parametric algorithms like (semi-supervised) K-Means to cluster instances and align clusters with classes to obtain final predictions. Since the clustering process is separated from supervised learning, the bias can be reduced. Inspired by this, we will adopt the two-stage design.

2.2 Contrastive Learning

Contrastive learning (CL) [5, 21, 11, 15, 42, 9, 32, 14, 48, 40, 41] is a common scheme for representation learning, which pulls together an anchor and its positive samples while pushing apart the anchor and its negative samples within the embedding space. CL can be applied to supervised, unsupervised, and semi-supervised settings. Most efforts of semi-supervised CL [52, 50, 46] are proposed under the closed-world setting and train a classification head with supervised losses for pseudo-labeling. Thus they tend to over-fit the labeled classes if we directly use them for open-world SSL. Therefore, we will design a semi-supervised CL scheme specifically tailored for the open-world SSL problem.

3 Problem Definition

Let $G = (\mathcal{V}, \mathcal{E}, \mathcal{X})$ be a graph, where $\mathcal{V}$ is the set of nodes, $\mathcal{E}$ is the set of edges, and $\mathcal{X}$ is the set of initial node features. $\mathcal{V}$ can be divided into two distinct sets: a labeled set $\mathcal{V}_l$ and an unlabeled set $\mathcal{V}_u$. We denote the set of classes associated with the nodes in $\mathcal{V}_l$ as $\mathcal{C}_l$ (i.e., the set of seen classes) and the set of classes associated with the nodes in $\mathcal{V}_u$ as $\mathcal{C}_u$. In the open-world SSL, $\mathcal{C}_l \neq \mathcal{C}_u$ and $\mathcal{C}_l \cap \mathcal{C}_u \neq \emptyset$. The set of novel classes $\mathcal{C}_n$ can be represented by $\mathcal{C}_n = \mathcal{C}_u \backslash \mathcal{C}_l$.

Problem 1 ***Open-world SSL on a Graph***. *Given a partially labeled graph $G = (\mathcal{V}_l, \mathcal{V}_u, \mathcal{E}, \mathcal{X}, \mathcal{Y}_l)$, where the set of available node labels is denoted as $\mathcal{Y}_l = \{y_i | v_i \in \mathcal{V}_l\}$. Let $\mathcal{Y}_u = \{y_i | v_i \in \mathcal{V}_u\}$ denote the set of class labels of $\mathcal{V}_u$, the goal is to learn a predictive function,*

$$\mathcal{F} : G = (\mathcal{V}_l, \mathcal{V}_u, \mathcal{E}, \mathcal{X}, \mathcal{Y}_l) \rightarrow \mathcal{Y}_u \tag{1}$$

In this work, we focus on a specific scenario where we have known the names of novel classes, while we lack labeled instances from novel classes and the detailed semantic information of these novel classes. In other words, we perform open-world SSL knowing the number of novel classes. We use the example depicted in Figure 1a to illustrate the situation. It is relatively easy for us to know what new research topics emerge, but in a coauthor network, providing detailed descriptions of the new fields and updating the label information promptly is challenging. As a result, we lack prior semantic knowledge and labeled data for the new research topics. In Appendix H, we discuss the cases where the number of novel classes is unknown.

4 Methodology

As motivated, the imbalance of intra-class variances between seen and novel classes can affect open-world SSL. Through a theoretical model, we build intuitions on the impacts of the imbalance issue. With a clearer picture in mind, we propose imbalance-reduced open-world SSL (OpenIMR).

4.1 Theoretical Motivation

In a latent feature space, we consider a binary clustering problem that involves N data samples. The data-generating distribution P_{XY} is characterized as a uniform mixture of two spherical Gaussian distributions. In particular, the class label Y equals either 1 or 2 with equal probability. Condition on $Y = 1$, $X|Y \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$, and $Y = 2$, $X|Y \sim \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$. K-Means aims to cluster data samples of the same class. Formally, K-Means (k=2) assigns each data sample to the nearest cluster by,

$$\hat{y} = \arg\min_j ||\boldsymbol{x} - \boldsymbol{\theta}_j||_2, j \in \{1, 2\} \tag{2}$$

where $\boldsymbol{x}$ denotes a data sample, $\hat{y}$ denotes the predicted cluster label, and $\boldsymbol{\theta}_j$ denotes the center of the j-th cluster. The cluster center is iteratively updated by the average of the data samples assigned to the cluster. Specially, we assume that the cluster IDs and class IDs are aligned, so that we can use $\hat{y}$ to denote either the cluster label or the predicted class label. It is straightforward to verify that the expectation of the converged cluster centers output by K-Means will be $\boldsymbol{\theta}_j = \mathbb{E}[X|\hat{Y} = j]$, and the expectation of clustering accuracy on each cluster will be $ACC_j = \mathbb{E}[\mathbb{1}(\hat{Y} = j)|Y = j]$.

Let σ_1^2 denote the largest eigenvalue of Σ_1, i.e., $\sigma_1^2 = \lambda_{max}(\Sigma_1)$, and σ_2^2 denote the largest eigenvalue of Σ_2, i.e., $\sigma_2^2 = \lambda_{max}(\Sigma_2)$. A large value of σ_1^2 or σ_2^2 indicates that the corresponding class has a large intra-class variance. We define the imbalance rate of intra-class variances among the two classes as $\gamma = \max(\sigma_1, \sigma_2)/\min(\sigma_1, \sigma_2)$. Intuitively, γ can affect the clustering accuracy because σ_1 and σ_2 are key parameters of the data generating distributions. Specially, when the latent classes are fairly separated with nearly no overlap, the cluster boundaries are easy to be found, and thus K-Means can correctly partition data samples with high possibility regardless of the imbalance rate. We model this by the following theorem and prove it in Appendix B.

Definition 1 *Given two spherical Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$, they are considered α-separated if*

$$||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||_2 = \alpha(\sqrt{\lambda_{max}(\Sigma_1)} + \sqrt{\lambda_{max}(\Sigma_2)})) = \alpha\,(\sigma_1 + \sigma_2)$$

where α can be represented by $||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||_2 / (\sigma_1 + \sigma_2)$, so a large value of α indicates that the two distributions are highly separated.

Now assume that $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$ are α-separated. Without loss of generality, suppose that $\sigma_1 < \sigma_2$, we have:

Theorem 1 *With $1 < \gamma < 2$, for any δ, there exists a constant $\overline{N}$, if the number of data samples $N \geq \overline{N}$, with a possibility at least 1-δ,*
(1) if $1 < \alpha < 3$, ACC_2 and σ_1 a.s. have a positive correlation;
(2) if $\alpha > 3$, $|1 - ACC_1| < 0.05$ and $|1 - ACC_2| < 0.05$.

Interpretations. The theorem illustrates two facts. ***(1) Increased imbalance rate can suppress the clustering accuracy on the class with a larger intra-class variance.*** Let σ_1 be the intra-class variance of the seen class, and σ_2 be that of the novel class, we have $\sigma_1 < \sigma_2$. Based on the first point of Theorem 1, a decrease in σ_1 leads to a decrease in the clustering accuracy ACC_2. As σ_1 decreases, the imbalance rate $\gamma = \max(\sigma_1, \sigma_2) / \min(\sigma_1, \sigma_2) = \sigma_2 / \sigma_1$ increases. Consequently, a negative correlation exists between ACC_2 and γ. In other words, an increased imbalance rate adversely affects the clustering accuracy of novel classes. ***(2) If classes are well-separated in the feature space, the clustering accuracy on each class is hard to be affected by the imbalance rate.*** If the seen class is effectively learned and can be clearly distinguished from the novel class, we can say that the two classes are α-separated, with α taking a large value. Based on the second point of Theorem 1, the imbalance rate hardly affects clustering results if α exceeds a threshold.

Hence a model is desired to (1) thoroughly learn the seen classes to encourage them to be separated from novel classes, and (2) reduce intra-class variances of novel classes to get a lower imbalance rate.

4.2 Overview of OpenIMR

Given a graph $G = (\mathcal{V}_l, \mathcal{V}_u, \mathcal{E}, \mathcal{X}, \mathcal{Y}_l)$, the inference of OpenIMR contains the following three steps:

- **Node embedding.** A GNN encoder such as GAT [39] is used to compute node representations $\mathcal{Z}_l = \{\boldsymbol{z}_i \in \mathbb{R}^d | v_i \in \mathcal{V}_l\}$ for labeled nodes and $\mathcal{Z}_u = \{\boldsymbol{z}_i \in \mathbb{R}^d | v_i \in \mathcal{V}_u\}$ for unlabeled nodes.
- **Clustering.** The classic algorithm K-Means++ [1] is applied to cluster nodes based on their representations $\mathcal{Z}_l \cup \mathcal{Z}_u$, where the number of clusters is set to be $|\mathcal{C}_l| + |\mathcal{C}_n|$.
- **ID alignment and prediction.** We perform alignment between class IDs and cluster IDs. To do this, we run the Hungarian optimal assignment algorithm [23] on the set of labeled nodes to find the optimal alignment function. Let $\mathcal{M}$ be the set of possible cluster-to-class mapping functions, the optimal alignment is found by

$$m^* = \text{argmax}_{m \in \mathcal{M}} \sum_{v_i \in \mathcal{V}_l} \mathbb{1}\left\{y_i = m\left(o_i\right)\right\}$$

  where $\mathbb{1}$ is an indicator function which takes value of 1 if $y_i = m\left(o_i\right)$ and 0 otherwise. o_i and y_i are respectively the cluster label and true class label of the i-th node in $\mathcal{V}_l$. With the optimal alignment m^*, we can predict class labels for the unlabeled nodes via $\hat{\mathcal{Y}}_u = m^*\left(\mathcal{O}_u\right) = \{m^*\left(o_i\right) | v_i \in \mathcal{V}_u\}$. Since the clusters contain both seen and novel classes, the number of clusters is more than that of seen classes. Hence a portion of cluster IDs will not match the seen class IDs.

Objective function. To optimize the GNN encoder, we employ (1) the powerful cross-entropy loss (CE) to exploit value label information, as well as (2) a **P**seudo **L**abel-enhanced **C**ontrastive **L**oss (PLCL) to enhance the learning of seen classes and reduce the intra-class variances of novel classes.

$$\mathcal{L}_{\text{OpenIMR}} = \mathcal{L}_{\text{PLCL}} + \eta \mathcal{L}_{\text{CE}} \tag{3}$$

where the hyper-parameter η is the scaling factor to weight the cross-entropy term.

4.3 Supervised Objective: CE

Based on Section 4.1, if seen classes are sufficiently learned and distinguished from novel classes, the negative impact of the imbalance issue can be mitigated. Considering that cross-entropy is a popular and powerful supervised loss, we adopt it to learn the valuable labeled data of seen classes.

$$\mathcal{L}_{\text{CE}} = -\frac{1}{|\mathcal{V}_l|} \sum_{\boldsymbol{z}_i \in \mathcal{Z}_l} \log \frac{\exp\left(W_{y_i}^\top \cdot \boldsymbol{z}_i\right)}{\exp\left(W_{y_i}^\top \cdot \boldsymbol{z}_i\right) + \sum_{j=1}^{|\mathcal{C}|} \mathbb{1}_{[j \neq y_i]} \exp\left(W_j^\top \cdot \boldsymbol{z}_i\right)} \tag{4}$$

where W is the weight matrix of a classification head, W_j is the j-th column of W, and $\mathcal{C} = \mathcal{C}_l \cup \mathcal{C}_n$.

4.4 Semi-Supervised Objective: PLCL

To enhance the learning of seen classes and reduce the intra-class variances of novel classes, we generate pseudo labels as strong supervision to guide the representation learning. A popular method of pseudo-labeling is to predict with a classification head. Since the classification head is trained with supervised losses, such a method tends to over-fit the seen classes and consequently produce unreliable pseudo labels. Therefore, we design the following pseudo-labeling method, and we will introduce how to utilize the pseudo labels for representation learning.

Pseudo-labeling. Setting the number of clusters to be $|\mathcal{C}_l| + |\mathcal{C}_n|$, we run K-Means++ over node representations $\mathcal{Z}_l \cup \mathcal{Z}_u$ to obtain cluster predictions $\mathcal{O}_l$ for labeled nodes and $\mathcal{O}_u$ for unlabeled nodes. Intuitively, samples near cluster centers are more likely to have reliable cluster predictions. Hence we define the prediction confidence of node v_i as inversely proportional to the Euclidean distance between $\boldsymbol{z}_i$ and $\boldsymbol{\theta}_{o_i}$ (i.e., inversely proportional to $||\boldsymbol{z}_i - \boldsymbol{\theta}_{o_i}||_2$), where o_i is the predicted cluster label of node v_i, and $\boldsymbol{\theta}_{o_i}$ denotes the corresponding cluster center.

To avoid erroneous training, we only use pseudo labels generated from confident predictions. According to the confidence metric, we sort $\mathcal{O}_l \cup \mathcal{O}_u$ in descending order and select the top $\rho\%$ of them as the reliable ones. Since a portion of samples have been labeled, we only supplement the predictions of unlabeled data from the above top-$\rho\%$ set, denoted as $\mathcal{O}_u^s$. With the optimal ID alignment function m^*, we derive the final pseudo labels $\hat{\mathcal{Y}}_u^s = m^*\left(\mathcal{O}_u^s\right)$.

Utilization of pseudo labels. Because novel classes lack label information, clusters of novel classes can not be aligned by Hungarian optimal assignment algorithm. As a result, the class IDs of novel classes are unordered. Such an issue makes the popular cross-entropy unsuitable for learning the pseudo-labeled data, as cross-entropy requires ordered IDs to index class prototypes for computing logits. In light of this, we convert to contrastive learning (CL) because CL learns whether two samples share the same class ID and thus does not require ordered class IDs. Specifically, our CL scheme includes embedding-level PLCL and logit-level PLCL.

Embedding-level PLCL. Given N_b randomly sampled nodes to form a mini-batch, we perform data augmentation on each sample twice to obtain positive pairs, so a batch contains $2N_b$ data points. Let (i, j) be the indices of a positive pair within the batch, we minimize the following objective,

$$\mathcal{L}_{\text{PLCL}}^{emb} = -\frac{1}{2N_b} \sum_{i=1}^{2N_b} \frac{1}{|\mathcal{P}(i)|} \sum_{j \in \mathcal{P}(i)} \log \frac{\exp\left(\boldsymbol{z}_i^\top \cdot \boldsymbol{z}_j / \tau\right)}{\sum_{k=1}^{2N_b} \mathbb{1}_{[k \neq i]} \exp\left(\boldsymbol{z}_i^\top \cdot \boldsymbol{z}_k / \tau\right)} \tag{5}$$

where $\boldsymbol{z}_i$ denotes the ℓ_2-normalized representation of the i-th data point in the batch, $\mathbb{1}_{[k \neq i]}$ is the indicator function which takes value of 1 if $k \neq i$ and 0 otherwise, and τ is the temperature parameter. Referring to SimCSE [10], we pass the same input to the GNN encoder twice. By using the standard dropout [34] twice, we can obtain the representations $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ as a "positive pair". Within a batch, $\mathcal{P}(i)$ denotes the indices of data points that share class label with the i-th data point. In fact, Eq. 5 and SupCon [21] have the same form. Different from the fully supervised SupCon, the semi-supervised PLCL constructs positive pairs based on both manual labels and pseudo labels. Specially for unlabeled nodes, Eq. 5 can be viewed as the InfoNCE loss because $|\mathcal{P}(i)|$ takes the value of 1.

Logit-level PLCL. In Eq. 4, CE enables label information of seen classes to influence the learning of classification head. As the classification head plays a key role in the learning process, we'd like to

improve its optimization with more information from unlabeled nodes. Considering that cross-entropy loss can not deal with the unordered IDs of novel classes, we design the following logit-level PLCL,

$$\mathcal{L}_{\text{PLCL}}^{logits} = -\frac{1}{2N_b}\sum_{i=1}^{2N_b}\frac{1}{|\mathcal{P}(i)|}\sum_{j\in\mathcal{P}(i)}\log\frac{\exp\left(\boldsymbol{e}_i^\top\cdot\boldsymbol{e}_j/\tau\right)}{\sum_{k=1}^{2N_b}\mathbb{1}_{[k\neq i]}\exp\left(\boldsymbol{e}_i^\top\cdot\boldsymbol{e}_k/\tau\right)} \tag{6}$$

where $\boldsymbol{e}_i = \text{l2\_norm}\left(W^T\boldsymbol{z}_i\right)$ is the ℓ_2-normalized logits of the i-th data point in the batch.

Finally, the PLCL objective combines the above two components,

$$\mathcal{L}_{\text{PLCL}} = \mathcal{L}_{\text{PLCL}}^{emb} + \mathcal{L}_{\text{PLCL}}^{logits} \tag{7}$$

The pseudocode of OpenIMR and its complexity are provided in Appendix C. We also discuss the differences between OpenIMR and the related work ORCA [2] and OpenCon [35] in Appendix D.

5 Experiments

Datasets. We conduct experiments on real-world graph benchmarks, including two citation networks *Citeseer* [22] and *ogbn-Arxiv* [19], three co-purchase networks *Amazon Photos* [31], *Amazon Computers* [31], and *ogbn-Products* [19], as well as two coauthor networks *Coauthor CS* [31] and *Coauthor Physics* [31]. Detailed information of these datasets are available in Appendix E.1.

For each dataset, we randomly sample 50% of classes as seen classes, and the remaining classes are treated as novel classes. For each seen class, we randomly sample 50 nodes (500 nodes for the large datasets *ogbn-Arxiv* and *ogbn-Products*) to the training set and another 50 nodes (500 nodes for *ogbn-Arxiv* and *ogbn-Products*) to the validation set. The remaining nodes in each dataset are assigned to the test set. For better evaluation, we use ten random seeds to derive ten data splits.

Baselines. We compare OpenIMR with end-to-end methods and two-stage methods. Appendix E.2 introduces how we select baselines for evaluation. Specifically, the end-to-end baselines include:

- **ORCA** [2] designs an uncertainty adaptive margin mechanism to control the supervised learning of labeled data, so that intra-class variances of seen classes can be similar to that of novel classes, thereby reducing the imbalance of intra-class variances between seen and novel classes.
- **ORCA-ZM** [2] removes the margin mechanism from ORCA (i.e., ORCA with Zero Margin).
- **OpenLDN** [28] leverages a classification head to generate pseudo labels for unlabeled data. Subsequently, it proceeds to train using the standard cross-entropy loss on the pseudo-labeled data.
- **OpenCon** [35] detects samples from novel classes and assigns pseudo labels to them. Then it utilizes the pseudo labels to learn more compact representations for novel classes via CL.

Consistent with OpenIMR, the two-stage baselines use K-Means++ for clustering, and then align clusters and classes for final prediction. The losses used for representation learning include:

- **InfoNCE** [36] adopts the unsupervised CL loss InfoNCE to learn both labeled and unlabeled data.
- **InfoNCE+SupCon** additionally use the supervised CL loss SupCon [21] to learn labeled data.
- **InfoNCE+SupCon+CE** further adds cross-entropy loss to enhance the learning of seen classes.

GCD [38], another work of open-world SSL in computer vision, designs a semi-supervised K-Means algorithm to force labeled samples of the same class to be clustered together. However, if a class has diverse features, such an operation could incorrectly group samples of different clusters. Based on preliminary experiments, we find that K-Means++ always performs better than the semi-supervised K-Means on the above graph benchmarks, so we adopt K-Means++ in the following experiments.

Parameter Settings. Detailed parameter settings are introduced in Appendix E.3. Here we introduce the metric for hyper-parameter selection. For closed-world problems, accuracy on the validation set is widely used for hyper-parameter selection. However, in the open-world SSL problem, relying on the validation accuracy can make models biased to seen classes as the validation set is composed of only seen classes. Hence, **we devise a new metric called SC&ACC for hyper-parameter selection**,

Table 1: Overall evaluation by accuracy (%). † denotes the two-stage variant of the original method. Bold numbers represent the best results, and underlined numbers represent the second best results.

	Citeseer			**Amazon Photos**			**Amazon Computers**			**Coauthor CS**			**Coauthor Physics**		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
ORCA-ZM	58.3	72.8	44.4	74.6	89.9	58.2	63.8	73.7	52.6	75.0	74.2	73.5	64.7	81.1	55.9
ORCA	58.2	68.0	49.0	76.2	87.1	64.9	60.9	67.8	53.7	73.9	81.6	68.3	66.2	84.8	58.2
OpenLDN	62.3	73.9	51.6	80.9	90.6	71.9	63.3	76.5	51.8	68.4	80.6	60.3	62.2	72.4	57.2
OpenCon	68.8	75.0	62.1	82.6	92.1	72.8	62.3	74.9	51.2	73.5	83.4	67.5	65.8	95.0	55.4
OpenCon†	66.7	73.7	60.0	82.9	87.9	78.1	59.4	69.0	53.2	71.0	81.9	64.8	62.6	83.8	54.4
InfoNCE	68.1	70.7	65.2	76.3	78.5	75.1	56.1	51.3	59.1	72.2	72.8	72.7	60.6	58.1	60.2
InfoNCE+SupCon	68.1	71.9	64.1	75.6	80.3	72.0	56.3	52.5	58.9	72.4	75.1	71.0	60.5	59.7	59.8
InfoNCE+SupCon+CE	68.1	73.6	62.6	76.4	80.5	72.9	55.8	54.7	56.5	74.4	77.1	73.0	62.8	79.4	56.1
OpenIMR	68.1	71.8	64.3	**83.6**	89.9	77.3	**67.8**	**77.8**	59.0	**77.1**	78.3	**75.9**	**78.0**	93.6	**72.2**

combining two popular metrics — **S**ilhouette **C**oefficient (SC) [29] and Clustering **Acc**uracy (ACC). SC is a clustering metric that measures the clustering quality based on the representations and the predicted cluster labels. SC&ACC takes into account both classification performance on seen classes as well as clustering performance on novel classes, thus alleviating the bias to seen classes.

Since open-world SSL is a transductive problem where unlabeled data is available during training, we calculate ACC on the validation set and SC on the union of validation and test sets. Under different combinations of hyper-parameters, we derive different (SC, ACC) value pairs. We normalize the values of SC via min-max normalization and also perform min-max normalization for values of ACC. Then, for each combination of hyper-parameters, we take the weighted sum of the normalized SC and the normalized ACC (the weight takes value 0.5) to obtain the value of SC&ACC. We will evaluate the effectiveness of proposed metric SC&ACC in Appendix F.

Evaluation Protocols. We adopt the widely-used clustering accuracy [38, 28, 2, 13] as the evaluation metric. Its calculation method is described in Appendix E.4. We repeat the experiments ten times with ten different data splits and report the averaged accuracy over the ten runs.

5.1 Overall Evaluation

From Table 1 and Table 2, we observe that: ***(1) Two-stage models can better balance the accuracy on seen and novel classes.*** OpenCon† is a two-stage variant of the competitive OpenCon, which runs K-Means++ over representations learned by OpenCon. Compared to OpenCon, the variant can reduce the accuracy gap between seen and novel classes. ***(2) OpenIMR obtains the best overall accuracy on most of the datasets.*** Especially on Coauthor Physics and ogbn-Products, OpenIMR respectively derives 11.8% and 12.5% gains of overall accuracy compared to the best baseline. ***(3) OpenIMR can alleviate the imbalance issue.*** Compared with InfoNCE, InfoNCE+SupCon and InfoNCE+SupCon+CE enhance the learning of seen classes, thus they achieve higher accuracy on seen classes, however, increased imbalance rates reduce the accuracy on novel classes. Conversely, compared with InfoNCE, OpenIMR not only achieves higher accuracy on seen classes but also derives better or comparable performance on novel classes. ***(4) OpenIMR is a more effective way to solve the imbalance issue.*** ORCA also notices the imbalance issue, and proposes to control the learning of seen classes. We observe that ORCA-ZM may obtain higher accuracy on seen classes compared with ORCA, indicating that controlling the learning of seen classes could suppress the utilization of valuable label information. Instead, OpenIMR reduces the imbalance issue by enhancing the learning of unlabeled novel classes, thereby guaranteeing good performance on both seen and novel classes. ***(5) OpenIMR also performs well on large datasets.***

Table 2: Evaluation on large datasets by accuracy (%).

	ogbn-Arxiv (169,343 nodes)			**ogbn-Products (2,449,029 nodes)**		
	All	Old	New	All	Old	New
ORCA-ZM	41.6	47.0	31.6	49.5	61.5	32.3
ORCA	41.6	44.7	34.6	46.8	55.5	34.3
OpenCon	32.2	31.8	31.6	43.7	46.0	43.0
OpenIMR	**43.6**	**49.2**	32.9	**62.0**	**73.6**	**44.3**

Table 3: Ablation studies by accuracy (%).

$\mathcal{L}_{\text{PLCL}}^{emb}$	$\mathcal{L}_{\text{PLCL}}^{logits}$	$\mathcal{L}_{\text{CE}}$	Citeseer			Amazon Photos			Amazon Computers			Coauthor CS			Coauthor Physics		
			All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
✓	✗	✓	**69.0**	72.4	65.2	82.8	90.0	75.5	66.4	75.5	56.1	**78.1**	80.3	76.2	64.0	94.6	51.6
✗	✓	✓	67.0	70.9	63.1	81.9	89.4	73.7	67.7	77.2	57.6	75.8	79.0	73.2	**82.5**	92.0	78.1
✓	✓	✗	67.8	71.1	64.5	80.8	81.7	80.8	55.8	54.1	57.6	76.0	77.4	75.2	58.8	58.9	57.0
✓	✗	✗	68.7	71.6	65.4	80.6	81.3	80.8	55.7	54.2	57.7	77.0	78.6	75.8	59.1	59.4	57.4
✗	✓	✗	67.2	71.0	63.7	79.7	80.2	80.0	56.5	52.7	59.3	73.4	74.9	72.7	54.6	50.1	55.5
✓	✓	✓	68.1	71.8	64.3	**83.6**	89.9	77.3	**67.8**	77.8	59.0	77.1	78.3	75.9	78.0	93.6	72.2

Table 2 shows the evaluation on larger datasets. Since we use mini batch-based clustering [30] for large datasets, the clustering effect could be impacted. Hence we refine OpenIMR for large datasets via 1) predicting with the classification head and 2) adding a widely used pair-wise clustering loss [2] to mitigate the over-fitting of seen classes. Here we observe that OpenIMR obtains the best overall accuracy on ogbn-Arxiv and ogbn-Products. Compared to OpenCon which generates pseudo labels only for samples from novel classes, OpenIMR assigns pseudo labels to unlabeled samples from both seen and novel classes, so OpenIMR augments richer semantic information to boost the performance.

5.2 Ablation Studies

We evaluate the loss of OpenIMR by testing each component in it. Table 3 shows that: (1) Every component is necessary because combining all of them derives the best or the second best performance on most of the datasets. (2) Cross-entropy (CE) ensures the effectiveness of OpenIMR. In particular, ignoring CE results in poor performance on Amazon Computers and Coauthor Physics. (3) On Amazon Photos, Amazon Computers, and Coauthor Physics, we search a larger η to enhance the learning of classification head. That is to say, the classification head plays a key role in the learning process. CE enables the classification head optimized with information from labeled data. Logit-level PLCL brings richer information from unlabeled data to improve the learning of classification head. Thus, combining logit-level PLCL and CE works well on the three datasets.

5.3 Effects of Hyper-Parameters

As shown in Figure 2, accuracy on seen classes increases with the scaling factor η ($\eta \in \{1, 10, 20\}$). However, a larger η may suppress accuracy on novel classes (Cf. results on Coauthor CS) due to the over-fitting of labeled seen classes. Specially for Coauthor Physics, increasing the value of η benefits performance on both seen and novel classes. On this dataset, a larger value of η brings a significant gain of accuracy on seen classes, indicating that the label information can be seriously underutilized if we set η to be small. Hence fully exploiting the valuable label information on Coauthor Physics improves model performance on both seen and novel classes.

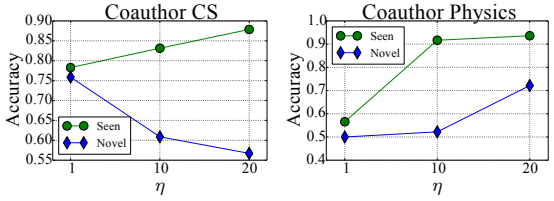


Figure 2: Effects of the scaling factor η.

We also discuss another important hyper-parameter — pesudo-label selection rate ρ in Appendix G.

6 Conclusion

This work studies open-world SSL on graph data, which is practical while presently under investigated. Based on the pilot study and the theoretical analysis, we identify that the imbalance of intra-class variances between seen and novel classes is a vital factor that impacts open-world SSL on graph data, and we propose OpenIMR to alleviate the imbalance issue. To motivate further studies, we point out the limitations of this paper and clarify the future work in Appendix I.

References

[1] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In Nikhil Bansal, Kirk Pruhs, and Clifford Stein, editors, *SODA*, pages 1027–1035. SIAM, 2007. 1, 4.2

[2] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. In *ICLR*, 2022. 1, 1, 2.1, 4.4, 5, 5.1, D, E.4, F, H

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9630–9640, 2021. 1

[4] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*, volume 9906, pages 52–68, 2016. 1, 2.1

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020. 1, 1, 2.2

[6] Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. Semi-supervised learning under class distribution mismatch. In *AAAI*, pages 3569–3576, 2020. 2.1

[7] Florent Chiaroni, Jose Dolz, Imtiaz Masud Ziko, Amar Mitiche, and Ismail Ben Ayed. Mutual information-based generalized category discovery. *CoRR*, 2022. 1, 2.1, E.2

[8] 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. 1, E.2

[9] Hongchao Fang and Pengtao Xie. CERT: contrastive self-supervised learning for language understanding. *CoRR*, 2020. 2.2

[10] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP*, pages 6894–6910, 2021. 4.4, E.3

[11] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In *NeurIPS*, 2020. 2.2

[12] Lan-Zhe Guo, Zhenyu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In *ICML*, volume 119, pages 3897–3906, 2020. 2.1

[13] Kai Han, Sylvestre-Alvise Rebuffi, Sébastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *ICLR*, 2020. 2.1, 5, E.4

[14] Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In *ICML*, pages 4116–4126, 2020. 2.2

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9726–9735, 2020. 2.2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 1

[17] Rundong He, Zhongyi Han, Yang Yang, and Yilong Yin. Not all parameters should be treated equally: Deep safe semi-supervised learning under class distribution mismatch. In *AAAI*, pages 6874–6883, 2022. 2.1

[18] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In *ICLR*, 2018. 2.1

[19] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *NeurIPS*, 2020. 5, E.1

[20] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 553–562, 2010. B.2, B.3

[21] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, pages 18661–18673, 2020. 1, 2.2, 4.4, 5

[22] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 1, 5, E.1

[23] Harold W. Kuhn. The hungarian method for the assignment problem. In *50 Years of Integer Programming 1958-2008 - From the Early Years to the State-of-the-Art*, pages 29–47, 2010. 4.2

[24] You Li, Kaiyong Zhao, Xiaowen Chu, and Jiming Liu. Speeding up k-means algorithm by gpus. *J. Comput. Syst. Sci.*, 79(2):216–229, 2013. C

[25] Zhaonan Li and Hongfu Liu. Zero-knowledge zero-shot learning for novel visual category discovery. *CoRR*, 2023. 1, 2.1

[26] Shichen Liu, Mingsheng Long, Jianmin Wang, and Michael I. Jordan. Generalized zero-shot learning with deep calibration network. In *NeurIPS*, pages 2009–2019, 2018. 1, 2.1

[27] Jiezhong Qiu, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongxia Yang, Ming Ding, Kuansan Wang, and Jie Tang. GCC: graph contrastive coding for graph neural network pre-training. In *KDD*, pages 1150–1160, 2020. 1

[28] Mamshad Nayeem Rizve, Navid Kardan, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Openldn: Learning to discover novel classes for open-world semi-supervised learning. In *ECCV*, volume 13691, pages 382–401, 2022. 1, 1, 2.1, 5, E.4, F

[29] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. 5

[30] D. Sculley. Web-scale k-means clustering. In *WWW*, pages 1177–1178, 2010. 5.1, C

[31] Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *CoRR*, 2018. 1, 5, E.1

[32] Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. A simple but tough-to-beat data augmentation approach for natural language understanding and generation. *CoRR*, 2020. 2.2

[33] Yu Song and Donglin Wang. Learning on graphs with out-of-distribution nodes. In *KDD*, pages 1635–1645, 2022. 2.1

[34] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. 4.4, E.3

[35] Yiyou Sun and Yixuan Li. Open-world contrastive learning. *CoRR*, 2022. 1, 1, 2.1, 4.4, 5, D, F

[36] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, 2018. 1, 5

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 1

[38] Sagar Vaze, Kai Hant, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *CVPR*, pages 7482–7491, 2022. 1, 2.1, 5, E.4, F, H

[39] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018. 4.2, E.3

[40] Petar Veličković, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. In *ICLR*, 2019. 2.2

[41] Yanling Wang, Jing Zhang, Shasha Guo, Hongzhi Yin, Cuiping Li, and Hong Chen. Decoupling representation learning and classification for gnn-based anomaly detection. In *SIGIR*, pages 1239–1248, 2021. 2.2

[42] Jason W. Wei and Kai Zou. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In *EMNLP-IJCNLP*, pages 6381–6387, 2019. 2.2

[43] Xin Wen, Bingchen Zhao, and Xiaojuan Qi. A simple parametric classification baseline for generalized category discovery. *CoRR*, 2022. 1

[44] Man Wu, Shirui Pan, and Xingquan Zhu. Openwgl: open-world graph learning for unseen class node classification. *Knowl. Inf. Syst.*, 63(9):2405–2430, 2021. 2.1

[45] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning - the good, the bad and the ugly. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3077–3086. IEEE Computer Society, 2017. 1, 2.1

[46] Fan Yang, Kai Wu, Shuyi Zhang, Guannan Jiang, Yong Liu, Feng Zheng, Wei Zhang, Chengjie Wang, and Long Zeng. Class-aware contrastive semi-supervised learning. In *CVPR*, pages 14401–14410, 2022. 2.2

[47] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In *NeurIPS*, pages 28877–28888, 2021. 1

[48] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In *NeurIPS*, pages 5812–5823, 2020. 2.2

[49] 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. *CoRR*, 2022. 1, 2.1, E.2

[50] Yuhang Zhang, Xiaopeng Zhang, Robert C. Qiu, Jie Li, Haohang Xu, and Qi Tian. Semi-supervised contrastive learning with similarity co-calibration. 2021. 2.2

[51] Weizhong Zhao, Huifang Ma, and Qing He. Parallel K-means clustering based on mapreduce. In *CloudCom*, volume 5931, pages 674–679, 2009. C

[52] Zhen Zhao, Luping Zhou, Lei Wang, Yinghuan Shi, and Yang Gao. Lassl: Label-guided self-training for semi-supervised learning. In *AAAI*, pages 9208–9216, 2022. 2.2

[53] 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*, pages 9462–9470, 2021. 2.1

A Calculation Method of Imbalance Rate and Separation Rate

Within an embedding space, imbalance rate measures the imbalance level of intra-class variances between seen and novel classes, and separation rate measures the extent of separation between seen and novel classes.

For each class, we calculate the mean and standard deviation of representations in the class. Given a seen class and a novel class, as well as their mean and standard deviation, we can derive the imbalance rate and the separation rate by

$$imbalance\_rate = \frac{\max\left(standard\_deviation_{seen}, standard\_deviation_{novel}\right)}{\min\left(standard\_deviation_{seen}, standard\_deviation_{novel}\right)}$$

$$separation\_rate = \frac{||mean_{seen} - mean_{novel}||_2}{standard\_deviation_{seen} + standard\_deviation_{novel}}$$

Then we calculate the averaged imbalance rate and averaged separation rate over all pairs of seen and novel classes.

B Proof of Theorem 1

B.1 Proof Skeleton

In the following proof, we use bold lowercase letters to represent vectors, regular lowercase letters to represent scalars, and regular uppercase letters to denote matrices. To simplify the problem, we first show that we can transform the K-Means clustering over d-dimension data (generated from a uniform mixture of spherical Gaussian distributions $\mathcal{N}\left(\boldsymbol{\mu_1}, \Sigma_1\right)$ and $\mathcal{N}\left(\boldsymbol{\mu_2}, \Sigma_2\right)$) to the K-Means clustering over 1-dimension data (generated from the uniform mixture of $\mathcal{N}\left(\mu_1, \sigma_1\right)$ and $\mathcal{N}\left(\mu_2, \sigma_2\right)$) without loss of generality. Then we bridge the accuracy $\{ACC_1, ACC_2\}$ and the distribution parameters $\{\mu_1, \mu_2, \sigma_1, \sigma_2\}$ to prove Theorem 1.

To do this, we introduce a partition threshold s and a function $h\left(s, \mu_1, \mu_2, \sigma_1, \sigma_2\right) = 2s - \theta_1 - \theta_2$. Specifically, θ_1 and θ_2 are cluster centers found by K-Means based on the partition threshold s. Therefore, θ_1 and θ_2 can be represented by $\mu_1, \mu_2, \sigma_1, \sigma_2$, and s. If $(\theta_1 + \theta_2)/2$ exactly equals s, the cluster centers will not change in the subsequent iterations of K-Means. In other words, the optimal partition threshold s^* is the solution of $h\left(s, \mu_1, \mu_2, \sigma_1, \sigma_2\right) = 0$. To prove Theorem 1.1, we prove that s^* and σ_1 have a negative correlation, and ACC_2 and s^* also have a negative correlation. Hence we can derive that ACC_2 is positively correlated to σ_1. To prove Theorem 1.2, we calculate value ranges of ACC_1 and ACC_2 based on value ranges of s^*, α, and γ.

B.2 Proof of Theorem 1.1

Given any d-dimension vectors $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$, there exists a rotation matrix R and a bias $\boldsymbol{b}$ that can transform $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ to x-axis, i.e.,

$$R\boldsymbol{\mu_1} + \boldsymbol{b} = \left(\mu_1, 0, ..., 0\right), R\boldsymbol{\mu_2} + \boldsymbol{b} = \left(\mu_2, 0, ..., 0\right) \tag{8}$$

Since K-Means clustering is based on the Euclidean distance metric, the clustering result will keep the same if all samples are rotated and moved with R and $\boldsymbol{b}$. Hence we can assume that the distribution P_{XY} is symmetry about the first dimension (x-axis), i.e., $\boldsymbol{\mu_1} = \left(\mu_1, 0, ..., 0\right), \boldsymbol{\mu_2} = \left(\mu_2, 0, ..., 0\right)$. Here we assume that $\mu_1 < \mu_2$ and $\mu_1 = 0$. For spherical Gaussian distributions, we can denote Σ_1 and Σ_2 by $\Sigma_1 = \mathbf{I}\sigma_1^2, \Sigma_2 = \mathbf{I}\sigma_2^2$. Without loss of generality, we assume that $\sigma_1 < \sigma_2$. A typical method for initializing cluster centers is to average the randomly selected samples. Since data samples are from P_{XY}, and P_{XY} is symmetry about the x-axis, the expectations of initial cluster centers are on the x-axis. And it is straightforward to verify that the expectations of the cluster centers predicted by each iteration of K-Means are also on the x-axis. Therefore, the expectations of final optimal cluster centers also locate on the x-axis, represented by $\boldsymbol{\theta_1^*} = \left(\theta_1^*, 0, ..., 0\right), \boldsymbol{\theta_2^*} = \left(\theta_2^*, 0, ..., 0\right)$. Up to now, we have transformed the K-Means on a d-dimension distribution to the K-Means on a 1-dimension distribution without loss of generality.

Suppose we have a d-dimensional variable $(x_1, x_2, ..., x_d)$. According to the property of marginal distribution in multivariate spherical Gaussian distribution, the probability density function of x_1 is $\mathbb{P}(x_1 = t) = \frac{1}{2}\mathbb{P}[\mathcal{N}(\mu_1, \sigma_1) = t] + \frac{1}{2}\mathbb{P}[\mathcal{N}(\mu_2, \sigma_2) = t]$. Given the partition threshold s found by K-Means, we can calculate the expectation of current cluster centers $\boldsymbol{\theta_1} = (\theta_1, 0, ..., 0)$ by

$$\begin{aligned}\theta_1 = \mathbb{E}[x_1|x_1 < s] &= \frac{\mathbb{E}[x_1 \mathbb{1}(x_1 < s)]}{\mathbb{P}(x_1 < s)} \\ &= \frac{\frac{1}{2}\mathbb{E}_{x_1 \sim \mathcal{N}(\mu_1, \sigma_1)}[x_1|x_1 < s] + \frac{1}{2}\mathbb{E}_{x_1 \sim \mathcal{N}(\mu_2, \sigma_2)}[x_1|x_1 < s]}{\frac{1}{2}\mathbb{P}_{x_1 \sim \mathcal{N}(\mu_1, \sigma_1)}(x_1 < s) + \frac{1}{2}\mathbb{P}_{x_1 \sim \mathcal{N}(\mu_2, \sigma_2)}(x_1 < s)}\end{aligned} \tag{9}$$

Let $\Phi(x)$ and $\varphi(x)$ be the cumulative distribution function (cdf) and probability density function (pdf) of the standard normal distribution respectively, we have

$$\mathbb{P}_{x_1 \sim \mathcal{N}(\mu_1, \sigma_1)}(x_1 < s) = \Phi\left(\frac{s - \mu_1}{\sigma_1}\right) \tag{10}$$

$$\mathbb{P}_{x_1 \sim \mathcal{N}(\mu_2, \sigma_2)}(x_1 < s) = \Phi\left(\frac{s - \mu_2}{\sigma_2}\right) \tag{11}$$

According to Lemma 1, we have

$$\mathbb{E}_{x_1 \sim \mathcal{N}(\mu_1, \sigma_1)}[x_1|x_1 < s] = \mu_1 \Phi\left(\frac{s - \mu_1}{\sigma_1}\right) - \sigma_1 \varphi\left(\frac{s - \mu_1}{\sigma_1}\right) \tag{12}$$

$$\mathbb{E}_{x_1 \sim \mathcal{N}(\mu_2, \sigma_2)}[x_1|x_1 < s] = \mu_2 \Phi\left(\frac{s - \mu_2}{\sigma_2}\right) - \sigma_2 \varphi\left(\frac{s - \mu_2}{\sigma_2}\right) \tag{13}$$

Substituting the above equations into Eq. 9, we can derive

$$\theta_1 = \frac{\mu_1 \Phi\left(\frac{s-\mu_1}{\sigma_1}\right) - \sigma_1 \varphi\left(\frac{s-\mu_1}{\sigma_1}\right) + \mu_2 \Phi\left(\frac{s-\mu_2}{\sigma_2}\right) - \sigma_2 \varphi\left(\frac{s-\mu_2}{\sigma_2}\right)}{\Phi\left(\frac{s-\mu_1}{\sigma_1}\right) + \Phi\left(\frac{s-\mu_2}{\sigma_2}\right)} \tag{14}$$

Similarly, we can derive

$$\theta_2 = \frac{\mu_1 - \mu_1 \Phi\left(\frac{s-\mu_1}{\sigma_1}\right) + \sigma_1 \varphi\left(\frac{s-\mu_1}{\sigma_1}\right) + \mu_2 - \mu_2 \Phi\left(\frac{s-\mu_2}{\sigma_2}\right) + \sigma_2 \varphi\left(\frac{s-\mu_2}{\sigma_2}\right)}{1 - \Phi\left(\frac{s-\mu_1}{\sigma_1}\right) + 1 - \Phi\left(\frac{s-\mu_2}{\sigma_2}\right)} \tag{15}$$

We define a function $h(s, \sigma_1, \sigma_2, \mu_1, \mu_2) = 2s - \theta_1 - \theta_2$. According to Eq. 14 and Eq. 15, θ_1 and θ_2 can be represented by $\mu_1, \mu_2, \sigma_1, \sigma_2$, and s. If s exactly equals $(\theta_1 + \theta_2)/2$, the cluster centers will not change in the subsequent iterations of K-Means. In other words, the optimal partition threshold s^* is the solution of $h(s, \mu_1, \mu_2, \sigma_1, \sigma_2) = 0$.

The optimal partition threshold s^* found by K-Means is the solution of $h(s, \sigma_1, \sigma_2, \mu_1, \mu_2) = 0$, as the cluster centers will not change in the subsequent iterations.

Now we target to discuss the monotonicity of functions $h(s)$ and $h(\sigma_1)$, which can help us explore the correlation between σ_1 and ACC_2. Before calculating the derivatives, we restrict s into a smaller interval to simplify our proof. As proved in previous studies, the optimal cluster centers found by K-Means are very close to the ground-truth cluster centers [20], that is

$$|\theta_1^* - \mu_1| < \epsilon, |\theta_2^* - \mu_2| < \epsilon \tag{16}$$

where ϵ is a rather small value ($<< 0.01(\mu_2 - \mu_1)$). Thus, we have

$$|s^* - \frac{\mu_1 + \mu_2}{2}| = |\frac{\theta_1^* + \theta_2^*}{2} - \frac{\mu_1 + \mu_2}{2}| \leq \frac{|\theta_1^* - \mu_1|}{2} + \frac{|\theta_2^* - \mu_2|}{2} < \epsilon \tag{17}$$

That is to say, $s^* \in \left[\frac{\mu_1+\mu_2}{2}-\epsilon, \frac{\mu_1+\mu_2}{2}+\epsilon\right]$. We will prove that $h(s)$ is a strictly monotone increasing function within the interval $\left[\frac{\mu_1+\mu_2}{2}-\epsilon, \frac{\mu_1+\mu_2}{2}+\epsilon\right]$.

For the second part of function $h(s)$, we calculate the derivative

$$\begin{aligned}\frac{\partial\theta_1}{\partial s} = &\frac{\frac{\mu_2-\mu_1}{\sigma_2}\varphi\left(\frac{s-\mu_2}{\sigma_2}\right)\Phi\left(\frac{s-\mu_1}{\sigma_1}\right)-\frac{\mu_2-\mu_1}{\sigma_1}\varphi\left(\frac{s-\mu_1}{\sigma_1}\right)\Phi\left(\frac{s-\mu_2}{\sigma_2}\right)}{\left(\Phi\left(\frac{s-\mu_1}{\sigma_1}\right)+\Phi\left(\frac{s-\mu_2}{\sigma_2}\right)\right)^2} \\ &+\frac{\frac{s-\mu_1}{\sigma_1}\varphi\left(\frac{s-\mu_1}{\sigma_1}\right)+\frac{s-\mu_2}{\sigma_2}\varphi\left(\frac{s-\mu_2}{\sigma_2}\right)}{\Phi\left(\frac{s-\mu_1}{\sigma_1}\right)+\Phi\left(\frac{s-\mu_2}{\sigma_2}\right)} \\ &+\frac{\left(\sigma_1\varphi\left(\frac{s-\mu_1}{\sigma_1}\right)+\sigma_2\varphi\left(\frac{s-\mu_2}{\sigma_2}\right)\right)\left(\frac{\varphi\left(\frac{s-\mu_1}{\sigma_1}\right)}{\sigma_1}+\frac{\varphi\left(\frac{s-\mu_2}{\sigma_2}\right)}{\sigma_2}\right)}{\left(\Phi\left(\frac{s-\mu_1}{\sigma_1}\right)+\Phi\left(\frac{s-\mu_2}{\sigma_2}\right)\right)^2}\end{aligned} \tag{18}$$

With the restriction of α-separated (Cf. Definition 1), we have $||\boldsymbol{\mu_1}-\boldsymbol{\mu_2}||_2 = \alpha(\sigma_1+\sigma_2)$. Since we have assumed that $\mu_2 > \mu_1$, we can derive

$$\mu_2-\mu_1 = \alpha(\sigma_1+\sigma_2) = \alpha(1+\gamma)\sigma_1 = \alpha(1+1/\gamma)\sigma_2 \tag{19}$$

As $1<\alpha<3, 1<\gamma<2$, we have

$$\frac{\mu_2-\mu_1}{\sigma_2} = \alpha(1+1/\gamma) > 2.25, \quad \frac{\mu_2-\mu_1}{\sigma_1} = \alpha(1+\gamma) > 3 \tag{20}$$

Then, we can obtain the inequalities below,

$$\varphi\left(\frac{s-\mu_2}{\sigma_2}\right) < \varphi\left(\frac{\left(\frac{\mu_1+\mu_2}{2}+\epsilon\right)-\mu_2}{\sigma_2}\right) = \varphi\left(\frac{\mu_2-\mu_1}{2\sigma_2}-\frac{\epsilon}{\sigma_2}\right) < \varphi(1.1) \tag{21}$$

$$\varphi\left(\frac{s-\mu_1}{\sigma_1}\right) < \varphi\left(\frac{\left(\frac{\mu_1+\mu_2}{2}-\epsilon\right)-\mu_1}{\sigma_1}\right) = \varphi\left(\frac{\mu_2-\mu_1}{2\sigma_1}-\frac{\epsilon}{\sigma_1}\right) < \varphi(1.45) \tag{22}$$

Similarly, we have

$$\Phi(1.5) < \Phi\left(\frac{s-\mu_1}{\sigma_1}\right) < \Phi(3), 1-\Phi(2) < \Phi\left(\frac{s-\mu_2}{\sigma_2}\right) < 1-\Phi(1.1) \tag{23}$$

According to the pdf values and cdf values of the standard normal distribution, for the first part of Eq. 18, we have

$$\frac{\frac{\mu_2-\mu_1}{\sigma_2}\varphi\left(\frac{s-\mu_2}{\sigma_2}\right)\Phi\left(\frac{s-\mu_1}{\sigma_1}\right)-\frac{\mu_2-\mu_1}{\sigma_1}\varphi\left(\frac{s-\mu_1}{\sigma_1}\right)\Phi\left(\frac{s-\mu_2}{\sigma_2}\right)}{\left(\Phi\left(\frac{s-\mu_1}{\sigma_1}\right)+\Phi\left(\frac{s-\mu_2}{\sigma_2}\right)\right)^2} < 0.8 \tag{24}$$

For the second part of Eq. 18, we have

$$\frac{\frac{s-\mu_1}{\sigma_1}\varphi\left(\frac{s-\mu_1}{\sigma_1}\right)+\frac{s-\mu_2}{\sigma_2}\varphi\left(\frac{s-\mu_2}{\sigma_2}\right)}{\Phi\left(\frac{s-\mu_1}{\sigma_1}\right)+\Phi\left(\frac{s-\mu_2}{\sigma_2}\right)} < 0.1 \tag{25}$$

For the third part of Eq. 18, we have

$$\frac{\left(\sigma_1\varphi\left(\frac{s-\mu_1}{\sigma_1}\right)+\sigma_2\varphi\left(\frac{s-\mu_2}{\sigma_2}\right)\right)\left(\frac{\varphi\left(\frac{s-\mu_1}{\sigma_1}\right)}{\sigma_1}+\frac{\varphi\left(\frac{s-\mu_2}{\sigma_2}\right)}{\sigma_2}\right)}{\left(\Phi\left(\frac{s-\mu_1}{\sigma_1}\right)+\Phi\left(\frac{s-\mu_2}{\sigma_2}\right)\right)^2} < 0.1 \tag{26}$$

Summing up the above inequalities together, we derive that $\frac{\partial\theta_1}{\partial s} < 1$. Similarly, we can prove that $\frac{\partial\theta_2}{\partial s} < 1$. Therefore, we have

$$\frac{\partial h(s)}{\partial s} = 2 - \frac{\partial\theta_1}{\partial s} - \frac{\partial\theta_2}{\partial s} > 2 - 1 - 1 = 0 \tag{27}$$

Up to now, we have proved that $h(s)$ is strictly monotone increasing within $\left[\frac{\mu_1+\mu_2}{2} - \epsilon, \frac{\mu_1+\mu_2}{2} + \epsilon\right]$. Next, we discuss the monotonicity of $h(\sigma_1)$ by calculating the derivative below,

$$\frac{\partial h(\sigma_1)}{\partial \sigma_1} = -\frac{\partial\theta_1}{\partial\sigma_1} - \frac{\partial\theta_2}{\partial\sigma_1} \tag{28}$$

We have assumed that $\mu_1 = 0$, so that we can derive

$$\begin{aligned} -\frac{\partial\theta_1}{\partial\sigma_1} = & \frac{\left(1 + \frac{(s-\mu_1)^2}{\sigma_1^2}\right)\varphi\left(\frac{s-\mu_1}{\sigma_1}\right)\left(\Phi\left(\frac{s-\mu_1}{\sigma_1}\right) + \Phi\left(\frac{s-\mu_2}{\sigma_2}\right)\right)}{\left(\Phi\left(\frac{s-\mu_1}{\sigma_1}\right) + \Phi\left(\frac{s-\mu_2}{\sigma_2}\right)\right)^2} \\ & + \frac{\left[\sigma_1\varphi\left(\frac{s-\mu_1}{\sigma_1}\right) - \mu_2\Phi\left(\frac{s-\mu_2}{\sigma_2}\right) + \sigma_2\varphi\left(\frac{s-\mu_2}{\sigma_2}\right)\right]\frac{s-\mu_1}{\sigma_1^2}\varphi\left(\frac{s-\mu_1}{\sigma_1}\right)}{\left(\Phi\left(\frac{s-\mu_1}{\sigma_1}\right) + \Phi\left(\frac{s-\mu_2}{\sigma_2}\right)\right)^2} \end{aligned} \tag{29}$$

The denominator of Eq. 29 is positive. For the only negative term in the numerator of Eq. 29, we can obtain that

$$\begin{aligned} & \mu_2\Phi\left(\frac{s-\mu_2}{\sigma_2}\right)\frac{s-\mu_1}{\sigma_1^2}\varphi\left(\frac{s-\mu_1}{\sigma_1}\right) \\ & = (\mu_2 - \mu_1)\,\Phi\left(\frac{s-\mu_2}{\sigma_2}\right)\frac{s-\mu_1}{\sigma_1^2}\varphi\left(\frac{s-\mu_1}{\sigma_1}\right) \\ & < \frac{1}{2}(\mu_2 - \mu_1)\frac{s-\mu_1}{\sigma_1^2}\varphi\left(\frac{s-\mu_1}{\sigma_1}\right)\left(\Phi\left(\frac{s-\mu_1}{\sigma_1}\right) + \Phi\left(\frac{s-\mu_2}{\sigma_2}\right)\right) \\ & < \left(1 + \frac{(s-\mu_1)^2}{\sigma_1^2}\right)\varphi\left(\frac{s-\mu_1}{\sigma_1}\right)\left(\Phi\left(\frac{s-\mu_1}{\sigma_1}\right) + \Phi\left(\frac{s-\mu_2}{\sigma_2}\right)\right). \end{aligned} \tag{30}$$

That is to say, the only negative term in the numerator is smaller than one of the positive term in the numerator. Therefore, we have $-\frac{\partial\theta_1}{\partial\sigma_1} > 0$. Similarly, we can derive that $-\frac{\partial\theta_2}{\partial\sigma_1} > 0$. Thus, we have $\frac{\partial h(\sigma_1)}{\partial\sigma_1} > 0$, i.e., $h(\sigma_1)$ is is a strictly monotone increasing function.

So far, we have proved that both $h(s)$ and $h(\sigma_1)$ are strictly monotone increasing functions. Now we aim to discuss the correlation between σ_1 and ACC_2 based on the above proved conclusions. Firstly, we formulate ACC_2 below,

$$\begin{aligned} ACC_2 = & \mathbb{E}\left[\mathbb{1}\left(\hat{Y} = 1\right)|Y = 1\right] = \mathbb{E}\left[\mathbb{1}(x_1 > s)\,|Y = 1\right] \\ = & \mathbb{P}_{x_1\sim\mathcal{N}(\mu_2,\sigma_2)}(x_1 > s) = 1 - \Phi\left(\frac{s-\mu_2}{\sigma_2}\right) \end{aligned} \tag{31}$$

We can prove that ACC_2 and σ_1 a.s. have a positive correlation via following steps:

- Given two candidate values σ_1' and σ_1'' for σ_1, suppose that $\sigma_1' < \sigma_1''$, we have $h_{\sigma_1'}(s) < h_{\sigma_1''}(s)$ because $h(\sigma_1)$ is strictly monotone increasing.
- Suppose that $s^{*\prime}$ is the solution of $h_{\sigma_1'}(s) = 0$, and $s^{*\prime\prime}$ is the solution for $h_{\sigma_1''}(s) = 0$, we have $h_{\sigma_1''}(s^{*\prime}) > h_{\sigma_1'}(s^{*\prime}) = h_{\sigma_1''}(s^{*\prime\prime}) = 0$. Considering that $h(s)$ is a strictly monotone increasing function, we have $s^{*\prime} > s^{*\prime\prime}$. Thus, if $\sigma_1' < \sigma_1''$, we have $s^{*\prime} > s^{*\prime\prime}$. In other words, s^* and σ_1 have a negative correlation.

- Substituting $s^{*\prime}$ and $s^{*\prime\prime}$ into Eq. 31 ($s^{*\prime} > s^{*\prime\prime}$), we have $1-\Phi\left(\frac{s^{*\prime}-\mu_2}{\sigma_2}\right) < 1-\Phi\left(\frac{s^{*\prime\prime}-\mu_2}{\sigma_2}\right)$. That is to say, ACC_2 and s^* also has a negative correlation.
- As we have proved (1) s^* and σ_1 have a negative correlation, and (2) ACC_2 and s^* also have a negative correlation, we finally prove that ACC_2 is positively correlated to σ_1.

According to the law of large numbers and Chebyshev inequality, when we perform sampling a large number of times, i.e., generating a large amount of data samples according to P_{XY}, the calculated results should be very close to the expectations. Therefore, there exits a constant $\overline{N}$, if the number of data samples $N \geq \overline{N}$, with a possibility at least 1-δ, ACC_2 and σ_1 a.s. have a positive correlation.

B.3 Proof of Theorem 1.2

As mentioned before, the optimal cluster centers found by K-Means are very close to the ground-truth cluster centers [20], so that we can derive $s^* \in \left[\frac{\mu_1+\mu_2}{2}-\epsilon, \frac{\mu_1+\mu_2}{2}+\epsilon\right]$ according to Eq. 16 and Eq. 17. Here we relax the value range to be $s^* \in \left[\frac{\mu_1+\mu_2}{2}-\frac{\sigma_1}{2}, \frac{\mu_1+\mu_2}{2}+\frac{\sigma_2}{2}\right]$. Based on the relaxed range of s^* and the restriction that $\alpha > 3$, we have

$$\begin{aligned}
ACC_1 &= \Phi\left(\frac{s-\mu_1}{\sigma_1}\right) > \Phi\left(\frac{\frac{\mu_1+\mu_2-\sigma_1}{2}-\mu_1}{\sigma_2}\right) = \Phi\left(\frac{\mu_2-\mu_1}{2\sigma_1}-\frac{1}{2}\right) \\
&= \Phi\left(\frac{\alpha\left(1+\gamma\right)-1}{2}\right) > \Phi\left(2.5\right) > 0.99 \\
ACC_2 &= 1-\Phi\left(\frac{s-\mu_2}{\sigma_2}\right) > 1-\Phi\left(\frac{\frac{\mu_1+\mu_2+\sigma_2}{2}-\mu_2}{\sigma_2}\right) = 1-\Phi\left(\frac{\mu_1-\mu_2}{2\sigma_2}+\frac{1}{2}\right) \\
&= \Phi\left(\frac{\alpha\left(1+1/\gamma\right)-1}{2}\right) > \Phi\left(1.75\right) > 0.95
\end{aligned} \tag{32}$$

Thus, we have proved that both $|1-ACC_1|$ and $|1-ACC_2|$ are less than 0.05.

B.4 Proof of Lemma 1

Lemma 1 *For $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, the expectation of truncated normal distribution is*

$$\mathbb{E}_{x\sim\mathcal{N}(\mu,\sigma^2)}\left[x|a<x<b\right] = \mu\left[\Phi\left(\beta\right)-\Phi\left(\alpha\right)\right] - \sigma\left[\varphi\left(\beta\right)-\varphi\left(\alpha\right)\right] \tag{33}$$

where $\alpha = \frac{a-\mu}{\sigma}$ and $\beta = \frac{b-\mu}{\sigma}$.

Proof: Converting the expectation into integral form, we have

$$\begin{aligned}
\mathbb{E}_{x\sim\mathcal{N}(\mu,\sigma)}\left[x|a<x<b\right] &= \int_a^b \frac{x}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\
&= \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \frac{(\sigma z+\mu)}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\
&= \frac{\sigma}{\sqrt{2\pi}} \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} z e^{-\frac{z^2}{2}} dz + \mu \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} dz \\
&= -\sigma \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}\Big|_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} + \mu\left[\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)\right] \\
&= \mu\left[\Phi\left(\beta\right)-\Phi\left(\alpha\right)\right] - \sigma\left[\varphi\left(\beta\right)-\varphi\left(\alpha\right)\right]
\end{aligned} \tag{34}$$

C Pseudocode and Complexity Analysis of OpenIMR

Algorithm 1 shows the pseudocode of OpenIMR, and Figure 3 illustrates the overview of OpenIMR. We denote the dimension of node representations as d, the number of nodes in a graph as N, the number of labeled nodes as M, the batch size as N_b, the number of clusters as K, and the number of iterations in K-Means as T. The complexity analysis of OpenIMR can be divided into five parts:

- The time complexity of CE-based supervised learning is $\mathcal{O}(MdK)$. If the size of labeled nodes is large, we can sample a subset of labeled nodes to reduce M;
- The time complexity of PLCL is $\mathcal{O}(NdN_b)$;
- The time complexity of K-Means clustering is $\mathcal{O}(TNdK)$;
- The time complexity of Hungarian optimal assignment algorithm is $\mathcal{O}(M^3)$. If the size of labeled nodes is large, we can sample a subset of labeled nodes for computation;
- The time complexity of selecting reliable pseudo labels is $\mathcal{O}(NlogN)$ because we need run a sorting algorithm.

Considering that N_b, d, K, T and M are relatively small, the complexity of OpenIMR is approximately about $O(NlogN)$. In addition, acceleration of K-Means has been widely studied [51, 24, 30], which can help us extend OpenIMR to larger scale datasets.

Algorithm 1 OpenIMR

Input: A graph $G = (\mathcal{V}_l, \mathcal{V}_u, \mathcal{E}, \mathcal{X}, \mathcal{Y}_l)$, the number of classes num_class, a feature encoder g_{ϕ_1}, a classification head f_{ϕ_2}, the scaling factor η, and the pseudo-label selection rate ρ.

Output: Predicted class labels $\hat{\mathcal{Y}}_u$ for the unlabeled nodes $\mathcal{V}_u$.

for $t \leftarrow 1$ to num_batch **do**
  $\mathcal{B} \leftarrow$ SampleMiniBatch $(\mathcal{V}_l \cup \mathcal{V}_u)$
  $\phi_1 \leftarrow$ Adam with loss $\mathcal{L}_{\text{InfoNCE}}(\phi_1; \mathcal{B})$
end for
$\hat{\mathcal{Y}}_u^s, \hat{\mathcal{Y}}_u \leftarrow$ Pseudo-labeling $(G, num\_class, \rho, g_{\phi_1})$;
while not MaxEpoch **do**
  for $t \leftarrow 1$ to num_batch **do**
    $\mathcal{B} \leftarrow$ SampleMiniBatch $(\mathcal{V}_l \cup \mathcal{V}_u)$
    Compute $\mathcal{L}_{\text{CE}}(\phi_1, \phi_2; \mathcal{V}_l)$
    Compute $\mathcal{L}_{\text{PLCL}}^{emb}\left(\phi_1; \mathcal{B}, \mathcal{Y}_l \cup \hat{\mathcal{Y}}_u^s\right)$
    Compute $\mathcal{L}_{\text{PLCL}}^{logits}\left(\phi_1, \phi_2; \mathcal{B}, \mathcal{Y}_l \cup \hat{\mathcal{Y}}_u^s\right)$
    $\phi_1, \phi_2 \leftarrow$ Adam $\left(\mathcal{L}_{\text{PLCL}}^{emb} + \mathcal{L}_{\text{PLCL}}^{logits} + \eta\mathcal{L}_{\text{CE}}\right)$
  end for
  $\hat{\mathcal{Y}}_u^s, \hat{\mathcal{Y}}_u \leftarrow$ Pseudo-labeling $(G, num\_class, \rho, g_{\phi_1})$
end while
Return: $\hat{\mathcal{Y}}_u$

Algorithm 2 Pseudo-labeling

Input: A graph $G = (\mathcal{V}_l, \mathcal{V}_u, \mathcal{E}, \mathcal{X}, \mathcal{Y}_l)$, the number of clusters $num\_cluster$, the pseudo-label selection rate ρ, a feature encoder g_{ϕ_1}.

Output: Pseudo labels $\hat{\mathcal{Y}}_u^s$ of a subset of unlabeled nodes, and predicted class labels $\hat{\mathcal{Y}}_u$ of all the unlabeled nodes.

Compute representations for both labeled and unlabeled nodes $\mathcal{Z}_l \cup \mathcal{Z}_u = g_{\phi_1}(G)$;
$\mathcal{O}_l \cup \mathcal{O}_u =$ K-Means++ $(\mathcal{Z}_l \cup \mathcal{Z}_u, num\_cluster)$;
Find the ID alignment $m^* \leftarrow \text{argmax}_{m \in \mathcal{M}} \sum_{v_i \in \mathcal{V}_l} \mathbb{1}\{y_i = m(o_i)\}$;
Select the top-$\rho\%$ confident cluster labels and filter that of labeled nodes to obtain $\mathcal{O}_u^s$;
$\hat{\mathcal{Y}}_u^s \leftarrow m^*(\mathcal{O}_u^s)$, $\hat{\mathcal{Y}}_u \leftarrow m^*(\mathcal{O}_u)$;
Return: $\hat{\mathcal{Y}}_u^s$ and $\hat{\mathcal{Y}}_u$

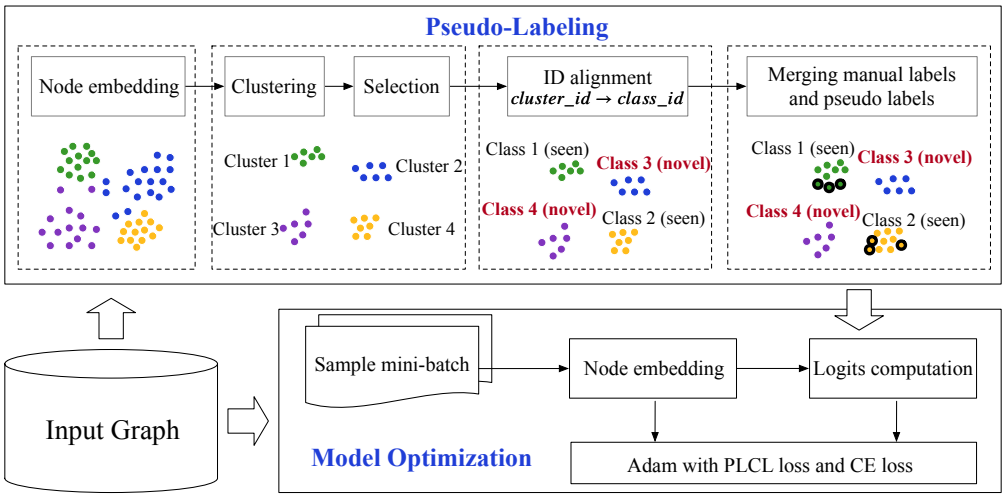


Figure 3: Overview of OpenIMR. The bold circles in the last phase of pseudo-labeling represent the manually labeled data.

D Discussion: OpenIMR vs. ORCA and OpenCon

ORCA [2] proposes the setting of open-world SSL and also explores the imbalance issue. The underlying idea of ORCA is to control supervised learning of labeled data, so that the intra-class variances of seen classes can be similar to that of novel classes, thereby reducing the imbalance rate. However, the intra-class variances of seen classes are affected to be larger than before. In other words, ORCA may not fully exploit the valuable label information. This limitation could be amplified on the more challenging graph-based tasks (Cf. comparison between ORCA-ZM and ORCA in Table 1). Conversely, OpenIMR emphasizes well learning seen classes when dealing with the imbalance issue.

Semi-supervised CL is a core technique in OpenIMR. OpenCon [35] also designs a semi-supervised CL scheme for solving the open-world SSL problem. However, it only generates pseudo labels for unlabeled instances from novel classes. In contrast, OpenIMR generates pseudo labels for unlabeled instances from both seen and novel classes, aiming to better learning seen classes. According to Theorem 1, learning more compact representations for instances of seen classes potentially separate seen classes from novel classes, which can help alleviate the negative impact of the imbalance issue.

E Experimental Settings

We run all the experiments on a server with NVIDIA RTX A6000 GPUs and 256 GB RAM.

E.1 Datasets

We conduct experiments on the following datasets, and Table 4 shows their statistics.

- **Citeseer** [22] and **ogbn-Arxiv** [19] are citation networks, where nodes represent scientific publications, and edges represent citation relationships.
- **Amazon Photos** [31], **Amazon Computers** [31], and **ogbn-Products** [19] are three Amazon co-purchase graphs, where nodes represent Amazon products, and edges represent that two products are frequently bought together.
- **Coauthor CS** [31] and **Coauthor Physics** [31] are two coauthor networks from the Microsoft academic graph, where nodes represent authors, and edges represent coauthor relationships.

E.2 Baselines

As an emergent and practical task in deep learning community, open-world SSL is being continuously researched in computer vision, but it remains under-explored in the domain of graph learning. We

Table 4: Statistics of the used datasets

Graph	#Nodes	#Edges	#Features	#Classes
Citeseer	3,327	4,676	3,703	6
Amazon Photos	7,650	119,082	745	8
Amazon Computers	13,752	245,861	767	10
Coauthor CS	18,333	818,94	6,805	15
Coauthor Physics	34,493	247,962	8,415	5
ogbn-Arxiv	169,343	1,166,243	100	40
ogbn-Products	2,449,029	61,859,140	128	47

Table 5: Hyper-parameters settings of OpenIMR on different datasets.

	Citeseer	Amazon Photos	Amazon Computers	Coauthor CS	Coauthor Physics	ogbn Arxiv	ogbn Products
max_epoch	20	20	20	20	20	20	10
batch size	2048	2048	2048	2048	2048	4096	4096
learning rate	1e-3	1e-2	1e-2	1e-4	1e-3	1e-2	1e-2
η	1	{10, 20}	{10, 20}	1	{10, 20}	1	1
τ	0.7	0.07	0.07	0.7	0.07	0.7	0.7
$\rho\,(\%)$	25	75	75	75	75	25	75

have surveyed some concurrent work [49, 7]. However, these concurrent studies are proposed for vision tasks and presently do not release their source code. More importantly, they require pre-trained feature encoders like ViT-B-16 [8]. Considering that graph domain presently lack generic pre-trained GNNs, these concurrent work can not well fit open-world SSL on graph datasets. In fact, we have compared OpenIMR with published open-world SSL methods which do not emphasize a powerful pre-trained feature encoder. These baselines are proposed for vision tasks. Thus, we extend them to solve our setting by replacing the vision encoder with a graph encoder.

E.3 Hyper-Parameter Settings

For all models, we use GAT [39] as the feature encoder and Adam optimizer with weight decay 1e-4 to optimize models. We set the number of GAT layers to 2, the hidden dimension to 128, the number of attention heads to 8, and the dropout rate [34] to 0.5. Due to the use of dropout strategy, we follow SimCSE [10] to pass the same input to the GNN encoder twice to obtain the positive pairs for CL.

Settings of maximal epochs and batch size. Empirically, we set the maximal training epochs (max_epoch) to 100 for end-to-end models and 20 for two-stage models. Specially, we set the max_epoch to 50 for ORCA and ORCA-ZM, since we find that max_epoch=50 would be better for the two baselines. For ogbn-Products which has millions of nodes, a model can converge within a small number of training epochs because the dataset can be partitioned into a large number of mini-batches during a training epoch. Hence we set max_epoch to 10 for all methods on ogbn-Products. For methods that adopt mini-batch training, we set the batch size to 2048 (4096 on large datasets ogbn-Arxiv and ogbn-Products).

Settings of model-specific parameters. For the baselines, the model-specific hyper-parameters are set to the empirical values given by the authors. For OpenIMR, we set the scaling factor η to 1, the temperature parameter τ to 0.7, and the pseudo-label selection rate $\rho\,(\%)$ to 75 by default. Specially, we observe that properly increasing the value of η can significantly boost the validation accuracy on Amazon Photos, Amazon Computers, and Coauthor Physics (validation accuracy gains $> 10\%$). Hence, we search $\eta \in \{10, 20\}$ for the three datasets. The significant gains of validation accuracy indicate the increased reliability of pseudo labels (especially for the pseudo labels of seen classes), so we use a smaller temperature parameter $\tau = 0.07$ to obtain more confident contrastive predictions. Besides, we observe that the values of validation accuracy on Citeseer and ogbn-Arxiv are obviously

lower than that on other datasets, indicating a higher learning difficulty on the two datasets. Hence we set ρ to a smaller value of 25 on Citeseer and ogbn-Arxiv to reduce the noisy from pseudo labels.

Settings of learning rate. We observe that the optimal learning rates of different end-to-end models can be very different on the same dataset, so we search the learning rate for the end-to-end models. Specially, we find that all the end-to-end baselines prefer a small learning rate of 1e-4 on Coauthor Physics, so we use such a learning rate for end-to-end models on Coauthor Physics. The CL-based two-stage models (i.e., InfoNCE, InfoNCE+SupCon, InfoNCE+SupCon+CE, and OpenIMR) share the same learning scheme, so they tend to share the same optimal learning rate on the same dataset. Intuitively, if instances have rich initial features, it could be easier for CL to capture the semantic similarity between them, and thus a smaller learning rate is preferred to enable a stable convergence. According to this, we set the default learning rate of the two-stage CL-based models to 1e-4 on Coauthor CS and Coauthor Physics, 1e-3 on Citeseer, Amazon Photos, and Amazon Computers, and 1e-2 on ogbn-Arxiv and ogbn-Products. Since a larger scaling factor of OpenIMR is used on Amazon Photos, Amazon Computers, and Coauthor Physics, the gradient produced by CE will be strengthened, so we increase the learning rate of OpenIMR by order of magnitude on the three datasets to also strengthen the gradients contributed by PLCL. Considering that the two-stage model OpenIMR can well fit a higher learning rate on Amazon Photos, Amazon Computers, and Coauthor Physics, and the Coauthor Physics is larger than Amazon Photos and Amazon Computers, we increase the learning rate to 1e-2 for the CL-based two-stage baselines on Coauthor Physics. Notably, we test other settings of learning rate for all the CL-based two-stage baselines and find that the above setting of learning rate is the optimal choice.

Summary. Hyper-parameters to be searched include: (1) the learning rate $lr \in \{$1e-2, 1e-3, 1e-4$\}$ of the end-to-end models, (2) the scaling factor $\eta \in \{1, 10, 20\}$, and (3) the best training epoches of each method. Table 5 shows the hyper-parameter settings of OpenIMR on different datasets.

E.4 Evaluation Metric: Clustering Accuracy

We employ the widely used clustering accuracy [38, 28, 2, 13] for evaluation. Given the ground-truth labels and the predictions, we run the Hungarian algorithm to obtain the best alignment between class IDs and cluster IDs. After the ID mapping, we calculate the overall accuracy. Following GCD [38], we run the Hungarian assignment only once across all classes, and calculate the resultant accuracy on seen and novel classes respectively. We repeat the experiments ten times using ten different data splits, and the reported accuracy is averaged over the ten runs.

F Evaluation of Metrics for Hyper-Parameter Search

How to search proper hyper-parameters is an under-explored task in the open-world problem. Recent studies set the hyper-parameters by searching according to validation accuracy [38] or directly using default values [2, 28]. Since the validation set only contains samples from seen classes, a model trained with the searched hyper-parameters is likely to be biased toward seen classes. Recently, OpenCon [35] proposes a validation strategy by spliting the labeled classes in into two parts (i.e., known classes and "novel" classes) to construct another open-world SSL task. Then the best hyper-parameters are selected according to the model performance on the new open-world SSL task. However, if the number of seen classes is relatively small, it is hard to construct an effective validation dataset. Besides, data distributions of the constructed open-world SSL task can be very different from that of the real open-world SSL task.

As introduced in Section 5, we propose the metric SC&ACC which additionally considers the clustering performance on novel classes to reduce the risk of over-fitting seen classes. Taking Amazon Photos and Amazon Computers as examples, Table 6 shows the performance of different metrics for hyper-parameter search. Similar experimental results can be found on other datasets. In specific, the compared metrics include the silhouette coefficient computed on the union of validation and test sets (SC) as well as the validation accuracy (ACC). We make the following observations. (1) Searching based on ACC tends to result in a larger accuracy gap between seen and novel classes, indicating that the resultant models are more biased toward seen classes. (2) The performance of SC or ACC varies across different methods. Clearly, ORCA-ZM, ORCA, and OpenLDN prefer ACC, while InfoNCE, InfoNCE+SupCon, and InfoNCE+SupCon+CE favor SC. On the contrary, SC&ACC performs more stably because every method can obtain a good performance with SC&ACC.

Table 6: Evaluation of metrics for hyper-parameter search on Amazon Photos and Amazon Computers (%). We also report the absolute gap between accuracy on seen classes and that on novel classes.

Method	**Metric**	**Amazon Photos**				**Amazon Computers**			
		All	Old	New	Gap	All	Old	New	Gap
ORCA-ZM	SC	54.4	67.3	39.0	28.3	44.1	38.0	32.8	5.2
	ACC	71.4	86.5	54.9	31.6	63.0	74.3	50.9	23.4
	SC&ACC	**74.6**	**89.9**	**58.2**	31.7	**63.8**	73.7	**52.6**	21.1
ORCA	SC	41.4	44.7	33.9	10.8	41.1	48.3	20.5	27.8
	ACC	73.3	85.8	60.3	25.5	61.0	71.4	50.4	21.0
	SC&ACC	**76.2**	**87.1**	**64.9**	22.2	60.9	67.8	**53.7**	14.1
OpenLDN	SC	48.6	48.9	46.0	2.9	43.4	42.7	28.8	13.9
	ACC	71.6	88.4	52.3	36.1	59.1	77.2	40.5	36.7
	SC&ACC	**80.9**	**90.6**	**71.9**	18.7	**63.3**	76.5	**51.8**	24.7
OpenCon	SC	83.6	90.8	76.0	14.8	59.8	69.7	51.2	18.5
	ACC	82.0	92.3	72.0	20.3	62.5	77.2	52.2	25.0
	SC&ACC	82.6	92.1	72.8	19.3	62.3	74.9	51.2	23.7
OpenCon†	SC	80.4	85.7	74.9	10.8	59.2	68.2	53.5	14.7
	ACC	81.2	91.5	71.8	19.7	60.8	72.7	53.2	19.5
	SC&ACC	**82.9**	**87.9**	**78.1**	9.80	59.4	69.0	53.2	15.8
InfoNCE	SC	77.0	77.1	77.5	0.40	56.8	52.8	58.7	5.9
	ACC	75.4	78.5	73.4	5.10	52.5	49.7	55.0	5.3
	SC&ACC	76.3	78.5	75.1	3.40	56.1	51.3	**59.1**	7.8
InfoNCE+SupCon	SC	77.2	77.5	77.3	0.20	55.7	51.1	58.4	7.3
	ACC	75.5	79.7	72.4	7.30	53.3	51.4	55.3	3.9
	SC&ACC	75.6	**80.3**	72.0	8.30	**56.3**	**52.5**	**58.9**	6.4
InfoNCE+SupCon+CE	SC	77.6	78.5	77.2	1.30	57.1	53.6	59.4	5.8
	ACC	75.5	79.7	71.8	7.90	52.9	53.2	54.3	1.1
	SC&ACC	76.4	**80.5**	72.9	7.60	55.8	**54.7**	56.5	1.8
OpenIMR ($\eta = 1$)	SC	80.9	81.1	81.7	0.60	54.8	52.4	57.9	5.5
	ACC	81.2	82.1	81.2	0.90	55.4	53.6	57.3	3.7
	SC&ACC	**81.2**	**82.1**	81.2	0.90	**55.4**	**53.6**	57.3	3.7
OpenIMR ($\eta \in \{10, 20\}$)	SC	83.3	89.3	77.1	12.2	67.8	77.9	58.7	19.2
	ACC	82.1	90.6	73.4	17.2	66.4	76.5	57.3	19.2
	SC&ACC	**83.6**	89.9	**77.3**	12.6	**67.8**	77.8	**59.0**	18.8

G Effects of Pseudo-Label Selection Rate

Pseudo-label selection rate ρ is an important hyper-parameter of OpenIMR. With the increase of ρ, accuracy on novel classes typically gets higher because more semantic information of novel classes is derived. Specially for Coauthor Physics, supplementing pseudo labels decreases the accuracy on seen classes. That is because the seen classes are already well learned (90%+ accuracy on seen classes) with a small number of pseudo labels. Hence increasing the value of ρ may amplify the negative impact of noisy pseudo labels.

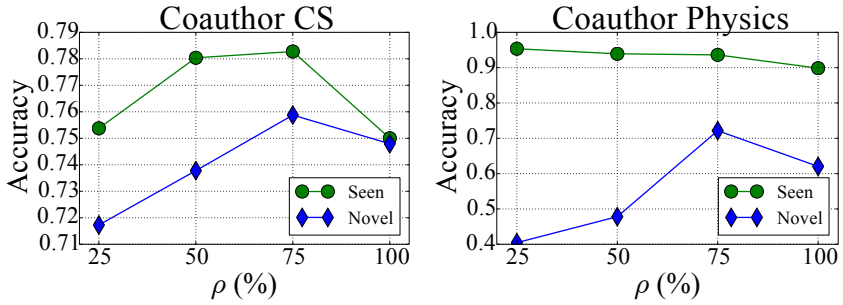


Figure 4: Effects of pseudo-label selection rate ρ.

Table 7: Evaluation of the heuristic method for estimating the number of novel classes.

	Citeseer	**Amazon Photos**	**Amazon Computers**	**Coauthor CS**	**Coauthor Physics**
Real number	3	4	5	7	3
Estimated number	2	4	6	8	4

Table 8: Overall evaluation by accuracy without knowing the number of novel classes (%).

	Citeseer			**Amazon Photos**			**Amazon Computers**			**Coauthor CS**			**Coauthor Physics**		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
ORCA-ZM	52.2	70.1	35.1	76.5	89.8	62.4	64.6	74.0	54.3	74.8	80.9	69.9	65.2	81.1	56.6
ORCA	52.7	65.6	40.0	77.6	87.5	67.1	62.3	71.1	53.4	74.8	83.6	68.3	66.3	89.3	56.8
OpenCon	52.7	73.8	34.5	83.5	91.1	75.4	61.9	75.8	51.0	72.9	83.6	66.4	63.6	94.3	53.0
OpenIMR	**64.8**	73.6	**55.9**	82.8	88.8	**76.3**	**67.0**	**77.3**	**58.1**	**77.1**	79.1	**75.5**	**76.3**	93.1	**70.2**

H OpenIMR without Knowing the Number of Novel Classes

Here we discuss how to perform OpenIMR without knowing the number of novel classes. Following previous efforts in computer vision [38], we estimate the number of clusters via a heuristic method. Our heuristic method includes following steps:

- ***Node embedding.*** We adopt the unsupervised InfoNCE loss to learn node representations.
- ***Estimating the number of novel classes.*** We run K-Means++ in the embedding space to cluster nodes varying the number of clusters ($num\_cluster = num\_novel\_class + num\_seen\_class$). Intuitively, an unreasonable number of clusters can misleading the clustering process, resulting in low clustering accuracy, so we select the number of clusters according to the averaged validation accuracy over ten runs. After determining the number of clusters, we can derive the estimated number of novel classes ($num\_estimated$) by subtracting the number of seen classes.

We report the real and estimated number of novel classes in Table 7. We can see that the heuristic method can estimate a proper number of novel classes (especially on Amazon Photos, the estimated number of novel classes equals the real value). For the end-to-end models, they are able to automatically prune redundant classes by not utilizing all initialized classification heads [2].

Considering that the estimated number of novel classes could be sub-optimal, we set the number of novel classes in interval $[num\_estimated - 1, num\_estimated + 1]$ and treat it as a hyper-parameter. Then we run OpenIMR and search the optimal number of novel classes according to the metric SC&ACC. Comparing OpenIMR with the most competitive baselines ORCA, ORCA-ZM, and OpenCon in Table 8, we can see that OpenIMR performs better on most of the datasets.

I Limitations and Future Work

This work explores the imbalance of intra-class variances between seen and novel classes, which is a key factor to impact open-word SSL on graph data. By alleviating the imbalance issue, we target to accurately classify samples from both seen and novel classes. We mainly considers the situation in which the number of novel classes have been known. Although Appendix H provides an intuitive extension of OpenIMR for open-world SSL without knowing the number of novel classes, we think a more promising solution is worth studying.

J Societal Impact

We think this work would not bring immediate negative impacts on the society, as this work does not involve privacy Instead, we target to help humans better recognize novel knowledge. Our work can accelerate the annotation processes, as the detected novel instances and novel classes can be regarded as prior knowledge to help annotators better understand a large amount of unlabeled data.