Learning Generalizable Contrastive Representations for Graph Zero-shot Learning

Siyu Yi, Zhengyang Mao, Kangjie Zheng, Zhiping Xiao, Ziyue Qiao, Chong Chen, Xian-Sheng Hua, *Fellow, IEEE,* Yongdao Zhou, Ming Zhang, and Wei Ju

***Abstract*—This paper studies the problem of graph zero-shot learning, which aims at recognizing novel classes of nodes on the graph that are never seen during training. The key to graph zero-shot learning is establishing the mathematical relationship to transfer the prior knowledge of nodes from seen classes to unseen classes. However, the problem is largely under-explored and existing methods typically focus on acquiring supervision signals from seen classes or simply establishing connections between classes based solely on a semantic description matrix, such that the learned representations lack generalizable properties to unseen classes. To address this issue, this paper proposes GraphGCR that learns generalizable contrastive representations from the perspective of uniformity and alignment. Technically, GraphGCR leverages graph diffusion to extend supervised contrastive learning, encouraging the representations of semantics from different classes to be distributed uniformly and meanwhile achieve the alignment of node features and class semantics with the assistance of graph structural information. Moreover, to effectively enhance model generalizability, we further develop a class generator to synthesize features of unseen classes by embedding propagation and interpolation, thereby enriching the diversity of classes. Theoretical analysis also shows that our proposed framework exhibits strong discriminative property, which significantly enhances graph zero-shot learning. Experimental findings reveal that our GraphGCR achieves significant performance improvements over state-of-the-art methods across various benchmark datasets.**

***Index Terms*—Graph Neural Networks, Zero-shot Learning, Contrastive Learning**

I. Introduction

GRAPH-STRUCTURED data is prevalent across a broad spectrum of real-world scenarios, such as social networks and molecular graphs. Understanding the interactions between entities in these networks is crucial. One key task in this field is node classification, where the goal is to classify the unlabeled node given a small set of labeled nodes in a graph. Recently, there has been significant interest in using graph neural networks (GNNs) for this task [1]–[4], with notable successes. These models have shown promise in domains like multimedia analysis, where understanding relationships between different types of media content is essential.

However, the graph typically evolves in dynamic with the emergence of nodes and edges, thereby inevitably arising novel classes. For example, when a new protein is discovered in the biological interaction network, the new role of this protein needs to be recognized to facilitate the study of the interaction mechanism between it and other proteins. Besides, in citation networks, the publication of new research papers often leads to new interdisciplinary subjects. Unfortunately, traditional GNNs typically assume that all classes are fixed and covered by the classes of labeled nodes [1]. When encountering newly emerging classes, GNNs are required to collect a large amount of labeled data for the new classes to achieve satisfactory performance. However, annotating these labels is a time-consuming and costly process [5]–[8]. In other words, vanilla GNNs show the inability to deal with these dilemmas. It thus naturally raises a question: *can we effectively predict the nodes for the newly emerging classes?*

To answer this question, we adopt the concept of zero-shot learning (ZSL) [9] to classify the novel classes of nodes on the graph that are never seen during training, which is a challenging yet promising task that remains largely unexplored before. We term this kind of task as *graph zero-shot learning*. ZSL has been extensively studied in computer vision [10]–[17], focusing on recognizing samples from new classes with auxiliary semantic information, e.g., category attributes [18]. The key to ZSL is establishing the relationship between the representation space and the semantic space, and generalizing this prior knowledge to unseen classes. However, applying ZSL techniques directly to graph domains poses a significant challenge, since graphs inherently are non-independent and identically distributed, the nodes in a graph interact and influence each other, and how to capture these complex connections is a crucial challenge.

To achieve effective graph zero-shot learning, there are a handful of works to conduct preliminary exploration [19]–[21]. DGPN [19] designed a generalized graph-based model using

This paper is partially supported by the Postdoctoral Fellowship Program (Grade A) of CPSF under Grant No. BX20240239, the China Postdoctoral Science Foundation under Grant No. 2024M762201, the Sichuan Science and Technology Program under Grant Numbers 2025ZNSFSC0808 and 2025ZNSFSC1506, the National Natural Science Foundation of China under Grant Numbers 62306014, 12131001 and 62276002, the Fundamental Research Funds for the Central Universities, LPMC, KLMDASR, and the Sichuan University Interdisciplinary Innovation Fund. (*Corresponding author: Wei Ju*)

Siyu Yi is with College of Mathematics, Sichuan University, Chengdu, 610065, China. (e-mail: siyuyi@scu.edu.cn)

Zhengyang Mao, Kangjie Zheng, and Ming Zhang are with School of Computer Science, National Key Laboratory for Multimedia Information Processing, Peking University-Anker Embodied AI Lab, Peking University, Beijing, China. (e-mail: zhengyang.mao@stu.pku.edu.cn, kangjie.zheng@gmail.com, mzhang_cs@pku.edu.cn)

Zhiping Xiao is with Department of Computer Science, University of Washington, Seattle, USA. (e-mail: patxiao@uw.edu)

Ziyue Qiao is with School of Computing and Information Technology, Great Bay University, Dongguan, 523000, China. (e-mail: ziyuejoe@gmail.com)

Chong Chen and Xian-Sheng Hua are with Terminus Group, Beijing, China. (e-mail: chenchong.cz@gmail.com, huaxiansheng@gmail.com)

Yongdao Zhou is with NITFID, School of Statistics and Data Science, Nankai University, Tianjin, 300071, China. (e-mail: ydzhou@nankai.edu.cn)

Wei Ju is with College of Computer Science, Sichuan University, Chengdu, 610065, China. (e-mail: juwei@scu.edu.cn)

locality and compositionality principles based on the obtained class semantic descriptions, DBiGCN [20] introduced two dual GCNs operating in opposite directions to facilitate mutual enhancement, while GraphCEN [21] utilized multi-granularity information through two-level contrastive learning to optimize representation learning and class assignments collaboratively. Nevertheless, there still exist some inherent issues. First, these methods primarily concentrate on establishing the connection among the classes based solely on a semantic description matrix, ignoring the informative properties of class and node representations or the utilization of graph structural information, which results in sub-optimal model generalizability. Second, existing methods typically acquire supervision signals from seen classes, and those for the unseen class semantics have not been fully explored, which easily suffer from the distribution-shift problem [22] when applied to disjoint unseen classes. As such, it is highly desirable to develop a simple yet effective approach that can alleviate the distribution-shift problem, and meanwhile achieve generalizable properties.

In this work, we present a novel framework GraphGCR to address these issues. The key idea of GraphGCR is to learn generalizable contrastive representations from the perspective of uniformity and alignment, and make the desired properties inherent in the learned representations have better generalization when facing the newly emerging classes. Specifically, GraphGCR is built on an optimization framework of supervised contrastive learning [23] coupled with graph diffusion to incorporate the discriminative property and graph structural information into the learned representations, which consists of two supervision signals to regularize the feature representation learning, *i.e.*, *class uniformity* and *feature alignment*. On the one hand, class uniformity encourages the representations of semantics from different classes to be distributed uniformly and preserves maximal information on the unit hypersphere, which is crucial when generalizing to unseen classes. On the other hand, feature alignment requires the representations of node features and class semantics from the same class to be as close as possible. When node features are well-aligned with class semantics, the model can more accurately associate new nodes with the correct classes based on their semantic similarities under the blessing of class uniformity, thus improving generalization to unseen classes for graph zero-shot learning task. Further, to effectively alleviate the distribution-shift problem, we develop a class generator to synthesize features of unseen classes by embedding propagation and interpolation from seen classes, which can provide additional supervision signals and thus well guarantee the generalization to newly emerging classes. We also provide a theoretical analysis showing that by incorporating graph structural information, our proposed framework exhibits strong discriminative capabilities under finite samples, thereby significantly enhancing graph zero-shot learning. By incorporating these desired properties into the learned representations, our experiments on benchmark datasets verify that our GraphGCR achieves state-of-the-art performance, and shows better closeness within classes and preserves more uniformity between class semantics. In summary, the primary contributions of our research can be summarized as follows:

- We study a promising yet largely under-explored problem: graph zero-shot learning, involving recognizing novel classes of nodes that are never seen during training. Despite its potential impact, very few algorithms have been developed to tackle this issue within the graph domain. Our work aims to shed light on this area, offering new perspectives that could inspire future research.
- We propose a novel model to incorporate uniformity and alignment into the learned representations and synthesize features of unseen classes by embedding propagation and interpolation to enhance model generalizability.
- Our theoretical analysis shows that with finite samples, our framework incorporating graph structural information can learn informative properties of nodes and classes, thereby enhancing graph zero-shot learning.
- We conduct experiments to showcase the efficacy of our GraphGCR across multiple benchmark datasets, comparing its performance against the state-of-the-art methods. Additionally, the node representations learned by our framework exhibit notable uniformity and alignment.

II. Related Work

A. Graph Neural Networks

Given their capacity to effectively model graph-structured data, graph neural networks (GNNs) have garnered significant research interest, spanning various applications including node classification [1], [4], [24], [25], graph clustering [26]–[29], and graph classification [30]–[34]. Most earlier efforts focus on spectral-based GNNs, which define graph convolution based on spectral graph theory [35], and many recent works move to the spatial-based GNNs which aggregate and transform local information via message-passing mechanism [36]. Despite the achievements in learning effective representations, GNNs fail to recognize newly emerging classes, while our proposed GraphGCR imposes desired properties on the learned representations to transfer semantic knowledge from the seen classes to the unseen classes.

B. Zero-shot Learning

Drawing inspiration from the human cognitive ability to generalize knowledge from known entities to novel concepts based on semantic descriptions and past recognition experiences, another relevant area of research is zero-shot learning (ZSL). ZSL aims to recognize new classes by transferring semantic knowledge from known classes [9]. Early ZSL approaches focused on establishing relationships between feature representations and corresponding class descriptions in a shared embedding space, enabling the transfer of semantic representations from seen to unseen classes [10], [37]–[40]. More recent studies have explored generative-based methods to synthesize features of unseen classes [11], [12], [14], [41] and focused on local region semantic mining [16], [17]. For example, AREN [16] integrated attentive region embedding and attentive compressed second-order embedding to effectively capture both local and global semantic information from images. RGEN [17] incorporated region-based relation reasoning through a graph convolutional network, effectively

modeling relationships between local image regions to improve semantic transfer and classification accuracy for unseen classes. Moreover, compared to traditional approaches such as word embedding methods and convolutional neural networks, pre-trained models like BERT [42], GPT-3 [43], and CLIP [44] leverage unsupervised or self-supervised learning on large-scale data to better capture semantic information, thereby demonstrating superior generalization capabilities when handling new classes. However, these methods struggle to adapt effectively to graph zero-shot learning with complex graph structures. This challenge arises from the inherent differences between traditional text or image data and graph data, which is characterized by non-Euclidean structures, intricate topological relationships, and varying levels of homophily or heterophily. Effectively utilizing graph structural information during training and guiding the learning process remain challenging issues. Our GraphGCR effectively integrates the topological structure properties of graphs to transfer class semantic knowledge from seen classes to unseen classes.

C. Contrastive Learning

Another closely related research area is that of contrastive learning (CL), which has recently aroused widespread interest and is proven to be the most dominant component in self-supervised learning. CL is built based on the task of instance discrimination [45] and its underlying concept is to explicitly compare pairs of sample embeddings to push away embeddings from different samples while pulling together those from augmentations of the same sample. Recent empirical works have successfully leveraged unlabeled data to learn effective feature representations that are broadly useful in downstream tasks [21], [46]–[52].

There are also some recent works trying to understand the CL [53]–[55]. They systematically analyze the behavior of contrastive learning and theoretically study the essence of its effectiveness. Different from the above works, our GraphGCR steps further and explores graph zero-shot learning from the perspective of uniformity and alignment, and further introduces a class generator to enhance model generalizability.

III. Problem Definition & Preliminaries

Notations. Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathbf{X}\}$ be an undirected graph consisting of N nodes, where $\mathcal{V} = \{v_1, \cdots, v_N\}$ is the set of nodes, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges, and $\mathbf{X} \in \mathbb{R}^{N \times d_f}$ is the node feature matrix. The adjacency matrix of the graph is denoted by $\mathbf{A} \in \{0,1\}^{N \times N}$, where $A_{ij} = 1$ if an edge (v_i, v_j) exists in $\mathcal{E}$, and $A_{ij} = 0$ otherwise. The class set is denoted as $\mathcal{C}$, and $\mathbf{S} \in \mathbb{R}^{|\mathcal{C}| \times d_c}$ represents the class semantic description (CSD) matrix for all classes [19], where $|\cdot|$ denotes the cardinality of a set and d_c is the dimension of the CSD. The CSD should be expressive enough to reflect the intricate relationships among the classes. This matrix can be obtained using Word2vec [56] from associated textual sources. Alternatively, pre-trained models like BERT can be used to generate the CSD, providing more accurate semantic information. In this paper, Word2vec is used to save the computational cost.

Graph Zero-shot Learning. Assume that the class set $\mathcal{C}$ is divided into seen class set $\mathcal{C}_s = \{c_1, \ldots, c_{|\mathcal{C}_s|}\}$ and unseen class set $\mathcal{C}_u = \{c_{|\mathcal{C}_s|+1}, \ldots, c_{|\mathcal{C}_s|+|\mathcal{C}_u|}\}$, which satisfy $\mathcal{C}_s \cap \mathcal{C}_u = \emptyset$ and $\mathcal{C}_s \cup \mathcal{C}_u = \mathcal{C}$. In graph zero-shot learning, the classes of the labeled nodes only come from the seen class set $\mathcal{C}_s$, while those of the unlabeled nodes only come from the unseen class set $\mathcal{C}_u$. All the node features and seen node labels can participate in training. The target of the task is to recognize the class labels of the unlabeled testing nodes from the unseen class set $\mathcal{C}_u$.

Graph Convolutional Network (GCN). GCN [1] is a popular type of GNN and can be interpreted as a message-passing network [36], which encodes attributive and structural information into node representations by aggregating the information of the neighbors and propagating it to the next layer. For example, the update process of a two-layer GCN can be formalized as

$$\mathbf{H} = \delta(\hat{\mathbf{A}}\delta(\hat{\mathbf{A}}\mathbf{X}\mathbf{W}^{(1)})\mathbf{W}^{(2)}), \quad \hat{\mathbf{A}} = \tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}} \tag{1}$$

where $\mathbf{H} \in \mathbb{R}^{N \times d}$ is the embedding matrix of nodes, $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, $\tilde{\mathbf{D}}$ is the degree matrix of $\tilde{\mathbf{A}}$, $\delta(\cdot)$ is the activation function, and $\mathbf{W}^{(1)}, \mathbf{W}^{(2)}$ are the trainable weight matrices.

Graph Contrastive Learning (GCL). Inspired by noise contrastive estimation [57], GCL recently revitalizes and has facilitated significant advances in self-supervised learning on graphs [51]. The basic objective of GCL is to maximize the agreement between the semantics-invariant positive augmentations of the nodes in the graph and minimize the correlation of the negative node pairs under augmentation. Specifically, given two augmented graph views $\{\mathbf{X}^1, \mathbf{A}^1\}$ and $\{\mathbf{X}^2, \mathbf{A}^2\}$ followed by a GNN encoder via Eq. (1), we map the obtained embeddings into a shared space by a projection head (*e.g.*, a multi-layer perceptron, MLP), yielding $\mathbf{z}_i^1$ and $\mathbf{z}_i^2$, $i = 1, \ldots, N$, which formulates the contrastive loss as

$$\mathcal{L} = -\sum_{i=1}^{N} \log \frac{e^{\text{sim}(\mathbf{z}_i^1, \mathbf{z}_i^2)/\tau}}{\sum_{j=1, j\neq i}^{N} e^{\text{sim}(\mathbf{z}_i^1, \mathbf{z}_j^1)/\tau} + \sum_{j=1}^{N} e^{\text{sim}(\mathbf{z}_i^1, \mathbf{z}_j^2)/\tau}},$$

where τ denotes the temperature parameter, $\text{sim}(\mathbf{z}_1, \mathbf{z}_2)$ is the cosine similarity $\frac{\mathbf{z}_1^\top \mathbf{z}_2}{\|\mathbf{z}_1\| \cdot \|\mathbf{z}_2\|}$, and $\|\cdot\|$ is the L_2-norm. Further, [54] identified two key properties of the contrastive loss, *i.e.*, (i) uniformity of the induced distribution of the normalized features on the hypersphere, (ii) the alignment of features from positive pairs, which can simultaneously preserve maximal information of the data and assign similar features to similar samples. [54] also asymptotically proved that the contrastive loss optimizes these two properties. In this paper, we leverage graph contrastive learning to assist with graph zero-shot learning.

IV. Methodology

A. Overview

In this section, we introduce our approach called GraphGCR for graph zero-shot learning as shown in Figure 1. The main purpose of GraphGCR is to learn discriminative node representations by exploring the relationship of all classes and achieving uniformity and alignment between the node

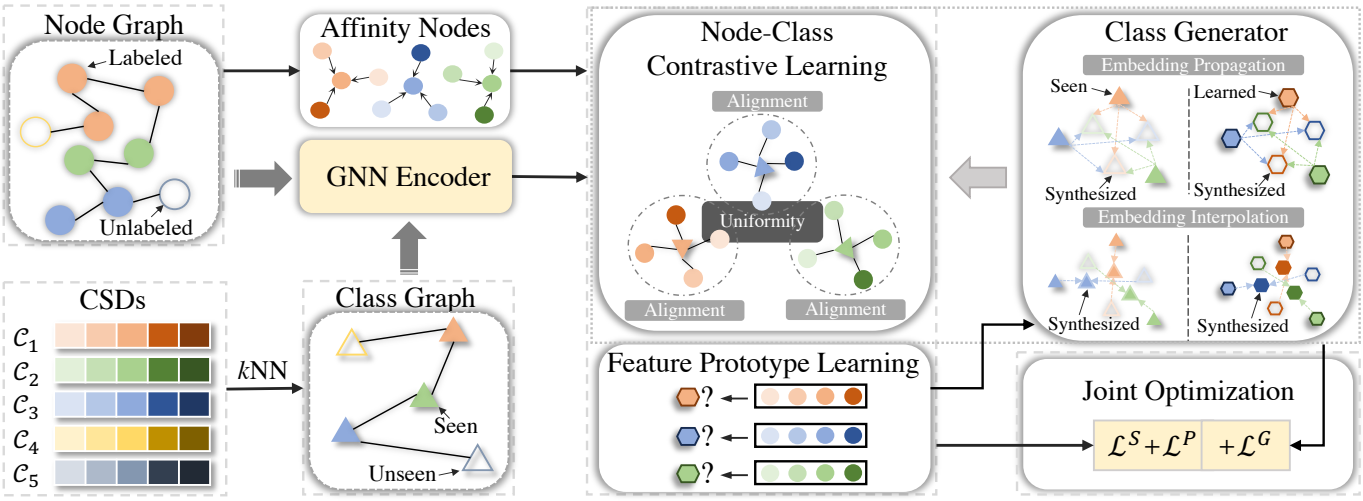


Fig. 1: Illustration of the proposed framework GraphGCR.

features and class semantics under the homophily assumption on graphs, such that the trained model can generalize well to newly emerging classes. Specifically, GraphGCR draws support from the supervised contrastive learning paradigm [23] to encourage the semantic embeddings from different seen classes to be distributed uniformly as well as realize the alignment between the node features and class semantics. Further, we synthesize features for unseen classes by our proposed class generator to improve model generalization, which enhances the diversity of the class augmentation by conducting embedding propagation and interpolation.

B. Node-Class Uniformity and Alignment Learning

In graph zero-shot learning, exploring the relationships among the classes and nodes in a unified framework for high-accuracy prediction is essential [20]. We attempt to leverage GNNs to capture both feature and relation information from node- and class-wise views. Specifically, based on the CSD matrix $\mathbf{S}$, we construct the adjacency matrix $\mathbf{A}^c$ of the classes by the k-nearest neighbor graph. With the two tuples $\{\mathbf{X}, \mathbf{A}\}$ and $\{\mathbf{S}, \mathbf{A}^c\}$ followed by a projection head to project $\mathbf{X}, \mathbf{S}$ to the same dimension, we feed them into a three-layer GCN to learn the node- and class-wise latent representations $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_N)^\top \in \mathbb{R}^{N\times d'}$ and $\mathbf{O} = (\mathbf{o}_1, \ldots, \mathbf{o}_{|\mathcal{C}|})^\top \in \mathbb{R}^{|\mathcal{C}|\times d'}$ respectively, where d' is the dimension of representations.

Note that the representations of node features and class semantics from the same class are naturally a pair of augmented forms that carry the same meaning, our approach hence leverages supervised contrastive learning [23] as a whole to formulate our framework in a uniform and aligned manner. To well adapt it to the graph domains and fully explore graph topology, we propose to adopt graph diffusion [58] to acquire richer global structural information. Formally, the graph diffusion matrix $\mathbf{F}$ is defined as

$$\mathbf{F} = \eta(\mathbf{I} - (1-\eta)\hat{\mathbf{A}})^{-1}, \tag{2}$$

which adopts the personalized PageRank [59] with $\eta \in (0, 1)$ being the teleport probability and $\hat{\mathbf{A}}$ is the normalized adjacency matrix defined in Eq. (1). The values in the i-th row of $\mathbf{F}$ can reflect the influence between node v_i and all the other nodes. We treat the nodes with $topk$ largest scores in the i-th row of $\mathbf{F}$ as the *affinity nodes* of v_i and they should possess similar semantic information to v_i. Moreover, the homophily assumption states that similar nodes typically should be close to each other and belong to the same class, these affinity nodes thus are expected to share the same class label. Then we insert the structural information upon the homophily assumption into the supervised contrastive learning to achieve both self- and affinity-alignment between nodes and classes. Accordingly, we formulate the supervised contrastive loss as

$$\begin{aligned}
\mathcal{L}^S &= -\sum_{i=1}^{N}\frac{1}{|A(i)|}\sum_{a\in A(i)}\log\frac{e^{\mathrm{sim}(\mathbf{z}_a,\mathbf{o}_{y_i})/\tau}}{\sum_{j\in\mathcal{C}_s}e^{\mathrm{sim}(\mathbf{z}_a,\mathbf{o}_j)/\tau}} \\
&= \sum_{i=1}^{N}\frac{1}{|A(i)|}\sum_{a\in A(i)} h(T_{i,a}^{\mathrm{unif}} + T_{i,a}^{\mathrm{align}}), \\
\text{with}\quad T_{i,a}^{\mathrm{unif}} &= \log\left(\sum_{j\in\mathcal{C}_s\setminus\{y_i\}}\mathrm{sim}(\mathbf{z}_a,\mathbf{o}_j)/\tau\right), \\
T_{i,a}^{\mathrm{align}} &= -\mathrm{sim}(\mathbf{z}_a,\mathbf{o}_{y_i})/\tau, \quad h(x) = \log(1+e^x),
\end{aligned} \tag{3}$$

where y_i is the ground-truth label of node v_i, $A(i)$ is the affinity node set of node v_i building on the explanation provided after Eq. (2) that always contains v_i, and τ is a scalar temperature hyper-parameter. Under the zero-shot setting, to make the participant class semantics align with at least one node representation, the defined loss in Eq. (3) only considers the seen class $\mathcal{C}_s$. Inspired by [54], $\mathcal{L}^S$ is decomposed into two regularization terms, *i.e.*, *class uniformity* ($T_{i,a}^{\mathrm{unif}}$) and *feature alignment* ($T_{i,a}^{\mathrm{align}}$). Let $\{\bar{\mathbf{z}}_a = \mathbf{z}_a/||\mathbf{z}_a||, a \in A(i)\}$ and $\{\bar{\mathbf{o}}_j = \mathbf{o}_j/||\mathbf{o}_j||, j \in \mathcal{C}_s\}$ be the normalized affinity node representations and normalized class semantic representations on the unit hypersphere, respectively. Then by theoretical analysis, we prove the following result with finite samples.

Theorem IV.1. *By minimizing the loss $\mathcal{L}^S$ in Eq.* (3) *with finite samples, $\{\bar{\mathbf{o}}_j, j \in \mathcal{C}_s\}$ are pushed away from each other to spread out over the latent space, and $\{\bar{\mathbf{z}}_a, a \in A(i)\}$ of node v_i that include $\bar{\mathbf{z}}_i$ are pulled to their corresponding normalized class semantic representation $\bar{\mathbf{o}}_{y_i}$ for any $i \in \{1, \ldots, N\}$.*

The proof of Theorem IV.1 is presented in the Appendix. Hence, minimizing $\mathcal{L}^S$ encourages the seen class representations to be uniformly distributed in the semantic space and also forces the pairs of positive augmentations (class semantics and node features from the same class, as well as their affinity nodes) to be mapped to nearby embeddings, thereby simultaneously achieving well-separated class semantic representations and discriminative node representations with maximal information preserving.

C. Feature Generation of Unseen Classes

Although uniformity and alignment are effectively achieved for the seen classes and corresponding node features, those for the unseen class semantics have not been fully explored, which inevitably limits the model generalization ability. To solve this crucial problem, we develop a class generator to produce synthetic class semantic representations by *embedding propagation* and *interpolation*. Then similar to Eq. (3), we uniformize and align the synthetic class semantics and the corresponding node features generated in the same way to learn more generalized and discriminative node representations for graph zero-shot learning.

To achieve the sharing of the synthesis way between class and node representations, we first learn a set of representative *feature prototypes* from the node representations $\mathbf{Z}$, which follows the paradigm in Eq. (3) and is supervised by

$$\mathcal{L}^P = -\sum_{i=1}^{N} \frac{1}{|A(i)|} \sum_{a \in A(i)} \log \frac{e^{\text{sim}(\mathbf{z}_a, \mathbf{w}_{y_i})/\tau}}{\sum_{j \in \mathcal{C}_s} e^{\text{sim}(\mathbf{z}_a, \mathbf{w}_j)/\tau}}, \quad (4)$$

where $\{\mathbf{w}_1, \ldots, \mathbf{w}_{|\mathcal{C}_s|}\}$ are the trainable feature prototypes with the same number as the seen classes. Each prototype should be relatively close to all nodes belonging to a certain class as well as their affinity nodes in the latent space. Meanwhile, they can be uniformly distributed under the contrastive learning framework. Further, depending on the idea that unseen classes usually share semantics with several seen classes, *e.g.*, bioinformatics combines the characteristics of biology and computer science in the citation network, we perform embedding propagation from seen class semantics and feature prototypes. Specifically, we calculate the similarities for all pairs in the semantic and prototype embeddings as $d^s_{i,j} = \text{sim}(\mathbf{o}_i, \mathbf{o}_j)$ and $d^p_{i,j} = \text{sim}(\mathbf{w}_i, \mathbf{w}_j), i, j = 1, \ldots, |\mathcal{C}_s|$. To effectively transfer knowledge from seen semantics and features to emerging ones, we utilize a weighted linear transformation to propagate new class semantics and feature prototypes in a synchronized manner,

$$\mathbf{o}'_i = \sum_{j=1}^{|\mathcal{C}_s|} w_{i,j} \mathbf{o}_j, \quad \mathbf{w}'_i = \sum_{j=1}^{|\mathcal{C}_s|} w_{i,j} \mathbf{w}_j,$$

where the weight $w_{i,j}$ plays a crucial role in governing the influence of existing class information ($\mathbf{o}_j$ and $\mathbf{w}_j$) on the updated representations ($\mathbf{o}'_j$ and $\mathbf{w}'_j$), defined as

$$w_{i,j} = \frac{e^{[(d^s_{i,j} + d^p_{i,j})/2]^{-1}}}{\sum_{j=1}^{|\mathcal{C}_s|} e^{[(d^s_{i,j} + d^p_{i,j})/2]^{-1}}}.$$

Such an operation makes the newly synthesized features fall around the original ones but not too close to them, since the representation distributions for the original seen classes and feature prototypes are roughly uniformly distributed through minimizing Eq. (3) and Eq. (4), and the negative correlation between the defined weight and similarity ensures the extension of the novel classes and features. As such, the correlative relationship across different classes is established and the cross-class transferability can be well enhanced, which can alleviate the domain-shift issue [22]. In addition, to increase the diversity of the synthesized features, we conduct embedding interpolation between the synthesized feature $\mathbf{o}'_i$ (*resp.* $\mathbf{w}'_i$) and the original one $\mathbf{o}_i$ (*resp.* $\mathbf{w}_i$), formulated as

$$\mathbf{o}''_i = \alpha \mathbf{o}_i + (1 - \alpha)\mathbf{o}'_i \quad \mathbf{w}''_i = \alpha \mathbf{w}_i + (1 - \alpha)\mathbf{w}'_i,$$

where α is a balance parameter following the uniform distribution $U[0, 1]$. Finally, based on the generated node- and class-wise features $\{\mathbf{o}'_i, \mathbf{o}''_i\}_{i=1}^{|\mathcal{C}_s|}$ and $\{\mathbf{w}'_i, \mathbf{w}''_i\}_{i=1}^{|\mathcal{C}_s|}$, we also inject the properties of uniformity and alignment by minimizing the following contrastive loss

$$\mathcal{L}^G = -\sum_{i=1}^{|\mathcal{C}_s|} \log \frac{e^{\text{sim}(\mathbf{o}'_i, \mathbf{w}'_i)/\tau} + e^{\text{sim}(\mathbf{o}''_i, \mathbf{w}''_i)/\tau}}{\sum_{j=1}^{|\mathcal{C}_s|} [e^{\text{sim}(\mathbf{o}'_i, \mathbf{w}'_j)/\tau} + e^{\text{sim}(\mathbf{o}''_i, \mathbf{w}''_j)/\tau}]}, \quad (5)$$

which contributes to establishing the connection between the unseen class semantics and node features, thereby promoting the generalization ability of the proposed GraphGCR.

D. Joint Optimization for Graph Zero-shot Learning

To enhance graph zero-short learning, we leverage the supervised contrastive learning framework to facilitate the uniformity and alignment of seen class semantics and node features as well as the synthesized unseen ones generated by embedding propagation and interpolation.

Formally, we unite the supervised contrastive losses for seen node-class pairs, node-prototype pairs, and newly synthesized node-class pairs in Eq. (3), Eq. (4), and Eq. (5) to optimize GraphGCR, *i.e.*, the total loss is

$$\mathcal{L} = \mathcal{L}^S + \epsilon \mathcal{L}^P + \eta \mathcal{L}^G, \quad (6)$$

where ϵ and η are the balance hyper-parameters to adjust the relative importance of each loss component. Furthermore, as discussed earlier and supported by Theorem IV.1, this scheme effectively mitigates the over-smoothing issue commonly observed in GNN models. By minimizing $\mathcal{L}$, it facilitates the acquisition of well-separated class semantic representations and discriminative node representations, thereby preventing the learned representations from becoming indistinguishable.

When the model reaches convergence, following the alignment paradigm, we predict the class label $\hat{y}$ of the unlabeled node from the unseen class set $\mathcal{C}_u$ by selecting the class whose

Algorithm 1 The Optimization of GraphGCR

Input: Graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathbf{X}\}$; Class set $\mathcal{C} = \mathcal{C}_s \cup \mathcal{C}_u$; Class semantic description matrix $\mathbf{S}$; Maximum iterations T_{max};
Output: Class assignments for unlabeled nodes from unseen class set $\mathcal{C}_u$;

Construct the class adjacency matrix $\mathbf{A}^c$ by kNN;
Construct the affinity node set $A(i)$ from Eq. (2);
Initialize the trainable parameters in GCN and the feature prototypes;
Set $t = 0$;
while $t \leq T_{max}$ **do**
  Update the node- and class-wise representations $\mathbf{Z}$ and $\mathbf{O}$ by GCN encoder;
  Generate the newly synthesized class semantics and feature prototypes by embedding propagation and interpolation;
  Calculate the losses $\mathcal{L}^S$, $\mathcal{L}^P$, and $\mathcal{L}^G$ in Eq. (3), Eq. (4), and Eq. (5), respectively;
  Conduct backpropagation and update the whole network in GraphGCR by minimizing $\mathcal{L}$ in Eq. (6);
  Update the feature prototypes in Eq. (4);
  Set $t = t + 1$;
end while

semantic representation has the maximum similarity with the testing node, *i.e.,*

$$\hat{y} = \text{argmax}_{y \in \mathcal{C}_u} \text{sim}(\mathbf{z}_{\text{test}}, \mathbf{o}_y),$$

where $\text{sim}(\cdot, \cdot)$ is the cosine similarity, and $\mathbf{z}_{\text{test}}$ and $\mathbf{o}_y$ are the learned representations from the GCN encoder for the testing node and the unseen class. The detailed training procedure for our proposed GraphGCR is outlined in Algorithm 1.

E. Computational Complexity Analysis

Given a graph dataset with N nodes and E edges, and assume the dimensions of the used L-layer GCN are $d_1, \ldots, d_L$. The time complexity of constructing the class adjacent matrix by the k-d tree is $O(|\mathcal{C}| \log |\mathcal{C}|)$. Suppose that the constructed class graph has E_c edges. Then we learn the representations by the GCN encoder in $O((E + E_c) \sum_{i=1}^{L} d_i d_{i-1})$. We perform embedding propagation and interpolation in $O(|\mathcal{C}_s|^2 d_L)$ and the complexity of computing the total loss is $O(N|\mathcal{C}_s| d_L topk + |\mathcal{C}_s|^2 d_L)$. Therefore, the total computational complexity of our proposed GraphGCR is $O((E + E_c) \sum_{i=1}^{L} d_i d_{i-1} + N|\mathcal{C}_s| d_L topk)$.

V. Experiment

A. Experimental Setup

Datasets. Following [19]–[21], we perform extensive experiments on three widely recognized citation datasets: Cora (7 classes), Citeseer (6 classes), and C-M10M (6 classes), where each node corresponds a publication, and the edges represent the citation relationships between two linked publications. To ensure fairness in comparison, we utilize identical settings for seen/unseen class splits as outlined in [19], *i.e.*, the labels of the same class only belong to one of the training set, validation set, and test set; the test set must contain class labels that do not appear in the training and validation sets; the number of classes in each set is finally guaranteed to sum up to the total number of classes of the data. The details of the split can be found in Table I. Other allocation of the number of classes is also feasible, as long as the aforementioned requirements are met. Additionally, we employ two types of CSDs: TEXT-CSDs (default) and LABEL-CSDs, which are generated by Bert-Tiny as described in [19] providing different class semantics.

TABLE I: The data split of three citation datasets.

Dataset	Class Split I [Train/Val/Test]	Class Split II [Train/Val/Test]
Cora	[3/0/4]	[2/2/3]
Citeseer	[2/0/4]	[2/2/2]
C-M10M	[3/0/3]	[2/2/2]

Baseline Methods. Our proposed GraphGCR is compared against various advanced methods in zero-shot learning including DAP and its variant DAP(CNN) [60], ESZSL [10], ZS-GCN and its variant ZS-GCN(CNN) [61], WDVSc [37], Hyperbolic-ZSL [62], AREN [16], and RGEN [17], which were originally developed for vision domains. Furthermore, we consider three recent methods specifically designed for graph zero-shot learning: DGPN [19], DBiGCN [20] and GraphCEN [21]. In addition, the RandomGuess is chosen as the naive baseline, which randomly assigns unseen classes to unlabeled nodes on the graph.

Implementation Details. We implement the proposed model using PyTorch and conduct all experiments using an NVIDIA GeForce RTX 3090. In all experiments, we use a three-layer GCN as the backbone encoder and we train the model from scratch by randomly initializing the network parameters, without using any pretraining. For our GraphGCR, we employ the grid search to tune the hyperparameters. The loss hyperparameters $\{\epsilon, \eta\}$ are set as $\{0.1, 1\}$ for Cora, $\{0.01, 0.5\}$ for Citeseer, $\{1, 1\}$ for C-M10M, respectively. In addition, the learning rate is selected from $\{1e\text{-}3, 1e\text{-}4, 1e\text{-}5\}$; the hidden dimensions are selected from {32, 64, 128, 256}; the number of neighbors k_n in kNN is chosen from {1, 2, ..., $|\mathcal{C}|$-1}; the number of affinity nodes $topk$ is chosen from {1, 10, 50, 100, 200}. For evaluating the performance, we adopt accuracy on the test set as the primary metric in our experiments.

B. Overall performance

In this section, we begin by comparing our GraphGCR with state-of-the-art methods for graph zero-shot learning. As outlined in Table II, our findings are summarized as follows:

- Our proposed GraphGCR consistently outperforms other strong baselines across all three datasets in different class split settings. Notably, GraphGCR achieves a remarkable improvement of 9.19% on C-M10M dataset under class split I and 4.37% on Citeseer dataset under class split II compared to the closest competitor. This indicates the superior model generalizability of our framework for graph zero-shot learning.

TABLE II: Comparison of the overall performance. The best results are in boldface and the second-best is underlined. 'Improve ↑' refers to the accuracy improvement rate of our GraphGCR relative to first-best baseline.

	Dataset	Cora	Citeseer	C-M10M
Class Split I	RandomGuess	25.35	24.86	33.21
	DAP	26.56	34.01	38.71
	DAP(CNN)	27.80	30.45	32.97
	ESZSL	27.35	30.32	37.00
	ZS-GCN	25.73	28.62	37.89
	ZS-GCN(CNN)	16.01	21.18	36.44
	WDVSc	30.62	23.46	38.12
	Hyperbolic-ZSL	26.36	34.18	35.80
	AREN	28.71	34.62	36.91
	RGEN	31.29	35.89	38.45
	DPGN	33.76	37.74	41.93
	DBiGCN	45.08	38.57	41.11
	GraphCEN	<u>48.43</u>	<u>40.77</u>	<u>44.17</u>
	GraphGCR (Ours)	**48.98**	**41.21**	**48.23**
	Improve ↑	+1.14%	+1.08%	+9.19%
Class Split II	RandomGuess	32.69	50.48	49.73
	DAP	30.22	53.30	46.79
	DAP(CNN)	29.83	50.07	46.29
	ESZSL	38.82	55.32	56.07
	ZS-GCN	29.53	52.22	55.28
	ZS-GCN(CNN)	33.20	49.27	51.37
	WDVSc	34.13	52.70	46.26
	Hyperbolic-ZSL	37.02	46.27	55.07
	AREN	36.58	49.52	55.16
	RGEN	40.37	52.63	57.33
	DPGN	48.31	58.86	61.68
	DBiGCN	46.95	58.37	66.12
	GraphCEN	<u>50.61</u>	<u>60.47</u>	<u>70.83</u>
	GraphGCR (Ours)	**52.48**	**63.11**	**72.95**
	Improve ↑	+3.69%	+4.37%	+2.99%

- Zero-shot learning methods for visions generally exhibit inferior performance compared with graph zero-shot learning methods (DPGN, DBiGCN and GraphCEN). This could be attributed to their limited ability to explore the relational information among nodes and capture the complex characteristics of graphs.
- Across all datasets, our GraphGCR surpasses existing graph zero-shot learning methods (DPGN, DBiGCN and GraphCEN) by a significant margin. This demonstrates the excellent superiority of modeling uniformity and alignment. Besides, it is beneficial to synthesize novel node features and class semantics for unseen classes to enhance generalization, further facilitating knowledge transfer from seen classes to unseen classes.

Scalability on the large-scale dataset. To assess the scalability of our approach, we test it on the large-scale ogbn-arxiv dataset. We compare the performance of our method, GraphGCR, against the latest baselines, DPGN, DBiGCN, and GraphCEN. Table III shows that our GraphGCR consistently outperforms these baselines, particularly in class split I. This indicates the superior discriminative power and generalization capabilities achieved through our learned desired representations of uniformity and alignment, as well as our developed class generator. These findings confirm that GraphGCR effectively learns effective and transferable representations and demonstrates excellent scalability.

TABLE III: The comparison (%) of DGPN, DBiGCN, GraphCEN, and our proposed GraphGCR on the large-scale ogbn-arxiv dataset for zero-shot node classification.

	DGPN	DBiGCN	GraphCEN	GraphGCR (Ours)
Class Split I	22.37	21.40	23.96	**25.78**
Class Split II	21.95	25.92	28.36	**29.47**

TABLE IV: Analysis of ablation study.

Model	Split	Cora	Citeseer	C-M10M
GraphGCR w/o $\mathcal{L}^S$	I	44.93	34.01	47.29
	II	46.30	57.19	72.12
GraphGCR w/o $\mathcal{L}^P$	I	37.44	37.36	41.74
	II	51.17	62.47	65.90
GraphGCR w/o $\mathcal{L}^G$	I	35.30	39.23	46.42
	II	49.76	60.75	71.44
GraphGCR (Ours)	I	48.98	41.21	48.23
	II	52.48	63.11	72.95

C. Ablation Study

In this part, we conduct an ablation study to verify the importance of three crucial components in GraphGCR, with the following contrast variants as follows:

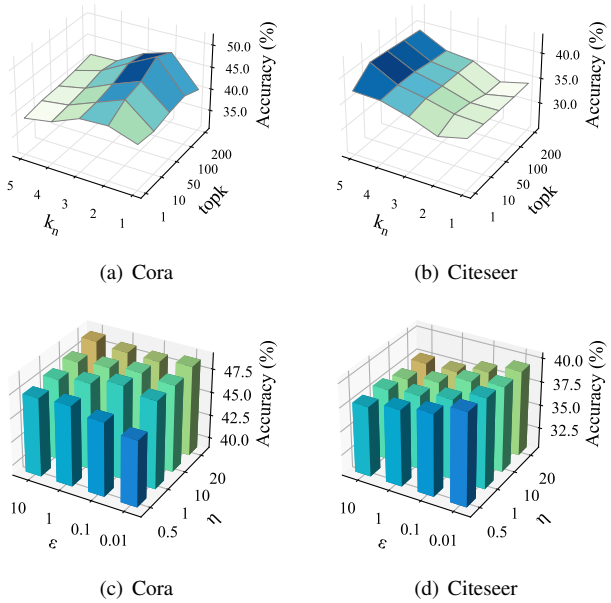
- GraphGCR w/o $\mathcal{L}^S$: Our full model without supervised contrastive loss for seen classes.
- GraphGCR w/o $\mathcal{L}^P$: Our full model without supervised contrastive loss for feature prototypes.
- GraphGCR w/o $\mathcal{L}^G$: Our full model without supervised contrastive loss for synthesized unseen classes.

The comparison results of different variants are presented in Table IV. It is evident from the table that our complete model, GraphGCR, achieves the highest performance across all three datasets under different class split settings. As such, it is necessary to build a joint framework to simultaneously capture uniformity and alignment for seen and unseen classes. In addition, we can see that different components play distinct roles for different datasets. Taking class split I as an example, on Cora dataset, GraphGCR w/o $\mathcal{L}^G$ exhibits the worst results, indicating the significance of class generator. While on Citeseer dataset, $\mathcal{L}^S$ is the most important. This variation in performance can be attributed to the inherent differences in the properties of the datasets, which results in different focuses in capturing information for our method. This further confirms the importance of each component in our proposed method, emphasizing that every component is indispensable.

Moreover, in Table V, we record the running time in seconds for our model GraphGCR and the three variant models under class split I. Under each variant, we omit the exclusive operations associated with the removed loss during training. Combining with Table IV, it can be observed that our model

TABLE V: Runtime comparison (in seconds) of ablation study under class split I.

Model	Cora	Citeseer	C-M10M
GraphGCR w/o $\mathcal{L}^S$	28.1883	53.4238	21.8016
GraphGCR w/o $\mathcal{L}^P$	29.1727	56.9088	21.5072
GraphGCR w/o $\mathcal{L}^G$	25.3496	51.7638	20.0701
GraphGCR (Ours)	33.9869	64.4500	26.4212



(a) Cora (b) Citeseer

(c) Cora (d) Citeseer

Fig. 2: The performance comparisons $w.r.t.$ different hyper-parameters on Cora and Citeseer under class split I.

GraphGCR achieves better prediction performance with only a slight increase in computational cost.

D. Sensitivity Analysis

We discuss the impacts of the hyper-parameters, *i.e.*, the number of neighbors k_n in kNN, the number of affinity nodes $topk$, and balance weights (loss hyperparameters) ϵ and η.

Effects of k_n ***and*** $topk$**.** As shown in Figures 2(a) and 2(b), we can observe that on Cora dataset, when $topk$ is fixed, the performance initially improves and then declines with k_n increasing, which indicates that class semantics in Cora exist significant discrepancies, and forcing connections can blur the distinctiveness. However, when k_n is fixed and $topk$ increases to 100, the result continues to improve, while a further increase leads to a decline. This could be due to the introduction of noise from selecting too many affinity nodes. On the other hand, increasing k_n is beneficial for Citeseer dataset as it better captures the relationships among classes, while $topk$ has a relatively less sensitive impact on Citeseer.

Effects of ϵ ***and*** η**.** As depicted in Figures 2(c) and 2(d), it can be seen that on Cora dataset, when η is fixed and relatively small, increasing ϵ to some extent positively impacts the model's predictions, highlighting the importance of the feature prototype. But further increase of ϵ may have negative consequences in some cases. Similarly, when ϵ is fixed and η is increased, the model performance also improves and exhibits greater stability. This is consistent with the ablation study, where removing $\mathcal{L}^G$ results in the most significant performance drop. However, on Citeseer dataset, the impact of these parameters shows an opposite phenomenon. When both η and ϵ are small, the model achieves the best results. This aligns with the ablation study and indicates that in Citeseer, $\mathcal{L}^S$ is the dominant factor influencing performance. Excessive η and ϵ can overshadow the effect of $\mathcal{L}^S$, leading to performance degradation.

E. Discussion on Different CSDs

In our GraphGCR, we construct class adjacency matrix $\mathbf{A}^c$ based on the CSD matrix $\mathbf{S}$. Since different CSDs can provide different semantic information for the class graph, the model performance on graph zero-shot learning that heavily relies on label semantics can be significantly influenced by the chosen CSD. Hence, we analyze the impacts of two types of CSDs [19], as shown in Table VI. From the table, we observe that TEXT-CSDs typically outperform LABEL-CSDs, which is consistent with previous researches [19], [20]. This highlights the better ability of natural language [63], [64] to capture relational information among label semantics. Moreover, regardless of the type of CSD used, our proposed GraphGCR achieves the best performance, further demonstrating the superiority of modeling uniformity and alignment as well as the generalizability of class generators.

F. Case Study

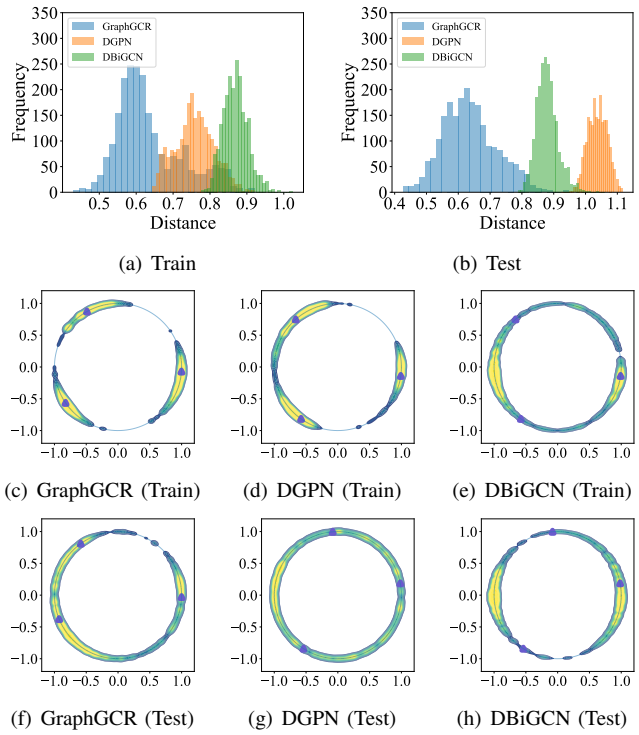
To validate the advantages of alignment and uniformity captured by our GraphGCR, we compare it with the competitive models DGPN and DBiGCN on the C-M10M dataset, as shown in Figure 3. Figures 3(a) and 3(b) illustrate the distribution of cosine distance between node features and corresponding class semantics to demonstrate alignment on the train and test sets, respectively. It is obvious that our method achieves the smallest mean of the distance distribution, indicating the best alignment. Additionally, DGPN exhibits better alignment compared to DBiGCN on the train set, but the opposite is observed on the test set. This is because DGPN models the consistency between node features and class semantics during the training but lacks the connection between seen and unseen classes. On the other hand, DBiGCN effectively captures the relationships among classes, leading to better generalization. For uniformity, we utilize principal component analysis and Gaussian kernel density estimation to visualize the distribution of features on the unit circle in Figures 3(c)-3(h). Taking Figure 3(c) as an example, it is evident that our method exhibits higher uniformity across different locations on the unit circle, with concentrated intra-class features and dispersed inter-class features, which demonstrates the discriminability of the learned contrastive representation.

VI. Conclusion

This work develops a novel framework GraphGCR for graph zero-shot learning, which explores the relationship between

TABLE VI: Accuracy (%) of zero-shot node classification w.r.t. different CSDs.

		Cora			Citeseer			C-M10M		
		TEXT-CSDs	LABEL-CSDs	Decline rate	TEXT-CSDs	LABEL-CSDs	Decline rate	TEXT-CSDs	LABEL-CSDs	Decline rate
Class Split I	DAP	26.56	25.34	-4.59%	34.01	30.01	-11.76%	38.71	32.67	-15.60%
	ESZSL	27.35	25.79	-5.70%	30.32	28.52	-5.94%	37.00	35.02	-5.35%
	ZS-GCN	25.73	23.73	-7.77%	28.62	26.11	-8.77%	37.89	33.32	-12.06%
	WDVSc	30.62	18.73	-38.83%	23.46	19.70	-16.02%	38.12	30.82	-19.15%
	Hyperbolic-ZSL	26.36	25.47	-3.38%	34.18	21.04	-38.44%	35.80	34.49	-3.66%
	DGPN	33.76	32.69	-3.17%	37.74	31.05	-17.73%	41.93	35.12	-16.24%
	DBiGCN	45.08	32.89	-27.04%	38.57	34.18	-11.38%	41.11	37.54	-8.68%
	GraphCEN	48.43	39.63	-18.07%	40.77	38.45	-5.69%	44.17	38.68	-12.43%
	GraphGCR (Ours)	**48.98**	**40.05**	-18.23%	**41.21**	**39.14**	-5.02%	**48.23**	**40.70**	-15.61%



(a) Train (b) Test

(c) GraphGCR (Train) (d) DGPN (Train) (e) DBiGCN (Train)

(f) GraphGCR (Test) (g) DGPN (Test) (h) DBiGCN (Test)

Fig. 3: Case study of alignment and uniformity.

seen and unseen classes, and captures the uniformity and alignment of the learned representations to enhance the model generalization ability. We first extend supervised contrastive learning for seen classes to capture desired properties as well as design a class generator to synthesize new features for unseen classes. The experimental results on real-world datasets demonstrate the superior performance of our method GraphGCR. Despite its advantages, the design of GraphGCR relies on the homophily assumption, which limits its effective applicability to heterogeneous graphs, where edges may connect nodes of different types. In our future, we aim to expand our framework to heterogeneous graphs with generalized graph zero-shot learning and combine advanced pre-training strategies for better knowledge transfer.

VII. Appendix

Proof of Theorem IV.1. Denote the node-class contrastive loss for any node v_i as $\mathcal{L}^S_{i,a} = h(T^{\text{unif}}_{i,a} + T^{\text{align}}_{i,a})$, where $T^{\text{unif}}_{i,a}$, $T^{\text{align}}_{i,a}$, and $h(\cdot)$ are defined in Eq. (3). With the fact that $||\bar{\mathbf{z}}_a - \bar{\mathbf{o}}_j||^2 = 2 - 2\bar{\mathbf{z}}_a^\top \bar{\mathbf{o}}_j \in [0, 4]$, intuitively, to minimize loss $\mathcal{L}^S_{i,a}$, the normalized representations of the affinity nodes of v_i that belongs to class y_i and the normalized class semantics of y_i should have a small L_2-norm to minimize $T^{\text{align}}_{i,a}$, which aligns the two terms on the unit hypersphere.

Let $M = |\mathcal{C}_s \setminus \{y_i\}|$. For uniformity, we perform the following discussion. First, we can derive that

$$
\begin{aligned}
\mathcal{L}^S_{i,a} =& \log\left(1 + \frac{\sum_{j\in\mathcal{C}_s\setminus\{y_i\}} e^{\text{sim}(\mathbf{z}_a,\mathbf{o}_j)/\tau}}{e^{\text{sim}(\mathbf{z}_a,\mathbf{o}_{y_i})/\tau}}\right) \\
\sim& e^{-\text{sim}(\mathbf{z}_a,\mathbf{o}_{y_i})/\tau} \cdot \sum_{j\in\mathcal{C}_s\setminus\{y_i\}} e^{\text{sim}(\mathbf{z}_a,\mathbf{o}_j)/\tau} \\
\sim& e^{||\bar{\mathbf{z}}_a-\bar{\mathbf{o}}_{y_i}||^2/\tau} \cdot \sum_{j\in\mathcal{C}_s\setminus\{y_i\}} e^{-||\bar{\mathbf{z}}_a-\bar{\mathbf{o}}_j||^2/\tau} \\
\le& e^{4/\tau} \cdot \sum_{j\in\mathcal{C}_s\setminus\{y_i\}} e^{-||\bar{\mathbf{z}}_a-\bar{\mathbf{o}}_j||^2/\tau} \\
\sim& \frac{1}{M} \sum_{j\in\mathcal{C}_s\setminus\{y_i\}} e^{-||\bar{\mathbf{z}}_a-\bar{\mathbf{o}}_j||^2/\tau} \\
=& e^{-\frac{1}{M\tau}\sum_{j\in\mathcal{C}_s\setminus\{y_i\}}||\bar{\mathbf{z}}_a-\bar{\mathbf{o}}_j||^2} + \sum_{j\in\mathcal{C}_s\setminus\{y_i\}} \frac{1}{M} e^{-4\lambda_j/\tau} \\
& - e^{-\frac{1}{M\tau}\sum_{j\in\mathcal{C}_s\setminus\{y_i\}} 4\lambda_j} \\
\le& e^{-\frac{1}{M\tau}\sum_{j\in\mathcal{C}_s\setminus\{y_i\}}||\bar{\mathbf{z}}_a-\bar{\mathbf{o}}_j||^2} \\
& + \sum_{j\in\mathcal{C}_s\setminus\{y_i\}} \frac{1}{M}[\lambda_j e^{-4/\tau} + (1-\lambda_j)e^0] \\
& - e^{-\frac{4}{M\tau}\sum_{j\in\mathcal{C}_s\setminus\{y_i\}}\lambda_j} \\
=& e^{-\frac{1}{M\tau}\sum_{j\in\mathcal{C}_s\setminus\{y_i\}}||\bar{\mathbf{z}}_a-\bar{\mathbf{o}}_j||^2} + \frac{e^{-4/\tau}}{M} \sum_{j\in\mathcal{C}_s\setminus\{y_i\}} \lambda_j \\
& + \frac{1}{M} \sum_{j\in\mathcal{C}_s\setminus\{y_i\}} (1-\lambda_j) - e^{-\frac{4}{M\tau}\sum_{j\in\mathcal{C}_s\setminus\{y_i\}}\lambda_j} \\
=& e^{-\frac{1}{M\tau}\sum_{j\in\mathcal{C}_s\setminus\{y_i\}}||\bar{\mathbf{z}}_a-\bar{\mathbf{o}}_j||^2} + pe^{-4/\tau} + q - e^{-4p/\tau} \\
\le& e^{-\frac{1}{M\tau}\sum_{j\in\mathcal{C}_s\setminus\{y_i\}}||\bar{\mathbf{z}}_a-\bar{\mathbf{o}}_j||^2} + pe^{-4/\tau} + q \\
& + 2q\left[\frac{q-p}{2q}e^{-4/\tau} - e^{\frac{q-p}{2q}(-4/\tau)+\frac{1}{2q}(-4p/\tau)}\right] \\
=& e^{-\frac{1}{M\tau}\sum_{j\in\mathcal{C}_s\setminus\{y_i\}}||\bar{\mathbf{z}}_a-\bar{\mathbf{o}}_j||^2} + pe^{-4/\tau} + q + (q-p)e^{-4/\tau} \\
& - 2qe^{-2/\tau} \\
\le& e^{-\frac{1}{M\tau}\sum_{j\in\mathcal{C}_s\setminus\{y_i\}}||\bar{\mathbf{z}}_a-\bar{\mathbf{o}}_j||^2} + e^{-4/\tau} + 1 - 2e^{-2/\tau},
\end{aligned}
$$

where '$\sim$' stands for equivalence, $\lambda_j \in [0, 1]$ is set to satisfy

$-||\bar{\mathbf{z}}_a - \bar{\mathbf{o}}_j||^2/\tau = -4\lambda_j/\tau$, $\sum_{j\in\mathcal{C}_s\setminus\{y_i\}} \lambda_j/M = p$, $p+q=1$, and the Jensen's inequality is used to build the inequalities. In the last line of the above derivation, all but the first term are independent of the optimization variables. It implies that maximizing $\sum_{j\in\mathcal{C}_s\setminus\{y_i\}} ||\bar{\mathbf{z}}_a - \bar{\mathbf{o}}_j||^2$ can effectively guarantee the minimization of the loss $\mathcal{L}^S_{i,a}$.

Taking a similar argument on the loss $\mathcal{L}^S$, with a summation layer over $\mathcal{L}^S_{i,a}$ with respect to all the nodes, *i.e.*, $\mathcal{L}^S = \sum_{i=1}^N \frac{1}{|A(i)|}\sum_{a\in A(i)} \mathcal{L}^S_{i,a}$, we can obtain that maximizing $\sum_{i=1}^N \sum_{a\in A(i)} \sum_{j\in\mathcal{C}_s\setminus\{y_i\}} ||\bar{\mathbf{z}}_a - \bar{\mathbf{o}}_j||^2$ can boost the optimization of the loss $\mathcal{L}_S$. Moreover, we also obviously have

$$\frac{1}{topkN\cdot M}\sum_{i=1}^{N}\sum_{a\in A(i)}\sum_{j\in\mathcal{C}_s\setminus\{y_i\}} ||\bar{\mathbf{z}}_a - \bar{\mathbf{o}}_j||^2 \geq \min_{i\in[1:N],a\in A(i),j\in\mathcal{C}_s\setminus\{y_i\}} ||\bar{\mathbf{z}}_a - \bar{\mathbf{o}}_j||^2,$$

which implies that maximizing $\min_{i,a,j} ||\bar{\mathbf{z}}_a - \bar{\mathbf{o}}_j||^2$ can facilitate the maximization of $\sum_i\sum_a\sum_j ||\bar{\mathbf{z}}_a - \bar{\mathbf{o}}_j||^2$. As for the objective $\max_{\mathbb{R}^{d'}} \min_{i,a,j} ||\bar{\mathbf{z}}_a - \bar{\mathbf{o}}_j||^2$, it geometrically pushes $\bar{\mathbf{z}}_a$'s and $\bar{\mathbf{o}}_j$'s as far apart from each other as possible to ensure that they are spread out over the latent space. It is consistent with the concept of the maximin design, a kind of space-filling design with good uniformity in the domain of experimental designs. Combining with the alignment property, an ideal scenario is that for any $i\in\{1,\ldots,N\}$, the affinity node representations $\{\bar{\mathbf{z}}_a, a\in A(i)\}$ of node v_i that include $\bar{\mathbf{z}}_i$ are pulled to their corresponding class semantic representation $\bar{\mathbf{o}}_{y_i}$, and for any $j\in\mathcal{C}_s$, the minimal L_2-norm between $\bar{\mathbf{o}}_j$ and $\{\bar{\mathbf{o}}_{j'}, j'\in\mathcal{C}_s\setminus\{j\}\}$ is the same, which results in a uniform distribution for normalized class semantic representations on the unit hypersphere.

REFERENCES

[1] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[2] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[3] X. Luo, W. Ju, Y. Gu, Y. Qin, S. Yi, D. Wu, L. Liu, and M. Zhang, "Toward effective semi-supervised node classification with hybrid curriculum pseudo-labeling," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 3, pp. 1–19, 2023.

[4] J. Lu, Z. Wu, L. Zhong, Z. Chen, H. Zhao, and S. Wang, "Generative essential graph convolutional network for multi-view semi-supervised classification," *IEEE Transactions on Multimedia*, 2024.

[5] W. Ju, S. Yi, Y. Wang, Q. Long, J. Luo, Z. Xiao, and M. Zhang, "A survey of data-efficient graph learning," in *International Joint Conference on Artificial Intelligence*, 2024.

[6] J. Luo, Z. Xiao, Y. Wang, X. Luo, J. Yuan, W. Ju, L. Liu, and M. Zhang, "Rank and align: towards effective source-free graph domain adaptation," in *International Joint Conference on Artificial Intelligence*, 2024.

[7] Y. Wang, X. Luo, C. Chen, X.-S. Hua, M. Zhang, and W. Ju, "Disensemi: Semi-supervised graph classification via disentangled representation learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[8] J. Luo, Y. Gu, X. Luo, W. Ju, Z. Xiao, Y. Zhao, J. Yuan, and M. Zhang, "Gala: Graph diffusion-based alignment with jigsaw for source-free domain adaptation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 01, pp. 1–14, 2024.

[9] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, pp. 1–37, 2019.

[10] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *International Conference on Machine Learning*. PMLR, 2015, pp. 2152–2161.

[11] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5542–5551.

[12] Y. Xian, S. Sharma, B. Schiele, and Z. Akata, "f-vaegan-d2: A feature generating framework for any-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 275–10 284.

[13] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2251–2265, 2018.

[14] X. Zhang, S. Gui, Z. Zhu, Y. Zhao, and J. Liu, "Hierarchical prototype learning for zero-shot recognition," *IEEE Transactions on Multimedia*, vol. 22, no. 7, pp. 1692–1703, 2019.

[15] Y. Li, Z. Liu, L. Yao, and X. Chang, "Attribute-modulated generative meta learning for zero-shot learning," *IEEE Transactions on Multimedia*, vol. 25, pp. 1600–1610, 2021.

[16] G.-S. Xie, L. Liu, X. Jin, F. Zhu, Z. Zhang, J. Qin, Y. Yao, and L. Shao, "Attentive region embedding network for zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9384–9393.

[17] G.-S. Xie, L. Liu, F. Zhu, F. Zhao, Z. Zhang, Y. Yao, J. Qin, and L. Shao, "Region graph embedding network for zero-shot learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 2020, pp. 562–580.

[18] S. Chen, Z. Hong, Y. Liu, G.-S. Xie, B. Sun, H. Li, Q. Peng, K. Lu, and X. You, "Transzero: Attribute-guided transformer for zero-shot learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 330–338.

[19] Z. Wang, J. Wang, Y. Guo, and Z. Gong, "Zero-shot node classification with decomposed graph prototype network," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 1769–1779.

[20] Q. Yue, J. Liang, J. Cui, and L. Bai, "Dual bidirectional graph convolutional networks for zero-shot node classification," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 2408–2417.

[21] W. Ju, Y. Qin, S. Yi, Z. Mao, K. Zheng, L. Liu, X. Luo, and M. Zhang, "Zero-shot node classification with graph contrastive embedding network," *Transactions on Machine Learning Research*, 2023.

[22] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, "Transductive multi-view zero-shot learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 11, pp. 2332–2345, 2015.

[23] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 661–18 673, 2020.

[24] J. Yuan, X. Luo, Y. Qin, Y. Zhao, W. Ju, and M. Zhang, "Learning on graphs under label noise," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[25] J. Yuan, X. Luo, Y. Qin, Z. Mao, W. Ju, and M. Zhang, "Alex: Towards effective graph transfer learning with noisy labels," in *Proceedings of the ACM International Conference on Multimedia*, 2023, pp. 3647–3656.

[26] S. Xiao, S. Du, Z. Chen, Y. Zhang, and S. Wang, "Dual fusion-propagation graph neural network for multi-view clustering," *IEEE Transactions on Multimedia*, 2023.

[27] W. Xia, Q. Wang, Q. Gao, M. Yang, and X. Gao, "Self-consistent contrastive attributed graph clustering with pseudo-label prompt," *IEEE Transactions on Multimedia*, 2022.

[28] W. Ju, Y. Gu, B. Chen, G. Sun, Y. Qin, X. Liu, X. Luo, and M. Zhang, "Glcc: A general framework for graph-level clustering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 4, 2023, pp. 4391–4399.

[29] S. Yi, W. Ju, Y. Qin, X. Luo, L. Liu, Y. Zhou, and M. Zhang, "Redundancy-free self-supervised relational learning for graph clustering," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[30] X. Luo, Y. Zhao, Y. Qin, W. Ju, and M. Zhang, "Towards semi-supervised universal graph classification," *IEEE Transactions on Knowledge and Data Engineering*, 2023.

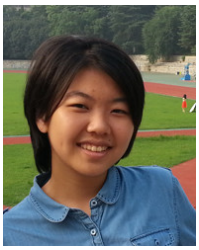
[31] Z. Mao, W. Ju, Y. Qin, X. Luo, and M. Zhang, "Rahnet: Retrieval augmented hybrid network for long-tailed graph classification," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 3817–3826.

[32] W. Ju, Z. Mao, S. Yi, Y. Qin, Y. Gu, Z. Xiao, Y. Wang, X. Luo, and M. Zhang, "Hypergraph-enhanced dual semi-supervised graph classification," *arXiv preprint arXiv:2405.04773*, 2024.
[33] X. Luo, Y. Zhao, Z. Mao, Y. Qin, W. Ju, M. Zhang, and Y. Sun, "Rignn: A rationale perspective for semi-supervised open-world graph classification," *Transactions on Machine Learning Research*, 2023.
[34] Y. Gu, Z. Chen, Y. Qin, Z. Mao, Z. Xiao, W. Ju, C. Chen, X.-S. Hua, Y. Wang, X. Luo *et al.*, "Deer: Distribution divergence-based graph contrast for partial label learning on graphs," *IEEE Transactions on Multimedia*, 2024.
[35] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," *Advances in Neural Information Processing Systems*, vol. 29, 2016.
[36] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1263–1272.
[37] Z. Wan, D. Chen, Y. Li, X. Yan, J. Zhang, Y. Yu, and J. Liao, "Transductive zero-shot learning with visual structure constraint," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
[38] Z. Yang, Y. Liu, W. Xu, C. Huang, L. Zhou, and C. Tong, "Learning prototype via placeholder for zero-shot recognition," *arXiv preprint arXiv:2207.14581*, 2022.
[39] X. Li, Z. Xu, K. Wei, and C. Deng, "Generalized zero-shot learning via disentangled representation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 1966–1974.
[40] L. Wu, J. Jiang, H. Zhao, H. Wang, D. Lian, M. Zhang, and E. Chen, "Kmf: knowledge-aware multi-faceted representation learning for zero-shot node classification," *arXiv preprint arXiv:2308.08563*, 2023.
[41] S. Pu, K. Zhao, and M. Zheng, "Alignment-uniformity aware representation learning for zero-shot video classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 968–19 977.
[42] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
[43] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.
[44] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.
[45] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3733–3742.
[46] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1597–1607.
[47] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.
[48] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 271–21 284, 2020.
[49] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, "Graph contrastive learning with augmentations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5812–5823, 2020.
[50] K. Liu, F. Xue, D. Guo, P. Sun, S. Qian, and R. Hong, "Multimodal graph contrastive learning for multimedia-based recommendation," *IEEE Transactions on Multimedia*, 2023.
[51] W. Ju, Y. Wang, Y. Qin, Z. Mao, Z. Xiao, J. Luo, J. Yang, Y. Gu, D. Wang, Q. Long *et al.*, "Towards graph contrastive learning: A survey and beyond," *arXiv preprint arXiv:2405.11868*, 2024.
[52] W. Ju, Y. Gu, Z. Mao, Z. Qiao, Y. Qin, X. Luo, H. Xiong, and M. Zhang, "Gps: Graph contrastive learning via multi-scale augmented views from adversarial pooling," *arXiv preprint arXiv:2401.16011*, 2024.
[53] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi, "A theoretical analysis of contrastive unsupervised representation learning," *arXiv preprint arXiv:1902.09229*, 2019.
[54] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9929–9939.
[55] F. Wang and H. Liu, "Understanding the behaviour of contrastive loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2495–2504.
[56] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
[57] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 297–304.
[58] J. Klicpera, S. Weißenberger, and S. Günnemann, "Diffusion improves graph learning," *arXiv preprint arXiv:1911.05485*, 2019.
[59] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.
[60] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453–465, 2013.
[61] X. Wang, Y. Ye, and A. Gupta, "Zero-shot recognition via semantic embeddings and knowledge graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6857–6866.
[62] S. Liu, J. Chen, L. Pan, C.-W. Ngo, T.-S. Chua, and Y.-G. Jiang, "Hyperbolic visual embedding learning for zero-shot recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9273–9281.
[63] J. Luo, X. Luo, X. Chen, Z. Xiao, W. Ju, and M. Zhang, "Semievol: Semi-supervised fine-tuning for llm adaptation," *arXiv preprint arXiv:2410.14745*, 2024.
[64] J. Yang, H. Xu, S. Mirzoyan, T. Chen, Z. Liu, Z. Liu, W. Ju, L. Liu, Z. Xiao, M. Zhang *et al.*, "Poisoning medical knowledge using large language models," *Nature Machine Intelligence*, pp. 1–13, 2024.

Siyu Yi is currently a postdoctoral researcher in Mathematics at Sichuan University, Chengdu, China. She received the B.S. and M.S. degrees in Statistics from Sichuan University, Sichuan, China, in 2017 and 2020, respectively. After that, she received the Ph.D. degree in Statistics from Nankai University, Tianjin, China, in 2024. Her research interests focus on graph machine learning, statistical learning, and subsampling in big data. She has published more than 20 papers.

Zhengyang Mao is currently a master's student at the School of Computer Science, Peking University. His research interests include graph representation learning and long-tailed learning.

Kangjie Zheng is currently a Ph.D. candidate in computer science from Peking University, Beijing, China. He received the B.S. degree in Computer Science from Harbin Institute of Technology Harbin, China, in 2020. His research interests include biological and chemical data representation learning and text generation.



Zhiping Xiao has graduated from Ph.D. program in Computer Science at University of California, Los Angeles in 2024. Her major is artificial intelligence, minor is data mining, did research in the area of multi-modality social-media data analysis, and her current research interests lie in AI for pathology. She is also interested in other interdisciplinary applications.

Ziyue Qiao is currently an Assistant Professor at the School of Computing and Information Technology, Great Bay University. Previously, he was a postdoc at The Hong Kong University of Science and Technology (Guangzhou) from 2022 to 2024. He received his Ph.D. degree in 2022 at the University of Chinese Academy of Sciences, and his B.S. degree in 2017 at Wuhan University, China. His research interests include data mining, graph learning, and AI for science, with a focus on graph representation/transfer learning and academic data mining.

Chong Chen is currently a research scientist in Terminus Group. He received the B.S. degree in Mathematics from Peking University in 2013 and the Ph.D. degree in Statistics from Peking University in 2019 under the supervision of Prof. Ruibin Xi. His research interests include image understanding, self-supervised learning, and data mining.

Xian-Sheng Hua (Fellow, IEEE) received the B.S. and Ph.D. degrees in applied mathematics from Peking University, Beijing, in 1996 and 2001, respectively. In 2001, he joined Microsoft Research Asia, as a Researcher, and has been a Senior Researcher at Microsoft Research Redmond since 2013. He became a Researcher and the Senior Director of Alibaba Group in 2015. He has authored or coauthored over 250 research articles and has filed over 90 patents. His research interests include multimedia search, advertising, understanding, and mining, pattern recognition, and machine learning. He was honored as one of the recipients of MIT35. He served as a Program Co-Chair for the IEEE ICME 2013, the ACM Multimedia 2012, and the IEEE ICME 2012, and on the Technical Directions Board for the IEEE Signal Processing Society. He is an ACM Distinguished Scientist.

Yongdao Zhou received his B.S. degree in Mathematics, M.S. and Ph.D. degrees in Statistics from Sichuan University in 2002, 2005, and 2008, respectively. After graduation, he joined Sichuan University and was a professor after 2015. In 2017, he joined Nankai University, where he is presently a professor in Statistics. His research agenda focuses on design of experiments and big data analysis. He published more than 60 papers and 5 monographs. He has won the best paper awards in WCE 2009 and Sci Sin Math in 2023.

Ming Zhang received her B.S., M.S. and Ph.D. degrees in Computer Science from Peking University respectively. She is a full professor at the School of Computer Science, Peking University. Prof. Zhang is a member of Advisory Committee of Ministry of Education in China and the Chair of ACM SIGCSE China. She is one of the fifteen members of ACM/IEEE CC2020 Steering Committee. She has published more than 200 research papers on Text Mining and Machine Learning in the top journals and conferences. She won the best paper of ICML 2014 and best paper nominee of WWW 2016. Prof. Zhang is the leading author of several textbooks on Data Structures and Algorithms in Chinese, and the corresponding course is awarded as the National Elaborate Course, National Boutique Resource Sharing Course, National Fine-designed Online Course, National First-Class Undergraduate Course by MOE China.

Wei Ju is currently an associate professor with the College of Computer Science, Sichuan University, Chengdu, China. Prior to that, he worked as a postdoc research fellow and received his Ph.D. degree in the School of Computer Science from Peking University, Beijing, China, in 2022. He received the B.S. degree in Mathematics from Sichuan University, Sichuan, China, in 2017. His current research interests lie primarily in the area of machine learning on graphs including graph representation learning and graph neural networks, and interdisciplinary applications such as recommender systems, bioinformatics, drug discovery, and spatio-temporal analysis. He has published more than 50 papers in top-tier venues and has won the best paper finalist in IEEE ICDM 2022.