

Structural Entropy Guided Unsupervised Graph Out-Of-Distribution Detection

Yue Hou^{1,2}, He Zhu¹, Ruomei Liu¹, Yingke Su², Jinxiang Xia¹, Junran Wu^{1*}, Ke Xu¹

¹State Key Laboratory of Complex & Critical Software Environment, Beihang University

²Shen Yuan Honors College, Beihang University

{hou_yue, roy_zh, rmliu, suyingke, jinxiangxia, wu_junran, kexu}@buaa.edu.cn

Abstract

With the emerging of huge amount of unlabeled data, unsupervised out-of-distribution (OOD) detection is vital for ensuring the reliability of graph neural networks (GNNs) by identifying OOD samples from in-distribution (ID) ones during testing, where encountering novel or unknown data is inevitable. Existing methods often suffer from compromised performance due to redundant information in graph structures, which impairs their ability to effectively differentiate between ID and OOD data. To address this challenge, we propose SEGO, an unsupervised framework that integrates structural entropy into OOD detection regarding graph classification. Specifically, within the architecture of contrastive learning, SEGO introduces an anchor view in the form of coding tree by minimizing structural entropy. The obtained coding tree effectively removes redundant information from graphs while preserving essential structural information, enabling the capture of distinct graph patterns between ID and OOD samples. Furthermore, we present a multi-grained contrastive learning scheme at local, global, and tree levels using triplet views, where coding trees with essential information serve as the anchor view. Extensive experiments on real-world datasets validate the effectiveness of SEGO, demonstrating superior performance over state-of-the-art baselines in OOD detection. Specifically, our method achieves the best performance on 9 out of 10 dataset pairs, with an average improvement of 3.7% on OOD detection datasets, significantly surpassing the best competitor by 10.8% on the FreeSolv-ToxCast dataset pair.

Code — <https://github.com/name-is-what/SEGO>

Introduction

Out-of-distribution (OOD) detection (Yang et al. 2024; Wu et al. 2024; Bao et al. 2024) is a crucial task in machine learning that aims to identify whether a given data point deviates significantly from the training distribution, especially for models deployed in real-world applications where encountering novel or unknown data is inevitable. In graph-based data, the challenge of OOD detection is heightened due to the complex structure and relationships inherent in graphs. In this context, a specific OOD detection model is

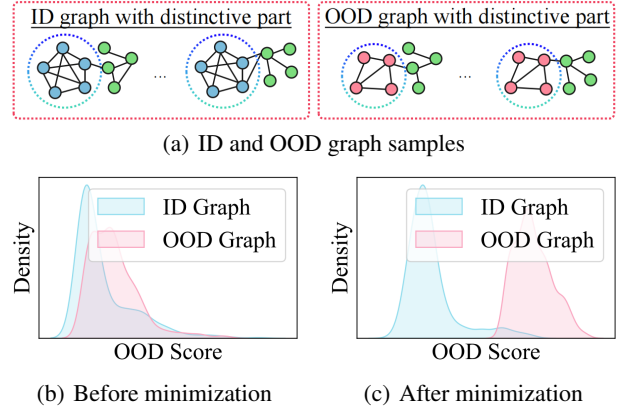


Figure 1: A toy example of ID and OOD graphs and scoring distributions before/after structural entropy minimization.

trained on in-distribution (ID) graphs and then predicts a score for each test sample to indicate its ID/OOD status.

Recent advancements (Wang et al. 2024; Guo et al. 2023; Liu et al. 2023; Yuan et al. 2024b) in graph OOD detection and generation have been explored with growing interest. Several studies (Wang et al. 2024; Guo et al. 2023) employ well-trained graph neural networks (GNNs) (Kipf and Welling 2017; Xu et al. 2019) to fine-tune OOD detectors to identify OOD samples. However, these methods require annotated ID data to pre-train GNNs, which limits their applicability in scenarios where labeled data is unavailable. In contrast, other research (Liu et al. 2023) focuses on training OOD-specific GNN models using only ID data, without relying on any labels or OOD data. They employ unsupervised learning techniques such as graph contrastive learning (GCL) to learn discriminative patterns of unlabeled ID data.

Despite the progress made in this area, a challenge still remains less explored. Due to the prevalent presence of redundant information in graph structures, current methods struggle to effectively capture and distinguish the essential structure between ID and OOD data. Without mechanisms to extract substantive information, models are susceptible to irrelevant features and structures that can mislead the learning process. Besides, GCL methods commonly adopt arbitrary augmentations, which may unexpectedly perturb both structural and semantic patterns of the graph, introducing

*Corresponding authors

undesired OOD samples and converting ID samples into OOD samples. Although methods like GOOD-D (Liu et al. 2023) attempt to mitigate the issue of structural perturbations through perturbation-free data augmentations, they fail to eliminate the interference of irrelevant information.

Structural entropy (Li and Pan 2016) provides a hierarchical abstraction of graphs to measure the complexity of structure. By minimizing structural entropy, the structural uncertainty of the graph is reduced, which aids in capturing essential information and identifying distinct patterns between ID and OOD samples. We argue that the key to OOD detection is eliminating redundant information in graph structure to focus on the most distinctive and effective information. Fig. 1(a) presents a toy example of ID and OOD graph data, where the light blue and pink shaded areas represent the distinctive parts (i.e., essential information) in ID and OOD graph structure, respectively. Capturing these distinctive parts of the graph can better differentiate OOD samples from ID graphs. We also compute the scoring distributions before and after structural entropy minimization on the BZR/COX2 dataset pair (with BZR as ID dataset and COX2 as OOD dataset). As illustrated in Fig. 1(b) and (c), after the minimization, the OOD scores exhibit smaller variance and a decrease in the overlap of scores between OOD and ID samples. The score frequency density plots show that structural entropy minimization effectively removes redundant information in graph samples, preserving the more distinctive parts of the graph, thus enabling the model to detect distributions more effectively.

In this paper, we propose a novel framework, Structural Entropy guided Graph contrastive learning for unsupervised OOD detection, termed **SEGO**, to address this challenge. Our approach introduces structural information theory to graph OOD detection for the first time, which can provide significant insights for future research in this field. By minimizing structural entropy, our method effectively removes redundant information of the graph while capturing essential structure. This allows the model to focus on substantive information that distinguishes ID data from OOD data, improving detection performance. Specifically, we extract a coding tree from the original graph using structural entropy minimization to obtain redundancy-eliminated structural information. Additionally, we theoretically demonstrate that minimizing our contrastive loss preserves the maximum mutual information associated with the ground-truth labels. Based on this foundation, we propose a multi-grained contrastive learning scheme using triplet views: the basic view of the original graph, the coding tree as the anchor view representing essential information, and a topological view enriched with topological features. Maximum agreement is achieved at local, global, and tree levels, encouraging the model to encode shared information between these views. Extensive experiments on both real-world datasets demonstrate the superiority of SEGO against state-of-the-art (SOTA) baselines. Our method shows an average improvement of 3.7% in OOD detection across 10 datasets, highlighting its effectiveness in capturing the essential information of graph data for OOD detection tasks. The main contributions of this work are as follows:

- Guided by structural entropy theory, we propose a novel framework for unsupervised graph OOD detection, termed SEGO, which can remove redundant information and capture the essential structure of graphs, significantly improving the model performance.
- To mitigate the information gap between node and graph embeddings, we employ a multi-grained contrastive learning scheme using triplet views, which includes coding tree as an anchor view and operates at local, global, and tree levels.
- Extensive experiments validate the effectiveness of SEGO, demonstrating superior performance over SOTA baselines in OOD detection.

Related Work

Graph Out-of-distribution Detection. OOD detection aims to identify OOD samples from ID ones and has gained increasing traction due to its wide application for vision (Sehwag, Chiang, and Mittal 2021; Wu et al. 2021; Liang, Li, and Srikant 2018) and language (Zhou, Liu, and Chen 2021) data. OOD detection on graph data can be broadly divided into two categories: graph-level (Liu et al. 2023) and node-level detection (Yang et al. 2024; Wu et al. 2024; Bao et al. 2024). Lots of existing methods (Zhu et al. 2024b; Li et al. 2024a,b) focus on improving the generalization ability of GNNs for specific downstream tasks like node classification through supervised learning, rather than identifying OOD samples. Compared to the works (Liang, Li, and Srikant 2018; Hendrycks and Gimpel 2017) relying on ground-truth labels, relatively less effort has been devoted to unsupervised graph-level OOD detection, which remains an urgent research problem. In this work, we provide a novel perspective for identifying OOD graphs by focusing on distinctive essential information based on structural entropy.

Graph Contrastive Learning. As an effective graph self-supervised learning paradigm (Liu et al. 2021, 2022), GCL has achieved great success on unsupervised graph representation learning (Velickovic et al. 2020; Sun et al. 2020; Hassani and Khasahmadi 2020; You et al. 2020; Zhu et al. 2021; Qiu et al. 2020; Zheng et al. 2022a,b; Ding et al. 2022). Typically, GCL methods involve generating diverse graph views through data augmentation techniques and optimizing the mutual agreement between these views to enhance the representation of samples with similar semantic semantics (You et al. 2020; Zhu et al. 2021; Hassani and Khasahmadi 2020; Zheng et al. 2022b; Ding et al. 2022). However, methods that perform augmentation on graph structures may inadvertently introduce undesired OOD samples within ID data, and views that enhance graph features often suffer from containing redundant information. To effectively capture the essential structure of original graphs, this paper introduces a GCL framework guided by structural entropy, innovatively incorporating triplet views and multi-grained contrast.

Structural Entropy. Structural entropy (Li and Pan 2016), an extension of Shannon entropy (Shannon 1948), quantifies system uncertainty by measuring the complexity of graph structures through the coding tree. Structural entropy has been widely applied in various domains (Wu et al. 2022a,b,

2023; Zhu et al. 2023, 2024a; Hou et al. 2024). In our work, we apply structural entropy in a self-supervised manner to capture the most distinctive part of graphs with essential information for unsupervised graph-level OOD detection.

Notations and Preliminaries

Before formulating the research problem, we first provide some necessary notations. Let $G = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ represent a graph, where \mathcal{V} is the set of nodes and \mathcal{E} is the set of edges. The node features are represented by the feature matrix $\mathbf{X} \in \mathbb{R}^{n \times d_f}$, where $n = |\mathcal{V}|$ is the number of nodes and d_f is the feature dimension. The structure information can also be described by an adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, so a graph can be alternatively represented by $G = (\mathbf{A}, \mathbf{X})$.

Unsupervised Graph-level OOD Detection. We consider an unlabeled ID dataset $\mathcal{D}^{id} = \{G_1^{id}, \dots, G_{N_1}^{id}\}$ where graphs are sampled from distribution \mathbb{P}^{id} and an OOD dataset $\mathcal{D}^{ood} = \{G_1^{ood}, \dots, G_{N_2}^{ood}\}$ where graphs are sampled from a different distribution \mathbb{P}^{ood} . Given a graph G from \mathcal{D}^{id} or \mathcal{D}^{ood} , our objective is to detect whether G originates from \mathbb{P}^{id} or \mathbb{P}^{ood} . Specifically, we aim to learn a model $f(\cdot)$ that assigns an OOD detection score $s = f(G)$ for an input graph G , with a higher s indicating a greater probability that G is from \mathbb{P}^{ood} . The model f is trained solely on the ID dataset $\mathcal{D}_{train}^{id} \subset \mathcal{D}^{id}$ and evaluated on a test set $\mathcal{D}_{test}^{id} \cup \mathcal{D}_{test}^{ood}$ (note that $\mathcal{D}_{test}^{id} \cap \mathcal{D}_{train}^{id} = \emptyset$, $\mathcal{D}_{test}^{id} \subset \mathcal{D}^{id}$, and $\mathcal{D}_{test}^{ood} \subset \mathcal{D}^{ood}$).

Structural Entropy. Structural entropy is initially proposed (Li and Pan 2016) to measure the uncertainty of graph structure, revealing the essential structure of a graph. The structural entropy of a given graph $G = \{\mathcal{V}, \mathcal{E}, \mathbf{X}\}$ on its coding tree T is defined as:

$$\mathcal{H}^T(G) = - \sum_{v_\tau \in T} \frac{g_{v_\tau}}{\text{vol}(\mathcal{V})} \log \frac{\text{vol}(v_\tau)}{\text{vol}(v_\tau^+)}, \quad (1)$$

where v_τ is a node in T except for root node and also stands for a subset $\mathcal{V}_\tau \in \mathcal{V}$, g_{v_τ} is the number of edges connecting nodes in and outside \mathcal{V}_τ , v_τ^+ is the immediate predecessor of v_τ and $\text{vol}(v_\tau)$, $\text{vol}(v_\tau^+)$ and $\text{vol}(\mathcal{V})$ are the sum of degrees of nodes in v_τ , v_τ^+ and \mathcal{V} , respectively.

Graph Contrastive Learning. In the general graph contrastive learning paradigm for graph classification, two augmented graphs are generated using different graph augmentation operators. Subsequently, representations are generated using a GNN encoder, and further mapped into an embedding space by a shared projection head for contrastive learning. A typical graph contrastive loss, InfoNCE (Chen et al. 2020; Zhu et al. 2021), treats the same graph G_i in different views G_i^α and G_i^β as positive pairs and other nodes as negative pairs. The graph contrastive learning loss \mathcal{L}_i of graph G_i and total loss \mathcal{L} can be formulated as:

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

$$\mathcal{L} = \frac{1}{2N} \sum_{i=1}^N [\ell(\mathbf{z}_i^\alpha, \mathbf{z}_i^\beta) + \ell(\mathbf{z}_i^\beta, \mathbf{z}_i^\alpha)], \quad (3)$$

where N denotes the batch size, τ is the temperature coefficient, and $\text{sim}(\cdot, \cdot)$ stands for cosine similarity function.

Methodology

In this section, we introduce the framework (see Fig. 2), termed **SEGO**. We first theoretically reveal that the coding tree with minimum structural entropy effectively captures the essential information of the graph. Additionally, we demonstrate that minimizing our contrastive loss preserves the maximum mutual information associated with ground-truth labels. Based on these insights, we propose a multi-grained contrastive learning using triplet views.

Essential View with Redundancy-eliminated Information

Redundancy-eliminated Essential Information. The key to effectively distinguishing between ID and OOD graphs lies in maximizing the elimination of redundancy while preserving essential information. According to graph information bottleneck theory (GIB) (Wu et al. 2020), retaining important information in a graph view should involve maximizing mutual information (MI) between the output and labels (i.e., $\max I(f(G); y)$) while reducing mutual information between input and output (i.e., $\min I(G; f(G))$). For unsupervised downstream tasks where ground-truth labels are unavailable, the objective of minimizing $I(G; f(G))$ is to generate an essential view that retains sufficient information while reducing uncertainty (i.e., redundant information) as much as possible. This can be expressed as follows:

$$\text{GIB: } \max I(f(G); y) - \beta I(G; f(G)) \Rightarrow \min I(G; f(G)), \quad (4)$$

where $I(\cdot; \cdot)$ denotes the MI between inputs.

Definition 1. *The anchor view with redundancy-eliminated essential information is supposed to be a distinctive substructure of the given graph.*

Now, let G^* be the target anchor view of graph G , the mutual information between G and G^* can be formulated as:

$$I(G^*; G) = \mathcal{H}(G^*) - \mathcal{H}(G^*|G), \quad (5)$$

where $\mathcal{H}(G^*)$ is the structural entropy of G^* and $\mathcal{H}(G^*|G)$ is the conditional entropy of G^* conditioned on G .

Theorem 1. *The information in G^* is a subset of information in G (i.e., $\mathcal{H}(G^*) \subseteq \mathcal{H}(G)$); thus, we have:*

$$\mathcal{H}(G^*|G) = 0. \quad (6)$$

Here, the mutual information between G and G^* can be rewritten as:

$$I(G^*; G) = \mathcal{H}(G^*). \quad (7)$$

Accordingly, to acquire the anchor view with essential information, we need to optimize:

$$\min I(G; f(G)) \Rightarrow \min \mathcal{H}(G^*). \quad (8)$$

Thus, we argue that the view obtained by minimizing structural entropy of a given graph represents the redundancy-eliminated information, serving as an anchor view that retains the graph's distinctive substructure.

matrix formed by concatenating random walk diffusion and Laplacian positional encoding, formally written as $\mathbf{p}_i = [\mathbf{p}_i^{(rw)} || \mathbf{p}_i^{(lp)}]$. Specifically, the random walk diffusion encoding is given by $\mathbf{p}_i^{(rw)} = [RW_{ii}, RW_{ii}^2, \dots, RW_{ii}^r] \in \mathbb{R}^r$, where $RW = \mathbf{A}\mathbf{D}^{-1}$ is the random walk transition matrix, and \mathbf{D} is the diagonal degree matrix. The Laplacian positional encoding is defined as $\mathbf{p}_i^{(lp)} = \Delta_{ii}$, where $\Delta = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$, with \mathbf{I} being the identity matrix. This strategy ensures discriminative representations for OOD detection.

These triplet views in SEGO, namely the basic view G_b , tree anchor view T , and topo. view G_t , collectively integrate multiple levels of information within the graph. Since the coding tree serves as an abstraction of the essential structure on entire graph, maximizing the MI between tree anchor view T and basic view G_b (i.e., $I(T; G_b)$) allows the model to focus on capturing global redundancy-eliminated patterns. However, relying solely on T and G_b might result in the overlooking of fine-grained node-level representations, as $I(T; G_b)$ primarily emphasizes coarse-grained graph-level information. Therefore, MI between basic view G_b and topo. view G_t (i.e., $I(G_b; G_t)$) is also required, which is captured in both node and graph embeddings. The introduction of G_t aligns the information from individual nodes with the overall graph structure, addressing the instability of information. The triplet views in our method mitigate the information gap between node and graph embeddings, effectively capturing both coarse- and fine-grained redundancy-eliminated essential information.

Triplet Views Representing Learning. To effectively extract embeddings from the basic view G_b and topo view G_t , we utilize two parallel and independent GNN encoders (i.e., encoder f_b for the basic view and f_t for the topo view) for representation learning, following the approach in GOOD-D (Liu et al. 2023). We employ GIN (Xu et al. 2019) as the backbone for its powerful expression ability. Taking f_b as an example, the propagation in the l -th layer can be expressed as, $\mathbf{h}_i^{(b,l)} = \text{MLP}^{(b,l)} \left(\mathbf{h}_i^{(b,l-1)} + \sum_{v_j \in \mathcal{N}(v_i)} \mathbf{h}_j^{(b,l-1)} \right)$, where MLP is a multilayer perceptron network with 2 layers, $\mathbf{h}_i^{(b,l)}$ is the interval embedding of node v_i at the l -th layer of encoder f_b , $\mathcal{N}(v_i)$ is the set of first-order neighbors of node v_i . After getting node embeddings, we employ a readout function to acquire the graph embedding: $\mathbf{h}_G^{(b)} = \sum_{v_i \in \mathcal{V}_G} \mathbf{h}_i^{(b)}$, where \mathcal{V}_G is the node set of G .

For the tree anchor view T , the encoder is designed to iteratively transfer messages from the bottom to the top. Formally, the l -th layer of the encoder can be written as, $\mathbf{x}_v^{(l)} = \text{MLP}^{(l)} \left(\sum_{u \in \mathcal{C}(v)} \mathbf{x}_u^{(l-1)} \right)$, where \mathbf{x}_v^i is the feature of v in the i -th layer of coding tree T , \mathbf{x}_v^0 is the input feature of leaf nodes, and $\mathcal{C}(v)$ refers to the children of v . Once the features reach the root node, a readout function is applied to obtain the tree-level embedding \mathbf{z}_T .

Multi-grained Contrastive Learning Objectives. To capture the multi-grained mutual information between views, we employ a multi-grained contrastive learning scheme that extracts features at three distinct levels: the

local-level for fine-grained feature extraction, the **global-level** for coarse-grained feature extraction, and the **tree-level** for capturing essential information of the entire graph. To maximize the agreement between node embeddings from different views of the same graph, we first map $\mathbf{h}_i^{(b)}$ and $\mathbf{h}_i^{(t)}$ into node-space embeddings $\mathbf{z}_i^{(b)}$ and $\mathbf{z}_i^{(t)}$ using MLP-based projection networks. The local-level contrast focuses on both inter-view and intra-view node relationships, defined as follows:

$$\mathcal{L}_{local} = \frac{1}{|\mathcal{B}|} \sum_{G_j \in \mathcal{B}} \frac{1}{2|\mathcal{V}_j|} \sum_{v_i \in \mathcal{V}_j} \left[\ell(\mathbf{z}_i^{(b)}, \mathbf{z}_i^{(t)}) + \ell(\mathbf{z}_i^{(t)}, \mathbf{z}_i^{(b)}) \right], \quad (11)$$

where \mathcal{B} is a training batch containing multiple graph samples, \mathcal{V}_j is the node set of graph G_j , $\ell(\mathbf{z}_i^{(t)}, \mathbf{z}_i^{(b)})$ and $\ell(\mathbf{z}_i^{(b)}, \mathbf{z}_i^{(t)})$ are calculated following Eq.2.

The global-level contrast allows the model to identify coarse-grained patterns that might be overlooked when focusing solely on finer details:

$$\mathcal{L}_{global} = \frac{1}{2|\mathcal{B}|} \sum_{G_i \in \mathcal{B}} \left[\ell(\mathbf{z}_{G_i}^{(b)}, \mathbf{z}_{G_i}^{(t)}) + \ell(\mathbf{z}_{G_i}^{(t)}, \mathbf{z}_{G_i}^{(b)}) \right], \quad (12)$$

where $\mathbf{z}_G^{(b)}$ and $\mathbf{z}_G^{(t)}$ are transformed from $\mathbf{h}_G^{(b)}$ and $\mathbf{h}_G^{(t)}$ using MLP-based projection networks.

Tree-level contrast operates at a higher level of abstraction, which can be calculated by:

$$\mathcal{L}_{tree} = \frac{1}{2|\mathcal{B}|} \sum_{G_i \in \mathcal{B}} \left[\ell(\mathbf{z}_{G_i}^{(b)}, \mathbf{z}_{T_i}) + \ell(\mathbf{z}_{T_i}, \mathbf{z}_{G_i}^{(b)}) \right]. \quad (13)$$

During the training phase, we introduce the standard deviation of prediction errors to adaptively adjust the balance of local and global information. This strategy automatically allocates the weights for loss and score terms. Concretely, the overall loss is calculated by:

$$\mathcal{L} = \mathcal{L}_{tree} + \sigma_l^\theta \mathcal{L}_{local} + \sigma_g^\theta \mathcal{L}_{global}, \quad (14)$$

where σ_l and σ_g are the standard deviations of predicted errors of the node and graph levels, respectively, and $\theta \geq 0$ is a hyper-parameter that controls the strength of self-adaptiveness, penalizing the term with a larger deviation. During the inference phase, to balance the scores of different levels, we employ z-score normalization based on the mean values and standard deviations of the predicted errors of training samples: $s_{G_i} = \frac{s_l - \mu_l}{\sigma_l} + \frac{s_g - \mu_g}{\sigma_g}$, where μ_l and μ_g represent the mean values of the predicted errors at the corresponding levels for the training samples.

Experiment

In this section, we empirically evaluate the effectiveness of the proposed SEGO. In particular, the experiments are unfolded by answering the following research questions:

- **RQ1:** How effective is SEGO compared with competitive baselines on identifying OOD graphs?
- **RQ2:** How transferable is SEGO to anomaly detection?
- **RQ3:** How do our multi-grained contrastive losses affect SEGO’s performance?
- **RQ4:** How about the parameter sensitivity of SEGO?

| ID dataset | BZR | PTC-MR | AIDS | ENZYMES | IMDB-M | Tox21 | FreeSolv | BBBP | ClinTox | Esol | A.A. | A.R. |
|------------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|--------------|------------|
| OOD dataset | COX2 | MUTAG | DHFR | PROTEIN | IMDB-B | SIDER | ToxCast | BACE | LIPO | MUV | | |
| PK-LOF | 42.22±8.39 | 51.04±6.04 | 50.15±3.29 | 50.47±2.87 | 48.03±2.53 | 51.33±1.81 | 49.16±3.70 | 53.10±2.07 | 50.00±2.17 | 50.82±1.48 | 49.63 | 12.9 |
| PK-OCSVM | 42.55±8.26 | 49.71±6.58 | 50.17±3.30 | 50.46±2.78 | 48.07±2.41 | 51.33±1.81 | 48.82±3.29 | 53.05±2.10 | 50.06±2.19 | 51.00±1.33 | 49.52 | 12.8 |
| PK-iF | 51.46±1.62 | 54.29±4.33 | 51.10±1.43 | 51.67±2.69 | 50.67±2.47 | 49.87±0.82 | 52.28±1.87 | 51.47±1.33 | 50.81±1.10 | 50.85±3.51 | 51.45 | 11.1 |
| WL-LOF | 48.99±6.20 | 53.31±8.98 | 50.77±2.87 | 52.66±2.47 | 52.28±4.50 | 51.92±1.58 | 51.47±4.23 | 52.80±1.91 | 51.29±3.40 | 51.26±1.31 | 51.68 | 10.4 |
| WL-OCSVM | 49.16±4.51 | 53.31±7.57 | 50.98±2.71 | 51.77±2.21 | 51.38±2.39 | 51.08±1.46 | 50.38±3.81 | 52.85±2.00 | 50.77±3.69 | 50.97±1.65 | 51.27 | 11.1 |
| WL-iF | 50.24±2.49 | 51.43±2.02 | 50.10±0.44 | 51.17±2.01 | 51.07±2.25 | 50.25±0.96 | 52.60±2.38 | 50.78±0.75 | 50.41±2.17 | 50.61±1.96 | 50.87 | 12.4 |
| OCGIN | 76.66±4.17 | 80.38±6.84 | 86.01±6.59 | 57.65±2.96 | 67.93±3.86 | 46.09±1.66 | 59.60±4.78 | 61.21±8.12 | 49.13±4.13 | 54.04±5.50 | 63.87 | 7.9 |
| GLocalKD | 75.75±5.99 | 70.63±3.54 | 93.67±1.24 | 57.18±2.03 | 78.25±4.35 | 66.28±0.98 | 64.82±3.31 | 73.15±1.26 | 55.71±3.81 | 86.83±2.35 | 72.23 | 5.1 |
| InfoGraph-iF | 63.17±9.74 | 51.43±5.19 | 93.10±1.35 | 60.00±1.83 | 58.73±1.96 | 56.28±0.81 | 56.92±1.69 | 53.68±2.90 | 48.51±1.87 | 54.16±5.14 | 59.60 | 8.5 |
| InfoGraph-MD | 86.14±6.77 | 50.79±8.49 | 69.02±11.67 | 55.25±3.51 | 81.38±1.14 | 59.97±2.06 | 58.05±5.46 | 70.49±4.63 | 48.12±5.72 | 77.57±1.69 | 65.68 | 7.4 |
| GraphCL-iF | 60.00±3.81 | 50.86±4.30 | 92.90±1.21 | 61.33±2.27 | 59.67±1.65 | 56.81±0.97 | 55.55±2.71 | 59.41±3.58 | 47.84±0.92 | 62.12±4.01 | 60.65 | 8.7 |
| GraphCL-MD | 83.64±6.00 | 73.03±2.38 | 93.75±2.13 | 52.87±6.11 | 79.09±2.73 | 58.30±1.52 | 60.31±5.24 | 75.72±1.54 | 51.58±3.64 | 78.73±1.40 | 70.70 | 5.3 |
| GOOD-D _{simp} | 93.00±3.20 | 78.43±2.67 | 98.91±0.41 | <u>61.89±2.51</u> | 79.71±1.19 | 65.30±1.27 | 70.48±2.75 | 81.56±1.97 | 66.13±2.98 | 91.39±0.46 | 78.68 | 3.2 |
| GOOD-D | 94.99±2.25 | <u>81.21±2.65</u> | 99.07±0.40 | 61.84±1.94 | 79.94±1.09 | <u>66.50±1.35</u> | <u>80.13±3.43</u> | <u>82.91±2.58</u> | <u>69.18±3.61</u> | <u>91.52±0.70</u> | <u>80.73</u> | <u>2.2</u> |
| SEGO | 96.66±0.91 | 85.02±0.94 | 99.48±0.11 | 64.42±4.95 | <u>80.27±0.92</u> | 66.67±0.82 | 90.95±1.93 | 87.55±0.13 | 78.99±2.81 | 94.59±0.94 | 84.46 | 1.1 |

Table 1: OOD detection results in terms of AUC (% , mean \pm std). The best and runner-up results are highlighted with **bold** and underline, respectively. A.A. is short for average AUC. A.R. implies the abbreviation of average rank. The results of baselines are derived from the published works.

Experimental Setups

Datasets. For OOD detection, we employ 10 pairs of datasets from two mainstream graph data benchmarks (i.e., TUDataset (Morris et al. 2020) and OGB (Hu et al. 2020)) following GOOD-D (Liu et al. 2023). We also conduct experiments on anomaly detection settings, where 5 datasets from TUDataset (Morris et al. 2020) are used for evaluation, where the samples in minority class or real anomalous class are viewed as anomalies, while the rest are as normal data.

Baselines. We compare SEGO with 14 competing baseline methods, including 6 GCL (Sun et al. 2020; You et al. 2020; Liu et al. 2023) based methods, 6 graph kernel based methods (Vishwanathan et al. 2010; Shervashidze et al. 2011), and 2 end-to-end graph anomaly detection methods (Zhao and Akoglu 2021; Ma et al. 2022).

Evaluation and Implementation. We evaluate SEGO with a popular OOD detection metric, i.e., area under receiver operating characteristic Curve (AUC). Higher AUC values indicate better performance. The reported results are the mean performance with standard deviation after 5 runs.

Performance on OOD Detection (RQ1)

To answer RQ1, we compare our proposed methods with 14 competing methods in OOD detection tasks. The AUC results are reported in Table 1. From the comparison results, we observe that SEGO achieves superior performance improvements over the baselines. Specifically, SEGO achieves the best performance on 9 out of 10 dataset pairs and ranks first on average among all baselines with an average rank (A.R.) of 1.1. Additionally, SEGO outperforms all the compared methods in terms of average AUC with a score of 84.46, which is 3.7% higher than the second-best method GOOD-D (Liu et al. 2023). Notably, on the FreeSolv/ToxCast dataset pair, SEGO surpasses the best competitor by 10.8%. Although SEGO nearly achieves optimal results on the IMDB-M/IMDB-B datasets, it falls short likely because its coding tree only approximates and doesn’t fully remove redundant information. Additionally, the high connectivity and edge density of social network datasets introduce more redundancy, making it harder to capture essential information. These results underscore the superiority of SEGO in

| Dataset | ENZYMES | DHFR | BZR | NCI1 | IMDB-B |
|------------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| PK-OCSVM | 53.67±2.66 | 47.91±3.76 | 46.85±5.31 | 49.90±1.18 | 50.75±3.10 |
| PK-iF | 51.30±2.01 | 52.11±3.96 | 55.32±6.18 | 50.58±1.38 | 50.80±3.17 |
| WL-OCSVM | 55.24±2.66 | 50.24±3.13 | 50.56±5.87 | 50.63±1.22 | 54.08±5.19 |
| WL-iF | 51.60±3.81 | 50.29±2.77 | 52.46±3.30 | 50.74±1.70 | 50.20±0.40 |
| GraphCL-iF | 53.60±4.88 | 51.10±2.35 | 60.24±5.37 | 49.88±0.53 | 56.50±4.90 |
| OCGIN | 58.75±5.98 | 49.23±3.05 | 65.91±1.47 | 71.98±1.21 | 60.19±8.90 |
| GLocalKD | 61.39±8.81 | 56.71±3.57 | 69.42±7.78 | 68.48±2.39 | 52.09±3.41 |
| GOOD-D _{simp} | 61.23±4.58 | <u>62.71±3.38</u> | 74.48±4.91 | 59.56±1.62 | 65.49±1.06 |
| GOOD-D | <u>63.90±3.69</u> | 62.67±3.11 | <u>75.16±5.15</u> | 61.12±2.21 | <u>65.88±0.75</u> |
| SEGO | 76.62±7.35 | 65.31±2.98 | 89.21±4.51 | <u>70.34±1.31</u> | 66.48±0.38 |

Table 2: Anomaly detection results in terms of AUC (% , mean \pm std). The best and runner-up results are highlighted with **bold** and underline, respectively.

OOD detection tasks, demonstrating its ability to capture essential information across different granular levels.

Performance on Anomaly Detection (RQ2)

To investigate if SEGO can generalize to the anomaly detection setting (Zhao and Akoglu 2021; Ma et al. 2022), we conduct experiments on 5 datasets following the benchmark in GLocalKD and GOOD-D (Ma et al. 2022; Liu et al. 2023), where only normal data are used for model training. The results are shown in Table 2. From the results, we find that SEGO shows significant performance improvements compared to other baseline methods. Capturing common patterns in the anomaly detection setting is crucial, which is directly reflected in the performance. Thus, we can conclude that SEGO indeed has a strong capability to learn the essential information of normal graph data.

Ablation Study (RQ3)

Ablation of Multi-grained Contrastive Loss. To address RQ3, we conducted ablation experiments on the OOD detection task by separately removing the different levels of contrastive losses, namely node-, graph-, and tree-level losses. The results are presented in Table 3. Firstly, we observe that applying contrastive loss across all three levels (the last row) achieves the best results on 7 out of 10 datasets and shows promising performance on the remaining datasets. This further elucidates that SEGO better captures the essential information, leading to superior performance in most OOD detection scenarios. Notably, we notice

| \mathcal{L}_{tree} | \mathcal{L}_{global} | \mathcal{L}_{local} | BZR | PTC-MR | AIDS | ENZYMES | IMDB-M | Tox21 | FreeSolv | BBBP | ClinTox | Esol |
|----------------------|------------------------|-----------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| | | | COX2 | MUTAG | DHFR | PROTEIN | IMDB-B | SIDER | ToxCast | BACE | LIPO | MUV |
| ✓ | - | - | 54.79±4.08 | 58.20±3.87 | 43.68±7.36 | 49.26±1.11 | 49.56±5.76 | 49.26±5.10 | 49.89±2.95 | 50.53±0.63 | 51.97±4.58 | 54.49±3.57 |
| - | ✓ | - | 87.44±4.66 | 77.84±3.71 | 97.60±1.05 | 56.74±1.96 | 75.22±1.91 | 65.07±1.32 | 78.40±6.44 | 77.66±2.29 | 70.11±2.44 | 89.57±2.80 |
| - | - | ✓ | 83.51±4.14 | 72.48±3.77 | 96.84±0.58 | 60.85±2.95 | 79.34±1.81 | 62.58±0.67 | 59.48±2.20 | 69.53±2.29 | 53.29±4.32 | 86.49±1.20 |
| ✓ | ✓ | - | 87.27±8.21 | 87.71±1.35 | 97.97±0.04 | 54.82±2.74 | 74.51±1.52 | 64.84±0.29 | <u>89.34±0.06</u> | 88.34±1.64 | 79.21±4.55 | <u>94.13±1.32</u> |
| ✓ | - | ✓ | 79.36±8.69 | 55.08±1.29 | 90.66±3.40 | <u>63.38±4.18</u> | 72.96±3.73 | 55.68±2.67 | 61.01±5.29 | 70.13±0.26 | 52.14±2.58 | 77.78±1.01 |
| - | ✓ | ✓ | 86.29±1.09 | 77.53±4.03 | <u>98.23±0.19</u> | 61.55±1.47 | 75.27±0.54 | <u>65.44±1.14</u> | 88.04±1.15 | 80.43±2.58 | 65.89±4.58 | 90.94±1.17 |
| ✓ | ✓ | ✓ | 96.66±0.91 | <u>85.02±0.94</u> | 99.48±0.11 | 64.42±4.95 | 80.27±0.92 | 66.67±0.82 | 90.95±1.93 | <u>87.55±0.13</u> | <u>78.99±2.81</u> | 94.59±0.94 |

Table 3: Ablation study results of SEGO and its variants in terms of AUC (%), mean \pm std.

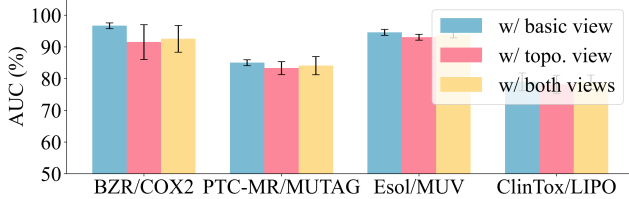


Figure 3: The effectiveness of different views.

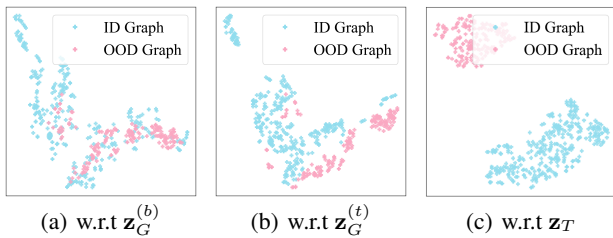


Figure 4: T-SNE visualization of embeddings.

that removing the local-level contrast \mathcal{L}_{local} improves performance on certain dataset pairs (e.g., PTC-MR/MUTAG, BBBP/BACE, and ClinTox/LIPO). These datasets primarily consist of biomolecular data, where molecular activity and interactions are more dependent on the overall topological structure rather than individual node features. Removing \mathcal{L}_{local} , which focuses on node-level feature extraction and optimization, reduces interference and allows the model to focus more on the global graph structure, enhancing performance on these datasets.

Effectiveness of MI in Triplet Views. SEGO utilizes $I(T; G_b)$ at the tree-level contrastive loss. Here, we explore the effectiveness of MI between the anchor view T and topo. view ($I(T; G_t)$), as well as both views ($I(T; G_b) \cup I(T; G_t)$) in identifying OOD graphs. As shown in Fig. 3, we observe that using $I(T; G_b)$ is more effective in eliminating redundancy from the original graph, whereas, in the view G_t , data augmentation reintroduces redundant information, leading to sub-optimal performance.

Visualization. We also visualize the embeddings on PTC-MR/MUTAG dataset pair learned by SEGO in triplet views via t-SNE (Van der Maaten and Hinton 2008). As shown in Fig. 4(a)-(c), the embeddings of ID and OOD graphs are well-separated across these views. Among them, the representation gap in z_T is the most pronounced, highlighting the effectiveness of coding tree in extracting essential structures.

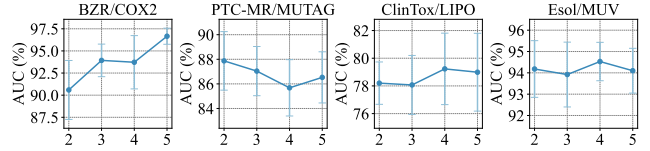


Figure 5: The natural hierarchy of graph.

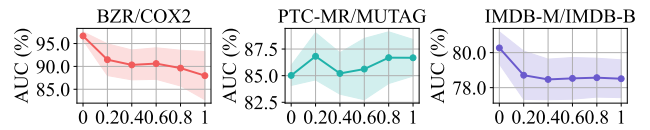


Figure 6: The sensitivity of self-adaptiveness strength θ .

Parameter Study (RQ4)

The Height k of Graph’s Natural Hierarchy. In the experimental setup, the height k of the coding tree is consistently set to 5, in alignment with the GNN encoder. Here, we delve deeper into the effect of the height k on the graph’s natural hierarchy. The specific performance of SEGO under each height k , ranging from 2 to 5, on OOD detection is shown in Fig. 5. We can observe that the optimal height k with the highest accuracy varies among datasets.

Self-adaptiveness Strength θ . To analyze the sensitivity of θ for SEGO, we alter the value of θ from 0 to 1. The AUC w.r.t different selections of θ is plotted in Fig. 6. We can observe that the variation in AUC with changes in θ is not entirely consistent across different datasets. This aligns with the findings from the ablation study in Section , where we noted that the essential information carried by different datasets varies in their dependence on local node information versus global graph information.

Conclusion

In this paper, we make the first attempt to introduce structural information theory into unsupervised OOD detection regarding graph classification. For this task, we propose a novel structural entropy guided graph contrastive learning framework, termed SEGO, that minimizes structural entropy to capture essential graph information while removing redundant information. Our SEGO employs a multi-grained contrastive learning at node, graph, and tree levels with triplet views, including a coding tree with minimum structural entropy as the anchor view. Extensive experiments on real-world datasets validate the effectiveness of SEGO, demonstrating superior performance over state-of-the-art baselines.

Acknowledgments

This work has been supported by the Guangxi Science and Technology Major Project, China (No. AA22067070), NSFC (Grant No. 61932002) and CCSE project (CCSE-2024ZX-09).

References

- Bachman, P.; Hjelm, R. D.; and Buchwalter, W. 2019. Learning Representations by Maximizing Mutual Information Across Views. In *Advances in Neural Information Processing Systems* 32, 15509–15519.
- Bao, T.; Wu, Q.; Jiang, Z.; Chen, Y.; Sun, J.; and Yan, J. 2024. Graph Out-of-Distribution Detection Goes Neighborhood Shaping. In *Forty-first International Conference on Machine Learning*.
- Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607. PMLR.
- Ding, K.; Wang, Y.; Yang, Y.; and Liu, H. 2022. Eliciting Structural and Semantic Global Knowledge in Unsupervised Graph Contrastive Learning. *arXiv preprint arXiv:2202.08480*.
- Guo, Y.; Yang, C.; Chen, Y.; Liu, J.; Shi, C.; and Du, J. 2023. A Data-centric Framework to Endow Graph Neural Networks with Out-Of-Distribution Detection Ability. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 638–648.
- Hassani, K.; and Khasahmadi, A. H. 2020. Contrastive multi-view representation learning on graphs. In *ICML*, 4116–4126. PMLR.
- Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *ICLR*.
- Hou, Y.; Chen, X.; Zhu, H.; Liu, R.; Shi, B.; Liu, J.; Wu, J.; and Xu, K. 2024. NC2D: Novel Class Discovery for Node Classification. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 849–859.
- Hu, W.; Fey, M.; Zitnik, M.; Dong, Y.; Ren, H.; Liu, B.; Catasta, M.; and Leskovec, J. 2020. Open graph benchmark: Datasets for machine learning on graphs. In *NeurIPS*, volume 33, 22118–22133.
- Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.
- Li, A.; and Pan, Y. 2016. Structural Information and Dynamical Complexity of Networks. *IEEE Transactions on Information Theory*, 62: 3290–3339.
- Li, H.; Wang, X.; Zhang, Z.; Chen, H.; Zhang, Z.; and Zhu, W. 2024a. Disentangled Graph Self-supervised Learning for Out-of-Distribution Generalization. In *Forty-first International Conference on Machine Learning*.
- Li, X.; Gui, S.; Luo, Y.; and Ji, S. 2024b. Graph Structure Extrapolation for Out-of-Distribution Generalization. In *Forty-first International Conference on Machine Learning*.
- Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *ICLR*.
- Liu, X.; Zhang, F.; Hou, Z.; Mian, L.; Wang, Z.; Zhang, J.; and Tang, J. 2021. Self-supervised learning: Generative or contrastive. *TKDE*.
- Liu, Y.; Ding, K.; Liu, H.; and Pan, S. 2023. Good-d: On unsupervised graph out-of-distribution detection. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 339–347.
- Liu, Y.; Jin, M.; Pan, S.; Zhou, C.; Zheng, Y.; Xia, F.; and Yu, P. 2022. Graph self-supervised learning: A survey. *TKDE*.
- Ma, R.; Pang, G.; Chen, L.; and van den Hengel, A. 2022. Deep Graph-level Anomaly Detection by Glocal Knowledge Distillation. In *WSDM*.
- Morris, C.; Kriege, N. M.; Bause, F.; Kersting, K.; Mutzel, P.; and Neumann, M. 2020. TUDataset: A collection of benchmark datasets for learning with graphs. In *ICML Workshop*.
- Poole, B.; Ozair, S.; van den Oord, A.; Alemi, A. A.; and Tucker, G. 2019. On Variational Bounds of Mutual Information. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 5171–5180. PMLR.
- Qiu, J.; Chen, Q.; Dong, Y.; Zhang, J.; Yang, H.; Ding, M.; Wang, K.; and Tang, J. 2020. GCC: Graph contrastive coding for graph neural network pre-training. In *SIGKDD*, 1150–1160.
- Sehwag, V.; Chiang, M.; and Mittal, P. 2021. SSD: A Unified Framework for Self-Supervised Outlier Detection. In *ICLR*.
- Shannon, C. E. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3): 379–423.
- Shervashidze, N.; Schweitzer, P.; Van Leeuwen, E. J.; Mehlhorn, K.; and Borgwardt, K. M. 2011. Weisfeiler-lehman graph kernels. *JMLR*, 12(9).
- Sun, F.-Y.; Hoffman, J.; Verma, V.; and Tang, J. 2020. InfoGraph: Unsupervised and Semi-supervised Graph-Level Representation Learning via Mutual Information Maximization. In *ICLR*.
- Tschannen, M.; Djolonga, J.; Rubenstein, P. K.; Gelly, S.; and Lucic, M. 2020. On Mutual Information Maximization for Representation Learning. In *Proceedings of the 8th International Conference on Learning Representations*.
- Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *JMLR*, 9(11).
- Velickovic, P.; Fedus, W.; Hamilton, W. L.; Liò, P.; Bengio, Y.; and Hjelm, R. D. 2020. Deep Graph Infomax. In *ICLR*.
- Vishwanathan, S. V. N.; Schraudolph, N. N.; Kondor, R.; and Borgwardt, K. M. 2010. Graph kernels. *JMLR*, 11: 1201–1242.
- Wang, L.; He, D.; Zhang, H.; Liu, Y.; Wang, W.; Pan, S.; Jin, D.; and Chua, T.-S. 2024. GOODAT: Towards Test-Time Graph Out-of-Distribution Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15537–15545.

- Wu, J.; Chen, X.; Shi, B.; Li, S.; and Xu, K. 2023. SEGA: Structural entropy guided anchor view for graph contrastive learning. In *International Conference on Machine Learning*. PMLR.
- Wu, J.; Chen, X.; Xu, K.; and Li, S. 2022a. Structural entropy guided graph hierarchical pooling. In *International Conference on Machine Learning*, 24017–24030. PMLR.
- Wu, J.; Li, S.; Li, J.; Pan, Y.; and Xu, K. 2022b. A simple yet effective method for graph classification. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, July 23-29, 2022*. ijcai.org.
- Wu, Q.; Chen, Y.; Yang, C.; and Yan, J. 2024. Energy-based Out-of-Distribution Detection for Graph Neural Networks. In *The Eleventh International Conference on Learning Representations*.
- Wu, T.; Ren, H.; Li, P.; and Leskovec, J. 2020. Graph information bottleneck. *Advances in Neural Information Processing Systems*, 33: 20437–20448.
- Wu, Z.-F.; Wei, T.; Jiang, J.; Mao, C.; Tang, M.; and Li, Y.-F. 2021. NGC: a unified framework for learning with open-world noisy data. In *CVPR*, 62–71.
- Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2019. How Powerful are Graph Neural Networks? In *ICLR*.
- Yang, S.; Liang, B.; Liu, A.; Gui, L.; Yao, X.; and Zhang, X. 2024. Bounded and Uniform Energy-based Out-of-distribution Detection for Graphs. In *Forty-first International Conference on Machine Learning*.
- You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph contrastive learning with augmentations. In *NeurIPS*, volume 33, 5812–5823.
- Yuan, H.; Sun, Q.; Fu, X.; Ji, C.; and Li, J. 2024a. Dynamic Graph Information Bottleneck. In *Proceedings of the ACM on Web Conference 2024*, 469–480.
- Yuan, H.; Sun, Q.; Fu, X.; Zhang, Z.; Ji, C.; Peng, H.; and Li, J. 2024b. Environment-aware dynamic graph learning for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 36.
- Zhao, L.; and Akoglu, L. 2021. On using classification datasets to evaluate graph outlier detection: Peculiar observations and new insights. *Big Data*.
- Zheng, Y.; Pan, S.; Lee, V. C.; Zheng, Y.; and Yu, P. S. 2022a. Rethinking and Scaling Up Graph Contrastive Learning: An Extremely Efficient Approach with Group Discrimination. In *NeurIPS*.
- Zheng, Y.; Zheng, Y.; Zhou, X.; Gong, C.; Lee, V.; and Pan, S. 2022b. Unifying Graph Contrastive Learning with Flexible Contextual Scopes. In *ICDM*.
- Zhou, W.; Liu, F.; and Chen, M. 2021. Contrastive Out-of-Distribution Detection for Pretrained Transformers. In *EMNLP*, 1100–1111.
- Zhu, H.; Wu, J.; Liu, R.; Hou, Y.; Yuan, Z.; Li, S.; Pan, Y.; and Xu, K. 2024a. HILL: Hierarchy-aware Information Lossless Contrastive Learning for Hierarchical Text Classification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 4731–4745.
- Zhu, H.; Zhang, C.; Huang, J.; Wu, J.; and Xu, K. 2023. HiTIN: Hierarchy-aware Tree Isomorphism Network for Hierarchical Text Classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7809–7821.
- Zhu, Y.; Shi, H.; Zhang, Z.; and Tang, S. 2024b. Mario: Model agnostic recipe for improving ood generalization of graph contrastive learning. In *Proceedings of the ACM on Web Conference 2024*, 300–311.
- Zhu, Y.; Xu, Y.; Yu, F.; Liu, Q.; Wu, S.; and Wang, L. 2021. Graph contrastive learning with adaptive augmentation. In *WWW*, 2069–2080.