Subgraph Invariant Learning Towards Large-Scale Graph Node Classification

Leilei Wang[1,2], Si Shi[1*], Fei Ma[1], Fei Richard Yu[2,3*], Pengteng Li[4], Ying Tiffany He[2]

[1]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China
[2]College of Computer Science and Software Engineering, Shenzhen University, China
[3]School of Information Technology, Carleton University, Canada
[4]HKUST (GZ), AI Thrust
{wangleilei, shisi, mafei}@gml.ac.cn, {yufei,heying}@szu.edu.cn, pengteng.li@connect.hkust-gz.edu.cn

Abstract

Graph Neural Networks (GNNs) have shown efficacy in graph node classification, but face computational challenges on large-scale graphs. Although existing graph reduction methods address these issues, they still require high computational resources and fail to prioritize robust performance on out-of-distribution data. To tackle these challenges, we introduce the subgraph invariant learning paradigm, inspired by the small-world phenomenon. This approach enables models trained on specific subgraphs to generalize across diverse subgraphs, reducing computational demands, and enhancing scalability. To promote generalization, we maximize the invariance log-likelihood by deriving a theoretical lower bound of it and formulating the InVar loss. This loss minimizes the discrepancy between node representations and their corresponding invariance representations while maximizing the entropy of the node representation. In response to InVar loss, we propose the Invariance Facilitation Model (IFM), comprising the Invariance Representation Encoder (IRE) and Node Representation Encoder (NRE). IRE, capturing the invariance representations, utilizes Invariance ATTention (InvarATT) to compress long-range dependencies, while NRE learns the node representation, by integrating invariance representations via Telematic ATTention (TeleATT) and exchanging local information within each subgraph through GNNs. Evaluations on four large-scale graph datasets demonstrate the effectiveness, computational efficiency, and interpretability of IFM for large-scale graph node classification.

Code — https://github.com/wleilei/SIL-IFM

Introduction

Graph structures are prevalent in complex systems across diverse domains, including social, biological, transportation, and communication networks. The widespread use of graph-structured data has driven the adoption of Graph Neural Networks (GNNs) (Kipf and Welling 2017; Veličković et al. 2018), significantly enhancing the modeling and analysis capabilities in various fields. However, the high computational demands of modeling large-scale graphs limit the deployment of GNNs, preventing their full potential from being realized in real-world applications.

A natural solution is graph reduction learning, which includes graph sparsification, graph coarsening, and graph condensation (Hashemi et al. 2024). These methods create a smaller graph that can be used during testing to achieve performance comparable to the original large-scale graph. However, transforming a large-scale graph into a smaller one faces three limitations. First, high computational resources are still required during the transformation process, as the large-scale graph remains necessary during training. Second, the smaller graph often fails to generalize to out-of-distribution graphs, since the goal of these methods is to match the performance of the large graph rather than to enhance generalization. Third, obtaining a large number of labeled nodes that represent the entire graph is often impractical, and in many cases, only a small subset of labeled nodes is available during training.

To address these limitations, we suppose that a large-scale graph can be viewed as an aggregation of numerous small-scale subgraphs, consistent with the small-world paradigm (Watts and Strogatz 1998; Kleinberg 2000). In this paradigm, nodes within each subgraph are densely interconnected, while connections between subgraphs are relatively sparse. Despite the potential differences among these subgraphs, it is possible to use models robust enough to generalize from individual subgraphs to the overall graph structure, guided by the principle of invariance (Jia et al. 2024; Ahuja et al. 2021). This principle suggests that capturing invariant information across varying conditions requires only a subset of the available data, and this invariance can be generalized across the entire graph. Based on this, we propose subgraph invariant learning, which enables the model to be trained on labeled subgraphs and tested on unlabeled subgraphs by extrapolating invariance from certain subgraphs to others with superior performance. Since models in subgraph invariant learning are applied exclusively to individual subgraphs, without needing the entire graph during either the training or testing phases, it offers two key advantages: a substantial reduction in computational resource consumption and the ability to increase model complexity due to the smaller scale of the subgraphs.

To ensure that invariance is learned effectively, we aim to maximize the log-likelihood of the invariance and derive its lower bound. Consequently, the InVariance (InVar) loss is strategically proposed as a proxy objective to max-

*Corresponding authors.

imize this lower bound. This loss function serves two purposes: minimizing the distance between a node's representation and its same-labeled invariant representation, while simultaneously maximizing the entropy of node representations to encourage feature diversity. Based on the InVar loss, we develop the corresponding Invariance Facilitation Model (IFM), comprising the Invariance Representation Encoder (IRE) that captures shared invariant features among nodes with the same label and the Node Representation Encoder (NRE) that provides each node with the necessary information to predict its label accurately.

IRE learns a unique invariance representation from nodes with the same label, ensuring this invariant representation encapsulates sufficient label-related information. As this information comes from nodes across the entire graph, including those at considerable distances, IRE can be considered as a compressor of long-range dependencies. In traditional GNNs, over-squashing (Topping et al. 2022; Di Giovanni et al. 2023) or over-globalizing (Xing et al. 2024) may occur due to information loss when compressing sparsely connected graph data to capture long-range dependencies across a wide receptive field. IFM addresses this issue by inherently limiting the scope of information propagation to dense subgraphs, with the long-range dependencies being captured by the Invariance ATTention (InvarATT) in IRE.

NRE equips each node with sufficient information to accurately predict its label. NRE consists of two modules: Telematic ATTention (TeleATT) and GNN. TeleATT enables nodes to extract global invariant information from the representations learned by IRE, while GNN facilitates the recursive exchange of structural information between adjacent nodes, fostering local invariance within each subgraph. Notably, IFM is adaptable to any base GNN model, enhancing the ability to capture long-range dependencies with global invariance. During inference, only NRE is required for the prediction with the learned invariant representations.

We conducted extensive experiments on two synthetic datasets and two real-world datasets to validate the effectiveness of InVar loss and IFM. Additionally, we performed visualization experiments to verify the invariance property and the explainability of the node representations. Our Contributions are summarized as follows:

- To the best of our knowledge, we are the first to develop subgraph invariant learning for large-scale graphs. This approach, inspired by the small-world phenomenon, emphasizes training on labeled subgraphs and generalizing to unlabeled ones with low computational costs.
- Based on the invariance principle, we derive the lower bound of the log-likelihood of invariance and introduce the InVar loss as a proxy objective for maximizing this bound. We subsequently design the IFM architecture, incorporating IRE and NRE, to minimize the InVar loss.
- Comprehensive experiments are conducted on four large-scale graph datasets. Our proposed method has demonstrated its effectiveness, efficiency, and interpretability through the empirical results.

Related Works

Graph Reduction Learning. To accelerate GNN models and enhance data processing efficiency, graph reduction learning simplifies large-scale graphs into more manageable small-scale counterparts while preserving the core structure (Hashemi et al. 2024). These methods can be categorized into three strategies: graph sparsification, graph coarsening, and graph condensation. Graph sparsification (Yu et al. 2022; Li et al. 2023) focuses on extracting a subset of essential nodes and edges. Graph coarsening (Taghibakhshi et al. 2021; Kumar et al. 2023) aims to aggregate raw nodes and edges into super nodes and super edges. Both of them preserve graph information without directly considering model performance. On the other hand, graph condensation (Zhang et al. 2024; Zheng et al. 2023) is designed to synthesize a small-scale graph that allows models trained on it to achieve results comparable to those trained on the original large-scale graphs (Gao et al. 2024). However, these graph reduction methods often involve high computational costs when transforming large-scale graphs into smaller ones, and face limitations in generalization capability as their primary goal is not to capture invariance across the entire graph. Additionally, a large number of labeled nodes are required during the training stage to produce the smaller graph.

Graph Invariant Learning. Graph invariant learning focuses on learning the invariance of distribution-shift graphs to enhance out-of-distribution generalization (Gui et al. 2022; Chen et al. 2023). Building on this foundation, recent works (Li et al. 2022; Yue et al. 2024) have introduced various methods to address feature and structure distribution shifts. However, these methods primarily address graphs in different environments and do not consider the issue of subgraph invariance within large-scale graphs. Although some studies (Li et al. 2022; Jia et al. 2024) have explored distribution shift of subgraphs, their main goal is to identify or extract subgraph environments for subsequent generalization, which still entails high training and testing costs on large-scale graphs. Alternatively, other methods (Buffelli, Lió, and Vandin 2022; Huang et al. 2024) aim to extrapolate invariance learned from a minor subgraph to its entire large-scale graph, but they may still incur significant computational demands due to the need to process the entire graph during model inference. Substantially, graph invariant learning either ignores distribution shifts across subgraphs or requires high computational costs when dealing with large-scale graphs.

Notations and Problem Formulation

Notations. A large-scale graph is supposed to be partitioned into a set of small-scale subgraphs, denoted as $\mathcal{G} = \{G_1, \ldots, G_m\}$, where m is the number of subgraphs. For each subgraph $G_i = \{X_i, A_i, Y_i\}$, $X_i \in \mathbb{R}^{N_i \times d_{in}}$ denotes the N_i numbers of nodes with d_{in}-dimensional features, $A_i \in \mathbb{R}^{N_i \times N_i}$ denotes the adjacency matrix, and $Y_i \in \mathbb{R}^{N_i \times C}$ denotes the C-classes of node labels. S^c are the c-labeled nodes in the training subgraphs.

Problem Formulation. Denoted $\mathbb{G}$ and $\mathbb{Y}$ as the graph and label space, we formulate subgraph invariant learning over the small-scale subgraphs of a large-scale graph as:

Problem 1. *The goal of subgraph invariant learning is the finding of an optimal predictor $\mathcal{P}^*(\cdot) : \mathbb{G} \to \mathbb{Y}$ that performs well on all subgraphs:*

$$\mathcal{P}^*(\cdot) = arg\min_{\mathcal{P}} \sup_{G_i \in \mathcal{G}} \mathcal{R}(\mathcal{P}|G_i) \tag{1}$$

where $\mathcal{R}(\mathcal{P}|G_i) = \mathbb{E}_{G_i}[\mathcal{L}(\mathcal{P}(G_i), Y_i)]$ is the expected risk of the predictor $\mathcal{P}$ on the subgraph G_i, and $\mathcal{L}(\cdot,\cdot) : \mathbb{Y} \to \mathbb{Y}$ is the loss function. Additionally, we can decompose $\mathcal{P}(\cdot) = \mathcal{C} \circ \mathcal{F}$, where $\mathcal{F}(\cdot) : \mathbb{G} \to \mathbb{R}^{d_h}$ is the feature representation learning function and $\mathcal{C}(\cdot) : \mathbb{R}^{d_h} \to \mathbb{Y}$ is the classifier.

Subgraphs $\mathcal{G}$ within the large-scale graph exhibit shifts in both feature and structure distributions due to their origins in diverse environments. These distribution shifts pose significant challenges in addressing Problem 1 as they impede the effective learning of valid feature representations.

To acquire robust feature representations, we posit that each node in a large-scale graph contains invariant information that maintains a consistent relationship with its label across various subgraphs, adhering to the principle of invariance (Jia et al. 2024; Ahuja et al. 2021). By effectively identifying and leveraging this invariance, a model can achieve strong out-of-distribution (OOD) performance. We formalize this assumption as follows:

Assumption 1. *Learned from the training set of small-scale subgraphs $\mathcal{G}_{train}$, the optimal invarianct feature encoder $\mathcal{F}^*(\cdot)$ does exist, satisfying:*

- *Invariance property: For any G_i and G_j, if $\hat{X}_i^c = \mathcal{F}^*(X_i, A_i)$ and $\hat{X}_j^c = \mathcal{F}^*(X_j, A_j)$, then $\hat{X}_i^c = \hat{X}_j^c$, where $\hat{X}^c$ denotes c-labeled node representations.*
- *Sufficiency property: a predictor $\mathcal{C}^*(\cdot)$ can be found and satisfies $Y_i = \mathcal{C}^*(\mathcal{F}^*(X_i, A_i)) + \epsilon$, $\epsilon \perp \mathcal{F}^*(X_i, A_i)$, where $\perp$ indicates statistical independence, and ϵ is random noise.*

The invariance property follows with the invariance principle in contrastive learning and OOD methods, by ensuring that the node representations remain robust across different conditions. It posits that the same-labeled node representations should be same even in different subgraphs and differ from different-labeled node representations. The sufficiency property ensures that the invariant representations can be effectively utilized by a simple classifier to predict node labels. Consequently, the problem of cultivating optimal $\mathcal{F}(\cdot)$ and $\mathcal{C}(\cdot)$ arises, which can be formulated as:

Problem 2. *Given a set of small-scale subgraphs $\mathcal{G}_{train}$, the task is designed to jointly capture the invariance of node representations among all subgraphs using $\mathcal{F}^*(\cdot)$ and learn a predictor $\mathcal{C}^*(\cdot)$ with the node representations, thereby achieving robust OOD generalization performance.*

Methodology

In this section, we first introduce the lower bound of the log-likelihood of invariance and develop the corresponding InVar loss as the proxy objective for maximizing it. Then we develop IFM to learn the invariance representations and the node representations as required in InVar loss. Finally, we compare IFM with traditional GNN models, focusing on differences in time complexity, model capacity, explainability and sparsity.

Invariance Loss

Under Assumption 1, the invariant representation $\hat{x}^c$ exists within the observed c-labeled node samples s^c. Following the likelihood-based approach, $\hat{x}^c$ can be derived by maximizing the invariance likelihood $p(\hat{x}^c)$ through optimizing the feature encoder $\mathcal{F}(\cdot)$. In the following theorems, we derive the lower bound of the log-likelihood $\log p(\hat{x}^c)$ and then decompose this lower bound into three terms with realistic meanings. Based on the decomposed terms, the InVar loss has been developed as the proxy objective of maximizing the lower bound.

Theorem 1. *The lower bound of log-likelihood $\log p(\hat{x}^c)$ is $\mathbb{E}_{q_\phi(x^c)}[\log \frac{p(x^c,\hat{x}^c)}{q_\phi(x^c)}]$.*

Proof. In order to obtain the lower bound of the log-likelihood $\log p(\hat{x}^c)$, we can decompose $\log p(\hat{x}^c)$ as follows:

$$\log p(\hat{x}^c) = \log p(\hat{x}^c) \int q_\phi(x^c)\, dx^c, \tag{2}$$

$$= \int q_\phi(x^c) \log p(\hat{x}^c)\, dx^c, \tag{3}$$

$$= \mathbb{E}_{q_\phi(x^c)}[\log p(\hat{x}^c)], \tag{4}$$

$$= \mathbb{E}_{q_\phi(x^c)}\left[\log \frac{p(x^c,\hat{x}^c) q_\phi(x^c)}{p(x^c|\hat{x}^c) q_\phi(x^c)}\right], \tag{5}$$

$$= \mathbb{E}_{q_\phi(x^c)}\left[\log \frac{p(x^c,\hat{x}^c)}{q_\phi(x^c)}\right] + \mathbb{E}_{q_\phi(x^c)}\left[\log \frac{q_\phi(x^c)}{p(x^c|\hat{x}^c)}\right], \tag{6}$$

$$= \mathbb{E}_{q_\phi(x^c)}\left[\log \frac{p(x^c,\hat{x}^c)}{q_\phi(x^c)}\right] + D_{KL}(q_\phi(x^c) \| p(x^c|\hat{x}^c)). \tag{7}$$

In this equation, $q_\phi(x^c)$ can be treated as NRE for transforming the node features into a node embedding space. Since the KL Divergence is always non-negative, we can infer that $\mathbb{E}_{q_\phi(x^c)}\left[\log \frac{p(x^c,\hat{x}^c)}{q_\phi(x^c)}\right]$ is the lower bound of $\log p(\hat{x}^c)$, i.e. $\log p(\hat{x}^c) \geq \mathbb{E}_{q_\phi(x^c)}\left[\log \frac{p(x^c,\hat{x}^c)}{q_\phi(x^c)}\right]$. Then, maximizing this lower bound becomes a proxy objective of capturing the invariance $\hat{x}^c$.

Theorem 2. *Minimizing the InVar loss $\mathcal{L}$ is the maximization of the lower bound $\mathbb{E}_{q_\phi(x^c)}\left[\log \frac{p(x^c,\hat{x}^c)}{q_\phi(x^c)}\right]$.*

Proof. By splitting the numerator, the lower bound can be rewritten as:

$$\mathbb{E}_{q_\phi(x^c)}\left[\log \frac{p(x^c,\hat{x}^c)}{q_\phi(x^c)}\right] = \mathbb{E}_{q_\phi(x^c)}\left[\log \frac{p_\theta(\hat{x}^c|x^c) p(x^c)}{q_\phi(x^c)}\right]. \tag{8}$$

In this formula, $p_\theta(\hat{x}^c|x^c)$ can be considered as IRE for unearthing the invariant features from the node features. Subsequently, we can further decompose this lower bound into three terms: $\mathbb{E}_{q_\phi(x^c)}[\log p_\theta(\hat{x}^c|x^c)]$, $\mathbb{E}_{q_\phi(x^c)}[\log p(x^c)]$

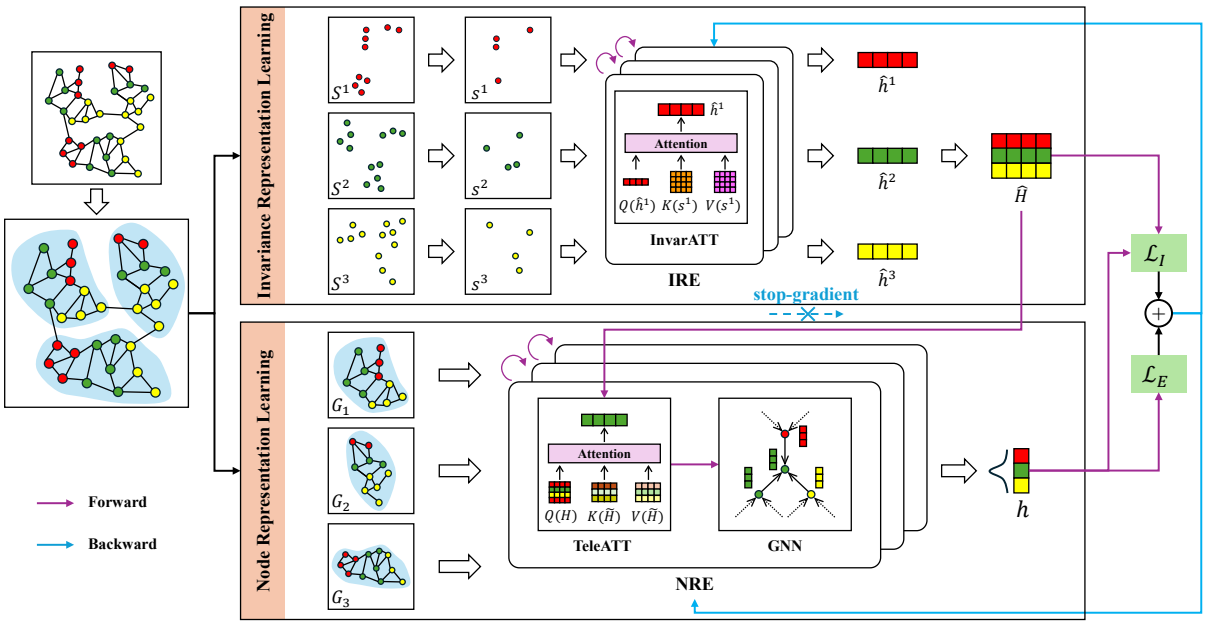


Figure 1: The overall architecture of IFM. Subgraph invariant learning partitions a large-scale graph into numerous subgraphs, aligning with the small-world phenomenon. IFM comprises two key components: IRE for invariance representation learning and NRE for node representation learning. (1) IRE employs InvarATT as a long-range dependency compressor to capture invariance representations from given node samples. (2) NRE leverages TeleATT to extract global invariance from these representations and utilizes GNN within subgraphs to foster local invariance, equipping nodes with the necessary invariance for label prediction.

and $-\mathbb{E}_{q_\phi(x^c)}[\log q_\phi(x^c)]$. All three terms of the decomposed lower bound have distinct purposes: the first term $\mathbb{E}_{q_\phi(x^c)}[\log p_\theta(\hat{x}^c|x^c)]$ measures the similarity between the invariant features and the node embeddings, ensuring that the invariant representations align closely with the actual node data. The second term $\mathbb{E}_{q_\phi(x^c)}[\log p(x^c)]$ assesses how well the node embeddings correspond to their respective labels, emphasizing the accuracy of label prediction. The third term $-\mathbb{E}_{q_\phi(x^c)}[\log q_\phi(x^c)]$ represents the entropy, which encourages diversity in the features, promoting a richer and more robust feature space.

Considering that each node embedding can be utilized for predicting its label through the invariance representations, we can apply InfoNCE (van den Oord, Li, and Vinyals 2019), a contrastive loss function, to simultaneously address the first two terms of the lower bound in Theorem 2. The InVar loss can then be developed as follows:

$$\mathcal{L} = \mathcal{L}_I - \lambda \mathcal{L}_E, \tag{9}$$

$$= -\log \frac{\exp(\text{sim}(x, \hat{x}^+)/\tau)}{\sum_{k=1}^{C} \exp(\text{sim}(x, \hat{x}^k)/\tau)} - \lambda \sum_{i=1}^{d} (-x_i \log x_i). \tag{10}$$

where the invariance representation $\hat{x}^+$ shares the same label with node x , τ is the temperature coefficient, andλ is the regularization weight. In addition, we normalize x and interpret the resulting term as a probability distribution.

Specifically, minimizing $\mathcal{L}_I$ encourages node representations to align more closely with their same-label invariance representations while remaining distinct from other invariance representations with different labels. This approach enforces the invariance property in Assumption 1 and enhances the explainability of predictions by bringing same-labeled node representations closer to their invariant counterparts. $\mathcal{L}_E$ is to maximize the entropy of node representations and helps maintain valid, numerically balanced features. This process avoids trivializing certain features (e.g., being near-zero after normalization) and encourages the model to retain a broader variety of distinctive characteristics, reinforcing the sufficiency property as outlined in Assumption 1.

Invariance Faciliation Model (IFM)

To implement the InVar loss, we propose the IFM architecture, as is shown in Figure 1. IFM consists of two main components: Invariance Representation Encoder (IRE) and Node Representation Encoder (NRE). The IRE, designed to capture invariant representations, is built from a stack of identical layers, each featuring an InvarATT sublayer and a feed-forward network. Similarly, the NRE, responsible for learning node representations, is composed of a series of repeated layers. Each layer in the NRE includes a feed-forward network, TeleATT for modeling global invariance from the IRE output, and GNN for capturing local invariance from neighboring nodes. During inference, only the NRE and the invariant representations are needed, with the softmax term in $\mathcal{L}_I$ acting as the prediction classifier.

Invariance Representation Encoder. We use the average of c-labeled node features in $\mathcal{G}_{train}$ as the initial invariance embedding $\hat{h}^c \in \mathbb{R}^{1\times d_h}$ and iteratively refine this embedding by incorporating a fixed number of random c-labeled

samples $s^c \in \mathbb{R}^{k \times d_h}$ from S^c, where $S^c \in \mathcal{S}$ represents c-labeled nodes in $\mathcal{G}_{train}$. Accordingly, IRE is implemented in a scalable, transformer-like architecture (Vaswani et al. 2023) using InvarATT, which is computed as:

$$\text{InvarATT} = \text{Attention}(Q(\hat{h}^c), K(s^c), V(s^c)). \tag{11}$$

InvarATT is an attention mechanism enabling the invariance representation $\hat{h}^c$ to assimilate information from same-labeled samples s^c, capturing label-specific invariant features across the graph. Since the nodes s^c are globally sampled from G_{train}, IRE can be viewed as a long-range dependency compressor, globally condensing invariance information into invariance representations, which are consistent across subgraphs yet specific to each label class.

Node Representation Encoder. With the invariance representations $\hat{H} \in \mathbb{R}^{C \times d_h}$ from IRE, TeleATT is designed to help each node capture its same-labeled invariance through attention. To ensure training stability, we apply a stop-gradient operation to $\hat{H}$ resulting in $\tilde{H}$, which prevents individual nodes from influencing the invariance. Meanwhile, GNN is used to capture complex relationships and dependencies within each subgraph. After mapping node representations X into $H \in \mathbb{R}^{N \times d_h}$, NRE is implemented in a scalable, transformer-like architecture using a combination of TeleATT and GNN, where TeleATT is computed as:

$$\text{TeleATT} = \text{Attention}(Q(H), K(\tilde{H}), V(\tilde{H})). \tag{12}$$

Specifically, NRE is designed to equip node representations with sufficient label correlation through this dual mechanism: (1) capturing global invariance from IRE-generated invariance representations via TeleATT, and (2) fostering local invariance from neighboring nodes within the subgraph via GNN. By employing TeleATT, NRE enhances any base GNN model's capacity to capture long-range dependencies using the compressed invariance representations. This approach effectively integrates both global and local graph structural information, resulting in more comprehensive and informative node representations.

Comparisons

We compare our proposed IFM, which operates on individual small-scale subgraphs, with traditional GNN models that run on the entire large-scale graph.

Time Complexity. The time complexity of GNN models is $O(n^2 d)$ per layer, where n is the number of nodes in the large-scale graph. In contrast, IFM operates on individual subgraphs, resulting in a time complexity of $O(ms^2 d)$ for the entire graph and $O(s^2 d)$ for a single subgraph, where m represents the number of subgraphs and $s \ll n$ is the maximum subgraph size. Thus, IFM has significantly lower time complexity compared to GNNs.

Model Capacity. According to scaling laws (Kaplan et al. 2020), the model capacity is positively correlated with the number of parameters. However, constrained by the large scale of graph data, GNN models cannot scale up their parameters effectively and struggle to capture long-range dependencies due to the limited receptive field. In contrast, IFM can scale up the number of parameters due to the smaller scale of subgraphs. IFM models long-range dependencies effectively by incorporating IRE as a long-range information compressor and using TeleATT to distribute long-range information to each node.

Explainability. To adhere to the invariance property in Assumption 1, IFM optimizes parameters to make the nodes more similar to their same-labeled invariance representations, guided by the InVar loss. This enhances explainability by bringing the c-labeled nodes closer to a fixed embedding for prediction. Conversely, GNN models use MLP as the predictor and the CrossEntropy loss for parameter optimization, which results in a lack of explainability.

Sparsity. Irregular graphs or non-hierarchical may lack clear clustering but exhibiting sparsity, which can hinder information flow in traditional GNNs and lead to over-squashing issues. IFM addresses this by removing sparse edges to partition large-scale graphs into smaller subgraphs, improving scalability. Additionally, the NRE module employs IRE and TeleATT to provide nodes with sufficient long-range information for accurate label prediction. Consequently, IFM is suitable for various large-scale graph types, even those not exhibiting the small-world property.

Experiments

In this section, we conduct the experiments on four large-scale graph datasets to answer the following research questions: ·**RQ1:** Is IFM effective for enhancing the performance of GNNs in subgraph invariant learning? ·**RQ2:** Is it necessary to incorporate $\mathcal{L}_E$ in the model training? ·**RQ3:** How sensitive is IFM to change in hyper-parameters? ·**RQ4:** Does IFM achieve the invariance property and the explainability?

Experimental Setup

Datasets. We conduct experiments on four graph node classification datasets, including two synthetic datasets and two real-world datasets. For the synthetic datasets, we utilize the **Reddit** and **Amazon** large-scale graph datasets (Zeng et al. 2020), employing a clustering algorithm (Chiang et al. 2019) to partition each graph into subgraphs. The real-world datasets, **PascalVOC** and **COCO**, are chosen from the Long Range Graph Benchmark (Dwivedi et al. 2022), where each subgraph represents an image and each node corresponds to a specific image region. To rigorously assess model performance, we divide the subgraphs of each dataset into non-overlapping training, validation, and testing subsets. Detailed dataset information is shown in Table 1.

Model Configurations. In NRE, GNN is used to aggregate information from subgraph nodes. Our method can be applied to various GNNs, including convolution-based and attention-based ones. For convolution-based GNNs, we use **GCN** (Kipf and Welling 2017), **SAGE** (Hamilton, Ying, and Leskovec 2017) (mean-pooling variant) and **SGC** (Wu et al. 2019). For attention-based GNNs, we incorporate **GAT** (Veličković et al. 2018), **GATv2** (Brody, Alon, and Yahav 2022) and **UniMP** (Shi et al. 2021). These models serve

Dataset	Subgraphs	Total Nodes	Nodes (Mean ± Std)	Edges (Mean ± Std)	Classes	Train/Val/Test
Reddit	400	232,965	582 ± 16.39	64,464 ± 66,026.99	41	80/160/160
Amazon	2,000	1,569,960	784 ± 20.29	15,461 ± 43,458.24	75	400/800/800
PascalVOC	11,355	5,443,545	479 ± 16.52	2,710 ± 96.67	41	8,498/1,428/1,429
COCO	123,286	58,793,216	477 ± 15.75	2,694 ± 93.58	81	113,286/5,000/5,000

Table 1: Statistics of the four datasets. Note that the data splitting for Pascal VOC-SP and COCO-SP follows LRGB (Dwivedi et al. 2022), and labels with no nodes are removed from Amazon (Zeng et al. 2020).

	Reddit		Amazon		PascalVOC		COCO	
	AUC-ROC	F1	AUC-ROC	F1	AUC-ROC	F1	AUC-ROC	F1
GCN	$84.74_{\pm 14.99}$	$79.79_{\pm 26.43}$	$\mathbf{97.08_{\pm 0.41}}$	$70.11_{\pm 1.85}$	$84.05_{\pm 14.44}$	$70.16_{\pm 20.68}$	$84.79_{\pm 16.32}$	$71.96_{\pm 22.82}$
GCN-IFM	$\mathbf{88.28_{\pm 11.20}}$	$\mathbf{79.87_{\pm 26.03}}$	$97.05_{\pm 0.42}$	$\mathbf{70.75_{\pm 1.83}}$	$\mathbf{84.56_{\pm 14.27}}$	$\mathbf{70.32_{\pm 20.26}}$	$\mathbf{85.07_{\pm 16.45}}$	$\mathbf{72.18_{\pm 23.35}}$
SAGE	$89.07_{\pm 14.55}$	$82.15_{\pm 24.03}$	$97.05_{\pm 0.43}$	$70.85_{\pm 1.83}$	$84.23_{\pm 14.92}$	$71.04_{\pm 20.33}$	$84.91_{\pm 16.62}$	$72.17_{\pm 23.09}$
SAGE-IFM	$\mathbf{91.18_{\pm 0.03}}$	$\mathbf{82.31_{\pm 23.89}}$	$\mathbf{97.11_{\pm 0.41}}$	$\mathbf{71.06_{\pm 1.71}}$	$\mathbf{84.41_{\pm 14.73}}$	$\mathbf{71.08_{\pm 20.28}}$	$\mathbf{85.85_{\pm 16.26}}$	$\mathbf{73.64_{\pm 21.77}}$
SGC	$88.77_{\pm 9.77}$	$81.33_{\pm 25.00}$	$97.07_{\pm 0.41}$	$70.21_{\pm 1.87}$	$83.12_{\pm 15.16}$	$69.10_{\pm 21.32}$	$85.19_{\pm 16.56}$	$72.66_{\pm 22.55}$
SGC-IFM	$\mathbf{89.97_{\pm 9.05}}$	$\mathbf{81.68_{\pm 25.16}}$	$\mathbf{97.10_{\pm 0.41}}$	$\mathbf{71.05_{\pm 1.76}}$	$\mathbf{83.74_{\pm 14.92}}$	$\mathbf{70.06_{\pm 20.84}}$	$\mathbf{85.28_{\pm 16.28}}$	$\mathbf{72.86_{\pm 22.72}}$
GAT	$82.81_{\pm 13.74}$	$78.16_{\pm 27.67}$	$97.03_{\pm 0.43}$	$69.96_{\pm 1.81}$	$83.90_{\pm 14.74}$	$70.70_{\pm 20.44}$	$85.24_{\pm 16.54}$	$73.04_{\pm 22.33}$
GAT-IFM	$\mathbf{87.40_{\pm 14.12}}$	$\mathbf{79.04_{\pm 26.45}}$	$\mathbf{97.03_{\pm 0.42}}$	$\mathbf{70.55_{\pm 1.82}}$	$\mathbf{84.60_{\pm 14.54}}$	$\mathbf{71.14_{\pm 20.84}}$	$\mathbf{85.34_{\pm 16.42}}$	$\mathbf{73.13_{\pm 22.50}}$
GATv2	$84.43_{\pm 17.23}$	$78.79_{\pm 27.51}$	$96.99_{\pm 0.43}$	$69.50_{\pm 1.99}$	$83.91_{\pm 15.02}$	$70.53_{\pm 20.93}$	$85.19_{\pm 16.47}$	$72.63_{\pm 22.78}$
GATv2-IFM	$\mathbf{86.89_{\pm 12.93}}$	$\mathbf{79.07_{\pm 26.13}}$	$\mathbf{97.04_{\pm 0.41}}$	$\mathbf{70.36_{\pm 1.79}}$	$\mathbf{84.55_{\pm 14.48}}$	$\mathbf{71.34_{\pm 20.42}}$	$\mathbf{86.01_{\pm 16.32}}$	$\mathbf{74.19_{\pm 21.65}}$
UniMP	$87.74_{\pm 15.92}$	$78.97_{\pm 28.74}$	$\mathbf{97.15_{\pm 0.40}}$	$70.41_{\pm 1.78}$	$85.79_{\pm 14.30}$	$72.74_{\pm 20.05}$	$85.94_{\pm 16.32}$	$73.36_{\pm 22.84}$
UniMP-IFM	$\mathbf{89.88_{\pm 10.85}}$	$\mathbf{80.95_{\pm 24.35}}$	$97.06_{\pm 0.42}$	$\mathbf{70.50_{\pm 1.83}}$	$\mathbf{85.89_{\pm 14.30}}$	$\mathbf{72.88_{\pm 19.97}}$	$\mathbf{86.47_{\pm 16.15}}$	$\mathbf{74.53_{\pm 21.73}}$

Table 2: Results on four datasets demonstrating the performance enhancement of traditional GNN models when augmented with InVar loss and IFM. The hidden size parameter is set to 256, 128, 32, and 64 for each dataset, respectively.

as our baseline, and we integrate them with IFM to validate the efficacy of our proposed methodology.

Evaluation and Implementation Details. We employ weighted AUC-ROC score and weighted macro F1 score as evaluation metrics. Since subgraph invariant learning aims for robust performance across subgraphs, we calculate the mean and standard deviation of these metrics across testing subgraphs, rather than globally for the entire test set. This approach provides a nuanced understanding of model performance and stability across different subgraph structures. All the experiments are conducted on a single RTX 3090 GPU, as models are trained on limited subgraphs instead of the entire graph. For fair comparison, all comparative models are trained under identical conditions, with both IRE and NRE configured with 12 layers. The parameters for predictive analysis are chosen based on the epoch with the best validation performance for the weighted macro F1 score.

Main Results (RQ1)

Table 2 shows the performance improvements achieved by incorporating IFM into various GNN models. Our method enhances the performance of base GNN models across both synthetic and real-world datasets.

For synthetic datasets, the improvements are statistically significant. On the Reddit dataset, GCN-IFM increases GCN's weighted AUC-ROC score from 84.74 to 88.28, while GAT-IFM improves GAT's weighted F1 score from 78.16 to 79.04. These enhancements also reduce standard deviations, suggesting that our IFM effectively captures invariance across subgraphs with complex structural shifts, leading to better generalization.

The effectiveness of IFM is also evident in real-world datasets, showing improved performance across all base GNN models. This enhancement underscores IFM's ability to capture long-range dependencies in node representations. The importance of this capability is further highlighted by using datasets from the Long Range Benchmark (Dwivedi et al. 2022), which is specifically designed to assess performance in capturing long-range dependencies.

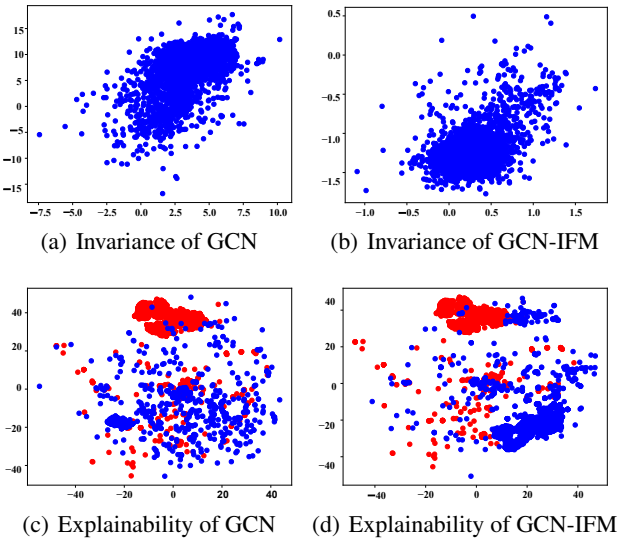
Across the four datasets, performance variations among IFM implementations with different GNN models highlight the impact of GNN architectures in modeling local invariance by enabling local information exchange and capturing structural properties. Besides, standalone GNN models underperform compared to those incorporating IRE, TeleATT, and InVar loss, testifying the improved out-of-distribution generalization achieved by the global invariance.

Ablation Study (RQ2)

We conduct ablation experiments to assess the impact of the entropy loss $\mathcal{L}_E$. The results in Table 3 show that adding $\mathcal{L}_E$ to the InVar loss leads to improved performance. This suggests that leveraging the lower bound of invariance enhances model performance. Specifically, our experiments demonstrate that maximizing node representation entropy improves prediction ability. This is likely because entropy maximiza-

Model	Loss	AUC-ROC	F1
GCN-IFM	w/o $\mathcal{L}_E$	$85.65_{\pm 9.93}$	$71.79_{\pm 29.57}$
	w/ $\mathcal{L}_E$	$\mathbf{86.36_{\pm 8.88}}$	$\mathbf{73.61_{\pm 28.75}}$
GAT-IFM	w/o $\mathcal{L}_E$	$84.83_{\pm 10.42}$	$68.02_{\pm 29.71}$
	w/ $\mathcal{L}_E$	$\mathbf{85.28_{\pm 11.12}}$	$\mathbf{71.07_{\pm 29.56}}$

Table 3: Ablation of the entropy component on Reddit, with the hidden size parameter set to 16 for a clear comparison.



(a) Invariance of GCN (b) Invariance of GCN-IFM

(c) Explainability of GCN (d) Explainability of GCN-IFM

Figure 2: Visualizations of node representations, which are the outputs of each model for prediction on Reddit. The upper two figures visualize two randomly selected attributes from node representations with the same labels. The lower two figures compare the distributions of node representations using t-SNE, with color indicating shared labels.

tion enriches the feature space, enabling the model to explore a wider range of features. These enriched features help node representations better satisfy the sufficiency property outlined in Assumption 1, thereby contributing to enhanced generalization ability.

Hyper-parameter Sensitivity Analysis (RQ3)

To validate the effectiveness of the long-range dependency compressor IRE, we analyze the sensitivity of performance to the number of samples s^c using experiments with GCN-IFM, as shown in Table 4. The results provide two key insights: (1) Our proposed model maintains strong performance even with a small number of samples s^c. This robustness is due to the final invariance representations $\hat{H}$, which can be globally utilized by adjusting IRE parameters through back-propagation. This adaptability ensures that learned representations are effectively leveraged, even with limited node samples, leading to consistent performance across varying sample sizes. (2) Performance improves as the number of samples s^c increases, highlighting the model's enhanced ability to capture long-range dependencies. This

Number of samples s^c	AUC-ROC	F1
32	$88.80_{\pm 11.77}$	$81.16_{\pm 25.27}$
64	$89.31_{\pm 11.07}$	$81.64_{\pm 24.76}$
128	$89.56_{\pm 10.87}$	$81.71_{\pm 25.32}$
256	$89.66_{\pm 10.74}$	$81.77_{\pm 24.84}$
512	$90.00_{\pm 10.31}$	$81.82_{\pm 24.79}$

Table 4: Performance analysis under varying settings of the number of samples per class s^c on Reddit.

improvement emphasizes the encoder's role in integrating information from distant parts of the data, which is crucial for overall performance enhancement.

Visualization (RQ4)

We devise two types of visualizations to verify the invariance property and explainability of our proposed method. The results in Figure 2 highlight two key aspects: (1) The upper two figures show that GCN-IFM node representations exhibit less variation and fewer outliers compared to GCN. This indicates that GCN-IFM produces more consistent and stable node representations, enhancing model robustness and achieving of the invariance property. (2) The lower two figures demonstrate that in GCN-IFM, same-labeled node representations cluster tightly around a central point, with clear separations between clusters of different labels. In contrast, GCN node representations form multiple clusters with overlapping boundaries between labels. This suggests that GCN-IFM achieves clearer and more distinct class separation, highlighting its superior performance in capturing invariance and maintaining explainability in label prediction based on node representations. These visualizations collectively support the effectiveness of our method in achieving invariance and enhancing model interpretability.

Conclusion

In this work, we develop the subgraph invariant learning paradigm to address computational challenges and out-of-distribution performance issues in large-scale graphs. Leveraging the small-world phenomenon, our approach reduces computational costs and enhances scalability by focusing on manageable subgraphs. We propose the InVar loss to maximize the lower bound of invariance likelihood. Comprising IRE and NRE, IFM aligns with this loss, which compresses long-range dependencies and facilitates global and local information exchange. Experiments on large-scale graphs demonstrate IFM's generalization capabilities, efficiency, and interpretability. Our approach addresses the limitations of graph reduction methods, offering a novel direction for graph node classification. Additionally, deriving a theoretical lower bound for invariance likelihood provides deeper insights into capturing invariance, bridging the gap between theoretical foundations and practical applications in invariance representation learning.

Acknowledgements

This research is supported in part by the National Natural Science Foundation of China (NSFC) under Grants Nos. 62406197, 62271324, 62231020, and 62371309, and the Shenzhen Science and Technology Program under Grant No. ZDSYS20220527171400002.

References

Ahuja, K.; Caballero, E.; Zhang, D.; Bengio, Y.; Mitliagkas, I.; and Rish, I. 2021. Invariance Principle Meets Information Bottleneck for Out-of-Distribution Generalization. In *Proceddings of the Thirty-Fifth Conference on Neural Information Processing Systems*, 3438–3450.

Brody, S.; Alon, U.; and Yahav, E. 2022. How Attentive are Graph Attention Networks? In *International Conference on Learning Representations*.

Buffelli, D.; Lió, P.; and Vandin, F. 2022. SizeShiftReg: a Regularization Method for Improving Size-Generalization in Graph Neural Networks. In *Proceddings of the Thirty-sixth Conference on Neural Information Processing Systems*, volume 35, 31871–31885.

Chen, Y.; Bian, Y.; Zhou, K.; Xie, B.; Han, B.; and Cheng, J. 2023. Does Invariant Graph Learning via Environment Augmentation Learn Invariance? In *Proceddings of the Thirty-seventh Conference on Neural Information Processing Systems*, 71486–71519.

Chiang, W.-L.; Liu, X.; Si, S.; Li, Y.; Bengio, S.; and Hsieh, C.-J. 2019. Cluster-GCN: An Efficient Algorithm for Training Deep and Large Graph Convolutional Networks. In *Proceedings of the Twenty-fifth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 257–266.

Di Giovanni, F.; Giusti, L.; Barbero, F.; Luise, G.; Liò, P.; and Bronstein, M. 2023. On Over-squashing in Message Passing Neural Networks: the Impact of Width, Depth, and Topology. In *Proceedings of the Fortieth International Conference on Machine Learning*.

Dwivedi, V. P.; Rampášek, L.; Galkin, M.; Parviz, A.; Wolf, G.; Luu, A. T.; and Beaini, D. 2022. Long Range Graph Benchmark. In *Proceedings of the Thirty-sixth Conference on Neural Information Processing Systems*, volume 35, 22326–22340.

Gao, X.; Yu, J.; Jiang, W.; Chen, T.; Zhang, W.; and Yin, H. 2024. Graph Condensation: A Survey. arXiv:2401.11720.

Gui, S.; Li, X.; Wang, L.; and Ji, S. 2022. GOOD: A Graph Out-of-Distribution Benchmark. In *Proceddings of the Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2059–2073.

Hamilton, W. L.; Ying, R.; and Leskovec, J. 2017. Inductive Representation Learning on Large Graphs. In *Proceedings of the Thirty-first International Conference on Neural Information Processing Systems*, 1025–1035.

Hashemi, M.; Gong, S.; Ni, J.; Fan, W.; Prakash, B. A.; and Jin, W. 2024. A Comprehensive Survey on Graph Reduction: Sparsification, Coarsening, and Condensation. arXiv:2402.03358.

Huang, Z.; Yang, Q.; Zhou, D.; and Yan, Y. 2024. Enhancing Size Generalization in Graph Neural Networks through Disentangled Representation Learning. In *Proceddings of the Forty-first International Conference on Machine Learning*.

Jia, T.; Li, H.; Yang, C.; Tao, T.; and Shi, C. 2024. Graph Invariant Learning with Subgraph Co-mixup for Out-Of-Distribution Generalization. In *Proceedings of the Thirty-seventh AAAI Conference on Artificial Intelligence*, 8562–8570.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling Laws for Neural Language Models. arXiv:2001.08361.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the Fifth International Conference on Learning Representations*.

Kleinberg, J. M. 2000. Navigation in a Small World. *Nature*, 406(6798): 845–845.

Kumar, M.; Sharma, A.; Saxena, S.; and Kumar, S. 2023. Featured Graph Coarsening with Similarity Guarantees. In *Proceedings of the Fortieth International Conference on Machine Learning*, 17953–17975.

Li, G.; Duda, M.; Zhang, X.; Koutra, D.; and Yan, Y. 2023. Interpretable Sparsification of Brain Graphs: Better Practices and Effective Designs for Graph Neural Networks. In *Proceedings of the Twenty-ninth ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1223–1234.

Li, H.; Zhang, Z.; Wang, X.; and Zhu, W. 2022. Learning Invariant Graph Representations for Out-of-Distribution Generalization. In *Proceddings of the Thirty-sixth Conference on Neural Information Processing Systems*, volume 35, 11828–11841.

Shi, Y.; Huang, Z.; Feng, S.; Zhong, H.; Wang, W.; and Sun, Y. 2021. Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification. arXiv:2009.03509.

Taghibakhshi, A.; MacLachlan, S.; Olson, L.; and West, M. 2021. Optimization-Based Algebraic Multigrid Coarsening Using Reinforcement Learning. In *Proceedings of the Thirty-Fifth Conference on Neural Information Processing Systems*, volume 34, 12129–12140.

Topping, J.; Giovanni, F. D.; Chamberlain, B. P.; Dong, X.; and Bronstein, M. M. 2022. Understanding Over-squashing and Bottlenecks on Graphs via Curvature. In *Proceedings of the Tenth International Conference on Learning Representations*.

van den Oord, A.; Li, Y.; and Vinyals, O. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2023. Attention Is All You Need. arXiv:1706.03762.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *Proceedings of the Sixth International Conference on Learning Representations*.

Watts, D. J.; and Strogatz, S. H. 1998. Collective Dynamics of 'Small-World' Networks. *Nature*, 393(6684): 440–442.

Wu, F.; Souza, A.; Zhang, T.; Fifty, C.; Yu, T.; and Weinberger, K. 2019. Simplifying Graph Convolutional Networks. In *Proceedings of the Thirty-sixth International Conference on Machine Learning*, 6861–6871.

Xing, Y.; Wang, X.; Li, Y.; Huang, H.; and Shi, C. 2024. Less is More: on the Over-Globalizing Problem in Graph Transformers. In *Proceedings of the Forty-first International Conference on Machine Learning*.

Yu, S.; Alesiani, F.; Yin, W.; Jenssen, R.; and Principe, J. C. 2022. Principle of Relevant Information for Graph Sparsification. In *Proceedings of the Thirty-eighth Conference on Uncertainty in Artificial Intelligence*, 2331–2341.

Yue, L.; Liu, Q.; Liu, Y.; Gao, W.; Yao, F.; and Li, W. 2024. Cooperative Classification and Rationalization for Graph Generalization. In *Proceedings of the Thirty-third ACM on Web Conference 2024*, 344–352.

Zeng, H.; Zhou, H.; Srivastava, A.; Kannan, R.; and Prasanna, V. 2020. GraphSAINT: Graph Sampling Based Inductive Learning Method. In *Proceedings of the Eighth International Conference on Learning Representations*.

Zhang, Y.; Zhang, T.; Wang, K.; Guo, Z.; Liang, Y.; Bresson, X.; Jin, W.; and You, Y. 2024. Navigating Complexity: Toward Lossless Graph Condensation via Expanding Window Matching. In *Proceddings of the Forty-first International Conference on Machine Learning*.

Zheng, X.; Zhang, M.; Chen, C.; Nguyen, Q. V. H.; Zhu, X.; and Pan, S. 2023. Structure-free Graph Condensation: From Large-scale Graphs to Condensed Graph-free Data. In *Proceddings of the Thirty-seventh Conference on Neural Information Processing Systems*, volume 36, 6026–6047.