Edge Self-Adversarial Augmentation Enhances Graph Contrastive Learning Against Neighborhood Inconsistency

Chunchun Chen[1], Xing Wei[2], Jiayi Yang[2], Chenrun Wang[2], Yiwei Fu[3], Yuxing Zhang[4], Xin Sun[5], Rui Fan[1, 2, 6], Wei Ye[1, 2*]

[1]Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, China
[2]College of Electronic and Information Engineering, Tongji University, China
[3]School of Mathematical Sciences, Peking University, China
[4]School of Computer Science, Shanghai Jiao Tong University, China
[5]Faculty of Data Science, City University of Macau, Taipa, Macau, China
[6]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University, China
{c2chen, xing627, 2111125, wcr0905, yew}@tongji.edu.cn, fuyw@stu.pku.edu.cn
zhangyuxing@sjtu.edu.cn, {sunxin1984, rui.fan}@ieee.org

Abstract

Recent studies have shown that unsupervised graph contrastive learning (GCL) is vulnerable to adversarial attacks. Automatic adversarial augmentation techniques are proposed to improve both the effectiveness and robustness of GCL. Existing methods typically regard unsupervised contrastive loss as the adversarial goal, essentially aiming to maximize inter-view instance-wise discrepancies between adversarial and original views. However, such attacks overlook intra-view neighborhood inconsistency, which hinders the robustness of GCL models against local neighborhood noises, resulting in performance degradation on low-homophily graphs. To tackle this issue, we propose a novel adversarial contrastive paradigm, named **E**dge self-a**D**versarial **A**ugmentation for **G**raph **C**ontrastive **L**earning (EDA-GCL). We theoretically establish that the adversarial objective of the intra-view neighborhood is equivalent to maximizing the discrepancy between bidirectional edge features. Hence, we build our adversarial framework based on edge self-adversarial learning. It generates pairwise adversarial augmentations from the original view by learning distinct neighborhood connectivity structures. The learned pairwise adversarial views are utilized for GCL model training in the minimization stage. Notably, this edge-level adversarial approach reduces the computational complexity to the level of the edge number. Experiments on various graph tasks and complex noise scenarios demonstrate the superiority and robustness of our EDA-GCL.

Code — https://github.com/CCChen-GEEX/EDA-GCL

Introduction

Graph-structured data can be found widely in real-world applications, such as citation networks and social networks. Graph Neural Networks (GNNs) (Kipf and Welling 2016; Veličković et al. 2018a; Hamilton, Ying, and Leskovec 2017; Veličković et al. 2018b) aim to encode graph-structured data into low-dimensional representations. They are effective methods for graph representation learning. However, most of them need sufficient sample labels to perform well in real-world scenarios. As a result, self-supervised learning (SSL) methods (Zhu, Peng, and Chen 2021; Wu et al. 2023; Xie et al. 2022; Liu et al. 2022) have garnered notable attention due to their independence from real labels.

Among SSL methods, contrastive learning (He et al. 2020; Chen et al. 2020) has emerged as a mainstream paradigm for learning expressive representations due to its effectiveness in capturing semantic similarities. Graph contrastive learning (GCL) (Zhu et al. 2020, 2021; Li et al. 2023; Xiao et al. 2023; He et al. 2024; Yang et al. 2025; Ning et al. 2025; Fu et al. 2025) involves generating multiple views through stochastic augmentations to form positive and negative pairs, and then training an encoder to maximize the agreement between embeddings of positive pairs while minimizing that of negative pairs. This paradigm has proved competitive with fully supervised models in various downstream tasks (Wu et al. 2023).

Recent work has revealed that GNNs are vulnerable to adversarial attacks (Zügner, Akbarnejad, and Günnemann 2018; Zügner and Günnemann 2019; Zhang et al. 2024); thus, the GNN-derived GCL paradigm is also inevitably affected. Subtle yet-crafted perturbations imposed on graphs can seriously degrade the quality of the representations learned by GCL models, which affects practical applications in unsupervised noise scenarios. This vulnerability highlights the importance of developing robust GCL frameworks for resisting adversarial distortions and maintaining reliable performance under attack.

Hence, many GCL models combine adversarial mechanisms to address this issue, enhancing robustness by learning invariant representations that are resilient to adversarial perturbations. For example, AD-GCL (Suresh et al. 2021) designs a learnable edge-dropping augmenter by minimizing mutual information between the original and adversarial views. GASSL (Yang, Zhang, and Yang 2021) generates adversarial views by adding learnable noise to input features and minimizing inter-view mutual information. SP-AGCL (In, Yoon, and Park 2023) utilizes gradient ascent (Zhang et al. 2022) to flip the edges to maxi-

*Corresponding Author

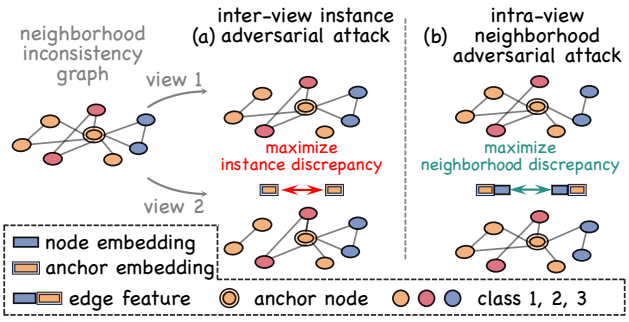

Figure 1: Comparison of adversarial strategies. Neighborhood inconsistency means that the neighbors of the anchor node from multiple classes. (a) Previous inter-view instance adversarial attack maximizes the discrepancy between anchor node embeddings across views, but uniformly perturbs all neighbors, potentially harming clean edges and linking the anchor to incorrect classes. (b) Our intra-view neighborhood adversarial attack focuses on edge-level asymmetry: by perturbing bidirectional edge features, it amplifies discrepancies for noise edges, leading to their down-weighting during aggregation. This improves neighborhood consistency by preserving clean edges while suppressing noisy ones.

mize contrastive loss and thereby obtain an adversarial view. ARIEL (Feng et al. 2024) uses the projected gradient descent (PGD) attack (Xu et al. 2019) to maximize contrastive loss. It is not hard to find that the essence of maximizing contrastive loss (or minimizing mutual information) between different augmented views is to decrease the similarity of the same instances across views (i.e., maximize instance discrepancy), as shown in Fig. 1 (a).

Despite these adversarial GCL methods have made progress, some challenges remain. **Firstly**, current methods ignore local neighborhood noises, e.g., not all neighboring nodes have the same label in heterophily graphs, resulting in degraded performance of GNNs. Existing adversarial GCL methods assume high homophily and overlook heterophily-induced noise, restricting their generalization. **Secondly**, gradient attack-based methods (e.g., ARIEL, SP-AGCL) suffer from high computational cost due to multiple forward/backward passes per iteration to generate adversarial perturbations, limiting their practicality.

To address these challenges, this paper seeks a novel adversarial paradigm: *intra-view neighborhood adversarial attack*. Through theoretical analysis, we find that the local neighborhood adversarial task is equivalent to the bidirectional edge feature self-adversarial problem. Building upon this, we further develop an approach, named **E**dge self-a**D**versarial **A**ugmentation for **G**raph **C**ontrastive **L**earning (**EDA-GCL**). Instead of maximizing the representation discrepancy of the same node across views, EDA-GCL implements adversarial attacks by maximizing the representation discrepancies between the anchor node and its neighboring nodes, as displayed in Fig. 1 (b). Specifically, our framework is an alternating min-max optimization problem: A learnable view generator constructs pairwise adversarial views by maximizing bidirectional edge features; and then the learned pairwise adversarial views with distinct neighborhood connectivity patterns are fed into a shared GNN encoder to minimize contrastive loss. As our paradigm operates at the edge level, its computational cost scales linearly with the number of edges, resulting in favorable scalability. In short, our work makes the following contributions.

- **Theoretical Analysis.** We reveal that the nature of traditional adversarial attack is to disrupt the alignment of the same instance across views. To defend against neighborhood noise, the paradigm evolves into maximizing intra-view neighborhood discrepancies and is theoretically equivalent to the edge self-adversarial attack problem.
- **Practical Solution.** We propose a novel edge self-adversarial attack paradigm, EDA-GCL, designed to handle neighborhood inconsistency, thereby addressing the critical limitation that existing adversarial GCL models are susceptible to local neighborhood noises.
- **Experimental Validation.** Comprehensive experiments demonstrate that EDA-GCL outperforms existing baselines across various graph downstream tasks and noise scenarios, indicating the effectiveness, efficiency, and robustness of the novel edge self-adversarial paradigm.

Related Work

Graph Contrastive Learning Existing GCL paradigm primarily follows the two-stage augmenting-contrasting framework (Chen, Lei, and Wei 2024; Zhao et al. 2025; Ren et al. 2025). For example, DGI (Veličković et al. 2018a) contrasts global graph embeddings with node local embeddings. MVGRL (Hassani and Khasahmadi 2020) utilizes graph diffusion to generate augmented views, and then contrasts first-order neighbors with the augmented view. GRACE (Zhu et al. 2020) constructs two augmented views by randomly dropping edges and masking features. GCA (Zhu et al. 2021) considers the importance of nodes to adaptively modify edges and features. SimGRACE (Xia et al. 2022) perturbs online networks to generate augmented outputs contrasted with the original outputs. CSGCL (Chen et al. 2023) considers the inherent community importance to guide the deletion of the node features and edges for augmented view constructions. However, existing studies have shown that GCL is susceptible to noise, and its robustness in neighborhood noise scenarios has rarely been studied.

Adversarial Graph Contrastive Learning Recently, some adversarial GCL methods have been proposed to alleviate the dilemma of being unable to effectively defend against noise in unsupervised settings. DGI-ADV (Xu et al. 2022) alternately optimizes DGI and adversarial attack training to learn robust representations. AD-GCL (Suresh et al. 2021) utilizes a learnable augmenter to construct an adversarial view by maximizing the contrastive loss between the same instance of the clean and adversarial views. ARIEL (Feng et al. 2024) learns an adversarial view by maximizing unsupervised contrastive loss via PGD attack. SP-AGCL (In, Yoon, and Park 2023) adds a k-NN augmented view based on ARIEL to preserve the node feature similarity, but its essence of adversarial attacks is still to maximize

the contrastive loss. Existing adversarial GCL methods exhibit limited robustness to neighborhood noise, because their adversarial mechanisms focus on disrupting consistency between the same instances across views. To address this problem, we propose a novel adversarial mechanism that considers local neighborhood noise.

Preliminaries

Notations Let us denote $\mathcal{G} = (\mathbf{A}, \mathbf{X})$ as an attributed graph with M directed edges, $\mathbf{A} \in \mathbb{R}^{N\times N}$ denote the adjacency matrix with $\mathbf{A}_{ij} = 1$ if and only if the i-th and the j-th nodes are connected by an edge, otherwise $\mathbf{A}_{ij} = 0$. $\mathbf{X} \in \mathbb{R}^{N\times F}$ and $\mathbf{Z} \in \mathbb{R}^{N\times d}$ denote the node feature and embedding matrices, respectively, where F and d are the dimensions of the feature and embedding.

Graph Contrastive Learning EDA-GCL consists of two stages: augmentation and contrast. First, given a clean graph input $\mathcal{G} = (\mathbf{A}, \mathbf{X})$, GCL model independently generates pairwise augmented views $\mathcal{G}^1 = (\mathbf{A}^1, \mathbf{X}^1)$ and $\mathcal{G}^2 = (\mathbf{A}^2, \mathbf{X}^2)$ by dropping edges or masking features. Then, the obtained augmented views are fed into a shared GNN encoder $f_\Psi : \mathcal{G} \to \mathbf{Z}$ with parameters Ψ to achieve the embeddings $\mathbf{Z}^1 = f_\Psi(\mathcal{G}^1)$ and $\mathbf{Z}^2 = f_\Psi(\mathcal{G}^2)$. Finally, the contrastive objective aims to pull together the embeddings of the same instance in different views and push away the embeddings of different instances. Consistent with the prior GCL method, the symmetric loss is defined as follows:

$$\begin{aligned}\mathcal{L}_{cont}(\mathbf{Z}^1, \mathbf{Z}^2) &= \frac{1}{2N}\sum\nolimits_{i=1}^{N}\left[l\left(\mathbf{z}_i^1, \mathbf{z}_i^2\right) + l\left(\mathbf{z}_i^2, \mathbf{z}_i^1\right)\right] \\ l\left(\mathbf{z}_i^1, \mathbf{z}_i^2\right) &= -\log\frac{e^{\theta\left(\mathbf{z}_i^1, \mathbf{z}_i^2\right)/\tau}}{\sum_k e^{\theta\left(\mathbf{z}_i^1, \mathbf{z}_k^2\right)/\tau} + \sum_{k\neq i} e^{\theta\left(\mathbf{z}_i^1, \mathbf{z}_k^1\right)/\tau}}\end{aligned} \tag{1}$$

where $\mathbf{z}_i^1$ and $\mathbf{z}_i^2$ stand for the i-th row of $\mathbf{Z}^1$ and $\mathbf{Z}^2$, respectively. τ is a temperature parameter. $\theta(\cdot,\cdot)$ is the similarity function, which in practice is typically the cosine similarity over their projected embeddings using a two-layer MLP as the projection head.

Rethinking Adversarial GCL

We first present the basic paradigm of existing adversarial GCL models. Then, we analyze the nature of adversarial attacks on GCL. Next, we examine the limitations of current models. Finally, we outline the motivation of our method.

Existing Adversarial Attack Paradigm The objective of most existing adversarial attacks (In, Yoon, and Park 2023; Feng et al. 2024) in GCL is to maximize the contrastive loss in Eq. (1) through adding subtle perturbations. To simplify notations for analysis, we use the asymmetric loss, which can be formally represented as follows:

$$\delta_p^* = \arg\max_{\delta_p\in\Delta} \frac{1}{N}\sum\nolimits_{i=1}^{N} l\left(\mathbf{z}_i, \mathbf{z}_i^{adv}\right) \tag{2}$$

where Δ is the perturbation set, and δ_p is perturbation budget. $\mathbf{Z} = f_\Psi(\mathcal{G})$ and $\mathbf{Z}^{adv} = f_\Psi(\mathcal{G} + \delta_p)$ are the embeddings of the original and adversarial views, respectively. Hence, Eq. (2) aims to find optimal perturbations imposed on the original graph that maximally increase contrastive loss. Previous work (Wang and Isola 2020) has shown that contrastive loss includes two parts: *alignment* between interview positive pairs and *uniformity* of the induced distribution within all samples. We refine Eq. (2) to separate two parts:

$$\begin{aligned}\delta_p^* = \arg\max_{\delta_p\in\Delta}\ &\underbrace{-\frac{1}{N}\sum\nolimits_{i=1}^{N}\theta\left(\mathbf{z}_i, \mathbf{z}_i^{adv}\right)/\tau}_{\mathcal{L}_{\text{align}}} \\ &+ \underbrace{\frac{1}{N}\sum_{i=1}^{N}\log\left(\sum\nolimits_k e^{\theta\left(\mathbf{z}_i, \mathbf{z}_k^{adv}\right)/\tau} + \sum\nolimits_{k\neq i} e^{\theta(\mathbf{z}_i, \mathbf{z}_k)/\tau}\right)}_{\mathcal{L}_{\text{uniform}}}\end{aligned} \tag{3}$$

Next, we explore the nature of adversarial attacks on GCL based on Eq. (3), that is, the adversarial process is dominated by the alignment part rather than the uniformity part.

Nature of Adversarial Attacks on GCL Adversarial attack often has high computational costs as computing gradients based on the full contrastive loss, and the main cost lies in the uniformity part (pairwise similarity computations between all samples). Here, we reveal that the adversarial process is dominated by the alignment rather than the uniformity, from both theoretical and empirical perspectives.

1) Theoretical Analysis. The existing adversarial GCL has verified that the adversarial process should be conducted under subtle attacks. The following theorem nicely confirms that such subtle attacks are capable of degrading alignment quality while preserving the overall uniformity of the distribution. The detailed proof can be found in the Appendix.

Theorem 1. *Let $\mathcal{G}^{(0)} = \mathcal{G}$ be the original graph, and $\mathcal{G}^{(t)}$ be the adversarial graph at the t-th iteration. Assume that $\mathcal{G}^{(t)}$ contains less information than $\mathcal{G}^{(t-1)}$, $\mathcal{G}^{(t)}$ should be less similar to $\mathcal{G}^{(0)}$ than $\mathcal{G}^{(t-1)}$, such that $\theta(\mathbf{z}_i, \mathbf{z}_i^{(t)}) < \theta(\mathbf{z}_i, \mathbf{z}_i^{(t-1)})$ for all node embeddings. Under subtle attacks, the alignment loss $\mathcal{L}_{align}$ is mainly perturbed, while the uniformity loss $\mathcal{L}_{uniform}$ relatively converges stably. Formally,*

$$\Delta\mathcal{L}_{align}^{t-1\to t} > 0, \ \ \Delta\mathcal{L}_{uniform}^{t-1\to t} < 0. \tag{4}$$

where $\Delta\mathcal{L}_*^{t-1\to t}$ denotes the difference in loss between the current and preceding iteration. A negative value indicates that the loss is decreasing, and vice versa.

Remark 1. Theorem 1 demonstrates that subtle attacks primarily disrupt the alignment property while preserving the uniformity of the distribution; however, computing the gradients of the uniformity loss entails substantial computational cost in the adversarial process.

2) Empirical Study. We further conduct an empirical study to verify whether subtle attacks primarily perturb the alignment loss rather than uniformity. We first train an adversarial GCL model, SP-AGCL (In, Yoon, and Park 2023), on the CiteSeer and Chameleon datasets. We follow perturbation settings in open source code to train the model, which means that all adversarial attacks are subtle. Then, we decompose and visualize the contrastive loss between the clean and adversarial views, as shown in Fig. 2. As expected, the alignment loss fluctuates during training, while the uniformity loss achieves smooth and fast convergence under adversarial attacks, which confirms our theoretical analysis.

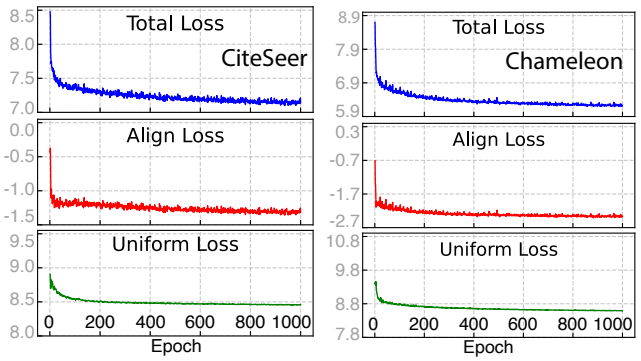

Figure 2: The contrastive loss between the original and adversarial views on SP-AGCL model reveals that the instability during the training process is mainly caused by $\mathcal{L}_{\text{align}}$.

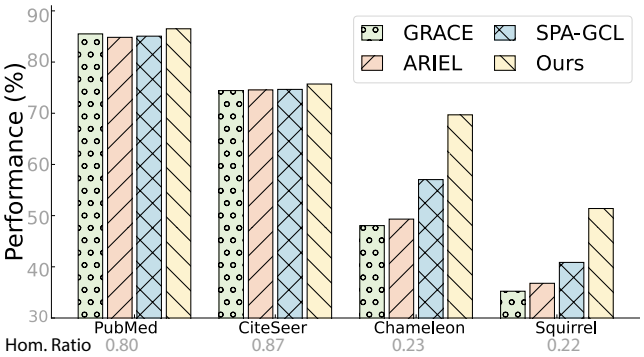

Figure 3: As Homophily Ratio decreases, the performance of adversarial GCL models degrades. Low-homophily graphs are typical neighborhood noise scenarios.

Limitation of Existing Method As analyzed above, the nature of adversarial attacks is to undermine the alignment property of GCL, whose goal is to promote the discrepancies of the same instance across views under subtle attacks. We term this mechanism as *inter-view instance-wise adversarial attack*. However, in real scenarios, GNNs are vulnerable to local neighborhood perturbations, thereby presenting a more significant challenge to the robustness of GCL models.

1) Empirical Study. To confirm our argument, we conduct experiments on adversarial GCL models, aiming to investigate whether or not the current adversarial mechanism indeed fails to defend against local neighborhood noises. Low homophily datasets serve as the ideal setting for studying local neighborhood noises; lower homophily indicates noisier neighborhoods. From the results in Fig. 3, we have two key observations: **1)** all model performance degrades with decreasing homophily ratio, revealing that local neighborhood noises indeed undermine the robustness of GCL; **2)** ARIEL and SP-AGCL yield only marginal performance gains, underscoring their vulnerability to local neighborhood noises. In contrast, our method delivers stable improvements, particularly on datasets with lower homophily ratios.

Motivation of Our Method Given the above analysis of the nature and limitations of existing adversarial GCL models, we are motivated to construct a novel adversarial paradigm that considers disrupting the feature alignment of local neighborhoods to achieve more efficient and robust adversarial contrastive learning. We term this novel mechanism as *intra-view neighborhood adversarial attack*.

Proposed Method

In this section, we propose a novel adversarial contrastive paradigm, named **E**dge self-a**D**versarial **A**ugmentation for **G**raph **C**ontrastive **L**earning (EDA-GCL), facilitating unsupervised graph representation learning. EDA-GCL achieves stronger adversarial robustness by challenging harder local neighborhood noise in an efficient manner.

Framework Overview The method contains two key modules: (1) *Edge Self-Adversarial Attack*, which generates two adversarial views with local neighborhood discrepancies via a learnable view generator including a GNN encoder g_Φ and an MLP_Ω. (2) *Contrastive Learning Training*, which optimizes another GNN encoder f_Ψ to maximize the agreement between the embeddings of the two adversarial augmented views. The framework is illustrated in Fig. 4.

Edge Self-Adversarial Attack Recall from the analysis in the previous section that the nature of adversarial attacks lies in disrupting alignment. As shown in (Jing et al. 2021), the squared loss formulation is equivalent to the standard cosine similarity for normalized embeddings in Eq. (3), if τ is set to 1. To further implement the *intra-view neighborhood adversarial attack*, we formally reformulate the problem as:

$$\delta_p^* = \arg\max_{\delta_p \in \Delta} -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{N}_i|}\sum_{k\in\mathcal{N}_i}^{|\mathcal{N}_i|} -\frac{1}{2}\|\mathbf{z}_i - \mathbf{z}_k\|_2^2 + \frac{1}{2} \quad (5)$$

where $\mathcal{N}_i$ denotes the neighbor set of the node i. Let $\mathbf{e}_{u,v} = [\mathbf{z}_u; \mathbf{z}_v]$ and $\mathbf{e}_{v,u} = [\mathbf{z}_v; \mathbf{z}_u]$ represent edge features in different directions between two nodes u and v, respectively. Therefore, the edge feature matrices of the two directions can be represented as $\widehat{\mathbf{E}} \in \mathbb{R}^{M\times 2d}$ and $\widetilde{\mathbf{E}} \in \mathbb{R}^{M\times 2d}$ respectively. Since the constant term can be ignored, Eq. (5) is rewritten as:

$$\begin{aligned}\delta_p^* &= \arg\max_{\delta_p\in\Delta}\sum_i \frac{1}{|\mathcal{N}_i|}\sum_{k\in\mathcal{N}_i}^{|\mathcal{N}_i|}\|[\mathbf{z}_i;\mathbf{z}_k] - [\mathbf{z}_k;\mathbf{z}_i]\|_2^2 \\ &= \arg\max_{\delta_p\in\Delta}\sum_i \frac{1}{|\mathcal{N}_i|}\sum_{k\in\mathcal{N}_i}^{|\mathcal{N}_i|}\|\mathbf{e}_{i,k} - \mathbf{e}_{k,i}\|_2^2 \\ &= \arg\max_{\delta_p\in\Delta}\|\widehat{\mathbf{E}} - \widetilde{\mathbf{E}}\|_F^2\end{aligned} \quad (6)$$

Through the above deduction, we find that the local neighborhood adversarial task is equivalent to the *edge self-adversarial* problem, that is, maximizing discrepancies between bidirectional edges. Next, we transform the above adversarial objective optimization into edge feature self-adversarial tasks.

View Generation Given an undirected graph $\mathcal{G}$, we first utilize the GNN encoder g_Φ to obtain node embeddings $\mathbf{Z} = g_\Phi(\mathcal{G})$. Then, we use a simple MLP_Ω to generate

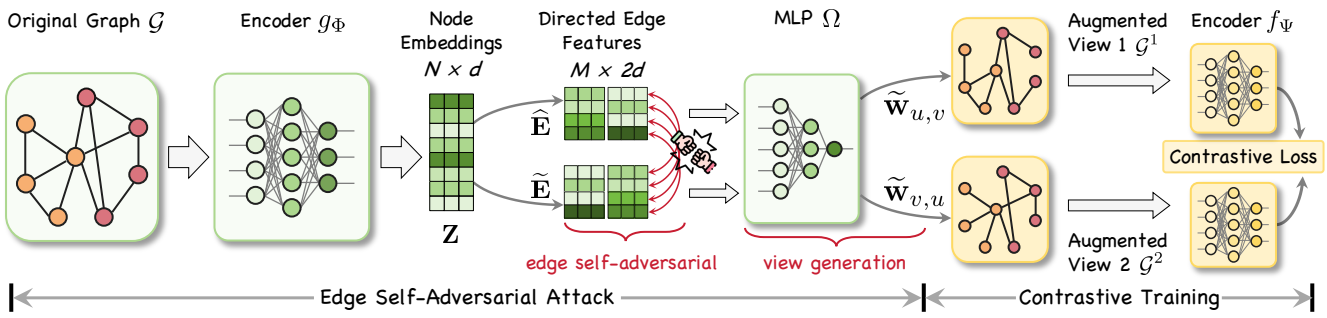

Figure 4: EDA-GCL is a Min-Max optimization problem. **Max**: The original graph $\mathcal{G}$ is fed into the encoder g_Φ to obtain node embeddings $\mathbf{Z}$. Then, two sets of edge weight vectors $\widetilde{\mathbf{w}}_{u,v}$ and $\widetilde{\mathbf{w}}_{v,u}$ are learned by maximizing the bidirectional edge feature discrepancies. **Min**: Pairwise adversarial views $\mathcal{G}^1$ and $\mathcal{G}^2$ are fed into a shared encoder f_Ψ to minimize graph contrastive loss.

the logit $w_{u,v}$ for each directed edge feature $\mathbf{e}_{u,v}$. The logit $w_{u,v}$ is used to generate soft edge weights $\widetilde{w}_{u,v}$ via Gumbel-Sigmoid relaxation (Maddison, Mnih, and Teh 2016), which allows for differentiable sampling of edge weights:

$$\widetilde{w}_{u,v} = \text{GumbelSigmod}(w_{u,v}) \tag{7}$$

where $w_{u,v} = \text{MLP}_\Omega(\mathbf{e}_{u,v})$. To ensure local neighborhood discrepancies between the two views, we apply a threshold mechanism:

$$\widetilde{w}_{u,v}, \widetilde{w}_{v,u} \leftarrow \begin{cases} (\widetilde{w}_{v,u}, 0) & \text{if } \widetilde{w}_{v,u} + \eta > \widetilde{w}_{u,v} + \eta \\ (0, \widetilde{w}_{u,v}) & \text{otherwise} \end{cases} \tag{8}$$

where $\eta \sim \mathcal{N}(0,1) \times 10^{-7}$ is a small noise added to ensure numerical stability. In the node feature aggregation phase of GNN, as edge weights only guide aggregation, they do not participate in gradient update, allowing the threshold mechanism to be applied. After thresholding, we obtain two sets of edge weight vectors $\widetilde{\mathbf{w}}_{u,v}$ and $\widetilde{\mathbf{w}}_{v,u} \in \mathbb{R}^{1\times M}$, which are used to reweight the original graph for generating two adversarial graphs $\mathcal{G}^1$ and $\mathcal{G}^2$ for graph contrastive learning.

Adversarial Objective Our overall training objective is formulated as a min-max game between the encoder f_Ψ and the adversarial view generator. Consistent with previous work, adversarial perturbations are expected to be appropriate. Therefore, we regularize the absolute difference between the directed edge weights. The specific form of the overall framework is as follows:

$$\begin{gathered} \min_{\Psi} \max_{\Phi,\Omega} \mathcal{L}_{cont}(\Psi) + \mathcal{L}_{edge}(\Phi,\Omega) - \lambda \mathcal{L}_{reg}(\Phi,\Omega), \\ \mathcal{L}_{edge} = \frac{1}{M}\|\widehat{\mathbf{E}} - \widetilde{\mathbf{E}}\|_F^2, \ \mathcal{L}_{reg} = \frac{1}{M}\|\widetilde{\mathbf{w}}_{u,v} - \widetilde{\mathbf{w}}_{v,u}\|_1 \end{gathered} \tag{9}$$

where $\mathcal{L}_{edge}$ and $\mathcal{L}_{reg}$ denote the edge self-adversarial loss and edge perturbation regularization loss, respectively. λ is the trade-off hyperparameter.

Contrastive Learning Training Based on representations obtained from different adversarial views, we train EDA-GCL using the pairwise contrastive loss provided in Eq. (1):

$$\min_\Psi \mathcal{L}_{cont}(\mathbf{Z}^1, \mathbf{Z}^2) \tag{10}$$

where $\mathbf{Z}^1 = f_\Psi(\mathcal{G}^1)$, $\mathbf{Z}^2 = f_\Psi(\mathcal{G}^2)$ denote the node embeddings of two adversarial views via the GNN encoder f_Ψ. The algorithm procedure of EDA-GCL is given in the Appendix.

Experiments

We conduct extensive experiments on various graph tasks to prove the effectiveness and robustness of EDA-GCL. We also present the superiority of EDA-GCL in computational efficiency. All dataset statistics and implementation details are provided in the Appendix.

Node Classification on Heterophilic Graphs

Datasets We evaluate the performance of EDA-GCL on five heterophilic graphs: Chameleon, Squirrel, Actor, Texas, and Wisconsin. The partition provided by Geom-GCN (Pei et al. 2020) is used, where the nodes for training/validation/testing are 48%/32%/20% of all nodes, respectively.

Competitors We compare EDA-GCL with nine baselines: DGI (Veličković et al. 2018b), MVGRL (Hassani and Khasahmadi 2020), BGRL (Thakoor et al. 2021), GRACE (Zhu et al. 2020), DGI-ADV (Xu et al. 2022), ARIEL (Feng et al. 2022), SP-AGCL (In, Yoon, and Park 2023), and GCL-JAM (Yang et al. 2025). SP-AGCL incorporates an additional k-NN augmented view. Here, we define SPAGCL-k as the ablated variant that excludes the k-NN component to further clarify the contribution from its adversarial part. We train a logistic regression classifier over learned embeddings. The mean accuracy with standard deviations is reported across ten runs for different data splits. The bold and underlined numbers denote the best and second-best performance, respectively.

Results Table 1 outlines the results. We observed that our EDA-GCL outperforms the SOTA graph contrastive learning baselines on 4 out of 5 datasets, with the best average ranking of 1.20. Compared with adversarial GCL methods (ARIEL and SP-AGCL), EDA-GCL achieves substantial robust performance gains over the baseline GRACE. Specifically, EDA-GCL achieves improvements by 12.65% (*Chameleon*), 10.53% (*Squirrel*), 6.49% (*Texas*), and 5.10% (*Wisconsin*) relative to the strong adversarial baseline SP-AGCL. When the k-NN view of SP-AGCL is ablated, the average rank dropped from 3.4 to 5.8, suggesting that its adversarial mechanism is still vulnerable to neighborhood noise. These results shed light on the robustness of our EDA-GCL against local neighborhood noises.

Method	Chameleon	Squirrel	Actor	Texas	Wisconsin	A.R.
DGI (2018)	39.95±1.75	31.80±0.77	29.82±0.69	60.59±7.56	55.41±5.96	7.80
MVGRL (2020)	42.34±2.11	33.49±0.84	31.21±0.53	61.70±3.94	50.64±5.89	7.20
BGRL (2021)	57.12±3.61	40.64±1.55	31.04±1.19	61.56±5.96	57.72±5.21	4.00
GRACE (2020)	48.05±1.81	35.23±1.03	29.49±0.51	57.57±5.68	51.50±5.83	8.80
DGI-ADV(2022)	53.37±2.21	40.13±1.60	26.50±0.89	58.41±6.08	57.29±4.88	6.60
ARIEL (2022)	49.32±2.37	36.78±1.21	29.60±0.29	58.39±4.67	53.25±7.17	7.60
SP-AGCL (2023)	57.04±2.34	40.86±1.68	31.27±1.30	62.16±8.11	55.88±5.63	3.40
SP-AGCL-k (2023)	52.70±3.31	40.25±1.65	31.23±1.21	58.56±5.10	51.63±6.47	5.80
GCL-JAM (2025)	66.37±2.37	49.84±0.93	31.01±0.46	65.83±3.61	**64.71±4.08**	2.60
EDA-GCL (Ours)	**69.69±1.83**	**51.39±0.68**	**31.33±0.90**	**68.65±3.86**	60.98±4.48	**1.20**

Table 1: Node classification accuracy on five heterophilous graphs. A.R. stands for the average rank.

Method	Cora	CiteSeer	PubMed	Photo	Computers	CS	Physics	A.R.
DGI (2018)	82.57±0.53	71.62±0.58	84.75±0.33	91.61±0.22	83.95±0.47	92.15±0.63	94.51±0.52	9.29
MVGRL (2020)	83.25±0.38	72.15±0.55	84.37±0.18	91.74±0.07	87.52±0.11	92.11±0.12	95.33±0.03	7.71
BGRL (2021)	82.69±1.03	73.43±1.04	85.09±0.23	92.08±0.42	88.05±0.39	92.36±0.20	95.21±0.08	6.57
GRACE (2020)	84.06±0.41	74.43±1.55	85.50±0.20	92.52±0.24	86.52±0.42	92.03±0.10	95.26±0.22	6.43
ARIEL (2022)	83.70±0.83	74.56±2.07	84.83±0.13	92.81±0.22	86.90±0.51	92.32±0.19	95.04±0.09	6.86
SP-AGCL (2023)	84.84±0.64	74.66±1.03	85.06±0.16	93.00±0.55	88.56±0.38	**93.55±0.13**	95.70±0.15	2.86
SP-AGCL-k (2023)	84.31±0.83	73.42±0.95	84.90±0.21	91.47±0.52	86.99±0.19	92.54±0.11	95.38±0.17	6.29
GRAPE (2024)	85.06±0.23	74.57±0.67	85.13±0.10	**93.32±0.00**	88.42±0.10	92.78±0.00	95.37±0.00	3.00
E2Neg (2025)	84.61±0.11	74.69±0.59	85.30±0.10	92.92±0.59	88.21±1.18	92.45±0.73	95.29±0.33	4.29
EDA-GCL (Ours)	**85.63±0.70**	**75.71±0.88**	**86.50±0.39**	93.08±0.46	**89.37±0.23**	92.96±0.09	**95.70±0.09**	**1.29**

Table 2: Node classification accuracy on seven homophilous graphs.

Method	Time Cost (s) per Epoch				
	Cora	Photo	Computers	PubMed	CS
ARIEL	0.1638	2.2089	9.6755	28.0601	22.6890
SP-AGCL	0.2236	2.0973	8.8529	22.2634	18.1842
EDA-GCL	**0.0613**	**0.5542**	**1.0882**	**0.5489**	**0.6144**
Ratio	3.65×	3.99×	8.89×	51.12×	36.93×

Table 3: Time cost (s) analyses on five datasets.

Node Classification on Homophilous Graphs

Datasets We evaluate performance across seven homophilous graphs: 3 citation networks (Zügner, Akbarnejad, and Günnemann 2018): Cora, CiteSeer, and PubMed, considering the largest connected component, 2 co-authorship networks: CS and Physics (Shchur et al. 2018), and 2 Amazon co-purchase networks: Computers and Photo (Shchur et al. 2018). For each dataset, we randomly select 10%/10%/80% nodes for training/validation/testing.

Competitors We select nine baselines: DGI (Veličković et al. 2018b), MVGRL (Hassani and Khasahmadi 2020), BGRL (Thakoor et al. 2021), GRACE (Zhu et al. 2020), ARIEL (Feng et al. 2022), SP-AGCL (In, Yoon, and Park 2023), GRAPE (Hao et al. 2024), E2Neg (Huang et al. 2025). For fair evaluation, we report average results with standard deviations across ten runs for different data splits.

Results The results across seven datasets are reported in Table 2. EDA-GCL still achieves best performance on 5 out of 7 datasets, and the optimal average ranking on all datasets. Specifically, EDA-GCL obtains the improvement by 0.57% (*Cora*), 1.02% (*CiteSeer*), 1.00% (*PubMed*), and 0.81% (*Computers*) over those second-best competitors. This shows that EDA-GCL also achieves compelling results for weak noise scenarios such as homophilous graphs. Furthermore, unlike previous adversarial methods, whose adversarial attacks are imposed on the entire graph, resulting in the complexity $\mathcal{O}(N^2)$, we utilize edge self-adversarial attacks to alleviate expensive costs, dropping to $\mathcal{O}(Md)$. To show the efficiency of EDA-GCL, we report the time costs of the attack process with ARIEL and SP-AGCL on the full Cora, Photo, Computers, PubMed, and Physics. The time costs are reported in Table 3. As shown, under the same settings, the running time of EDA-GCL is about 3.65× to 51.12× faster than the adversarial methods ARIEL and SP-AGCL, highlighting its superior computational efficiency.

Robustness Evaluation

We further evaluate the robustness under two adversarial settings: metattack (Zügner and Günnemann 2019) and random attack. Each attack type falls into one of two categories: poisoning attacks or evasion attacks. In poisoning attacks, the clean graph is perturbed prior to the model training phase, whereas evasion attacks introduce perturbations exclusively during the testing phase.

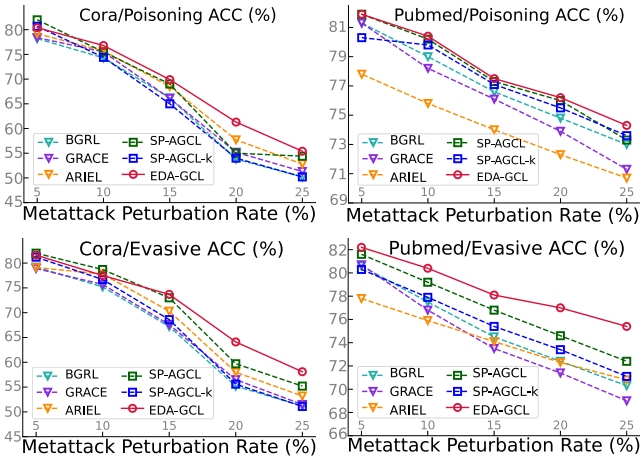

Figure 5: Node classification accuracy under Metattack.

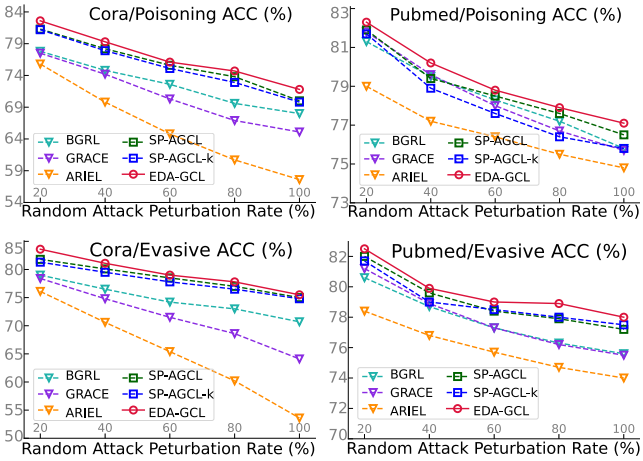

Figure 6: Node classification accuracy under random attack.

Against Metattack We utilize the public Metattack graph datasets (Jin et al. 2020) for Cora and PubMed datasets. The results are reported in Fig. 5. We have two observations: (1) Our method has demonstrated robustness under poisoning and evasion attacks, even under high perturbation rates. (2) Compared to the adversarial baselines, EDA-GCL achieves stable performance improvement. These results suggest the importance of learning local neighborhood noises.

Against Random Attack We further test the robustness of EDA-GCL under random attack. Following the protocol from (Jin et al. 2020), we inject random edge noise at intensities ranging from 20% to 100% of the original edge count, in increments of 20%. As shown in Fig. 6, our EDA-GCL still obtains outstanding robustness over adversarial baselines. Combined with the results of Fig. 5, our new adversarial mechanism is demonstrated to be robust against noise in various perturbation scenarios.

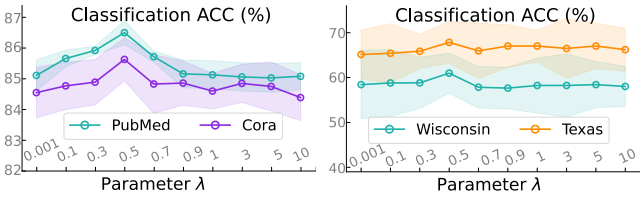

Figure 7: Sensitivity analysis with respect to parameter λ.

Method	Cora	CiteSeer	PubMed	Chameleon	Texas
w/o *Threshold*	$83.77_{\pm 0.14}$	$74.08_{\pm 1.01}$	$84.49_{\pm 0.05}$	$68.93_{\pm 2.10}$	$62.16_{\pm 4.77}$
	↓ 1.86%	↓ 1.63%	↓ 2.01%	↓ 0.76%	↓ 6.49%
w/o $\mathcal{L}_{reg}$	$84.86_{\pm 0.70}$	$72.99_{\pm 0.37}$	$85.09_{\pm 0.47}$	$69.43_{\pm 2.43}$	$66.22_{\pm 7.88}$
	↓ 0.77%	↓ 2.72%	↓ 1.41%	↓ 0.26%	↓ 2.43%
w/o $\mathcal{L}_{reg}$	$83.98_{\pm 0.47}$	$73.79_{\pm 1.08}$	$84.46_{\pm 1.08}$	$69.01_{\pm 2.19}$	$62.43_{\pm 5.62}$
& *Threshold*	↓ 1.65%	↓ 1.92%	↓ 2.04%	↓ 0.68%	↓ 6.22%
EDA-GCL	**85.63**$_{\pm 0.70}$	**75.71**$_{\pm 0.88}$	**86.50**$_{\pm 0.39}$	**69.69**$_{\pm 1.83}$	**68.65**$_{\pm 3.86}$

Table 4: Ablation study of EDA-GCL.

Sensitivity Analysis & Ablation Studies

Sensitivity Analysis We evaluate EDA-GCL under various choices of the expected perturbation hyperparameter λ, as shown in Fig. 7. We observe that too small or too large values (corresponding to too strong or too weak perturbations, respectively) lead to suboptimal performance. When the value of λ is about 0.5, EDA-GCL can be improved obviously on both heterophilous and homophilous graphs. Overall, within a reasonable range, our EDA-GCL is stable.

Ablation Studies In Table 4, we conduct an investigation into the contribution of each part. Our framework can remove two parts, namely the threshold mechanism in Eq. (8) and the regularization term in Eq. (9). When the threshold mechanism is ablated, the performance of all five datasets degrades, which explains the importance of reinforcing local neighborhood discrepancies. Besides, when the regularization term is removed, the performance is also reduced, indicating that regularization is essential for stable adversarial training. When both are removed, there is also a significant performance degradation. In summary, our EDA-GCL is a simple yet effective adversarial technique.

Conclusion

In this paper, we propose a novel edge self-adversarial attack paradigm, EDA-GCL, for robust graph contrastive learning. By theoretically connecting intra-view neighborhood inconsistency to bidirectional edge feature discrepancies, our EDA-GCL generates pairwise adversarial views through edge-level perturbations to enhance local neighborhood noise robustness. The resulting adversarial mechanism achieves superior performance and scalability with linear complexity, demonstrating significant effectiveness and resilience across various graph tasks and noise scenarios.

Acknowledgments

We thank the anonymous reviewers for their valuable and constructive comments. This work was supported partially by the National Natural Science Foundation of China under Grants 62176184 and 62473288, the Fundamental Research Funds for the Central Universities, and the Science and Technology Development Fund, Macao SAR No. 0006/2024/RIA1.

References

Chen, H.; Zhao, Z.; Li, Y.; Zou, Y.; Li, R.; and Zhang, R. 2023. CSGCL: community-strength-enhanced graph contrastive learning. In *International Joint Conference on Artificial Intelligence*.

Chen, J.; Lei, R.; and Wei, Z. 2024. PolyGCL: Graph contrastive learning via learnable spectral polynomial filters. In *International Conference on Learning Representations*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*.

Feng, S.; Jing, B.; Zhu, Y.; and Tong, H. 2022. Adversarial graph contrastive learning with information regularization. In *Proceedings of the Web Conference*.

Feng, S.; Jing, B.; Zhu, Y.; and Tong, H. 2024. Ariel: Adversarial graph contrastive learning. *ACM Transactions on Knowledge Discovery from Data*, 18(4): 1–22.

Fu, Y.; Zhang, Y.; Chen, C.; JianwenMa, J.; Yuan, Q.; Tu, R.-C.; Huang, X.; Ye, W.; Luo, X.; and Deng, M. 2025. Mark: Multi-agent collaboration with ranking guidance for text-attributed graph clustering. In *Findings of the Association for Computational Linguistics: ACL 2025*, 6057–6072.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In *Annual Conference on Neural Information Processing Systems*.

Hao, Z.; Xin, H.; Wei, L.; Tang, L.; Wang, R.; and Nie, F. 2024. Towards expansive and adaptive hard negative mining: Graph contrastive learning via subspace preserving. In *Proceedings of the Web Conference*, 322–333.

Hassani, K.; and Khasahmadi, A. H. 2020. Contrastive multi-view representation learning on graphs. In *International Conference on Machine Learning*.

He, D.; Zhao, J.; Huo, C.; Huang, Y.; Huang, Y.; and Feng, Z. 2024. A new mechanism for eliminating implicit conflict in graph contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 12340–12348.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

Huang, Y.; Zhao, J.; He, D.; Jin, D.; Huang, Y.; and Wang, Z. 2025. Does GCL Need a Large Number of Negative Samples? Enhancing Graph Contrastive Learning with Effective and Efficient Negative Sampling. *arXiv preprint arXiv:2503.17908*.

In, Y.; Yoon, K.; and Park, C. 2023. Similarity preserving adversarial graph contrastive learning. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Jin, W.; Ma, Y.; Liu, X.; Tang, X.; Wang, S.; and Tang, J. 2020. Graph structure learning for robust graph neural networks. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Jing, L.; Vincent, P.; LeCun, Y.; and Tian, Y. 2021. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*.

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

Li, W.-Z.; Wang, C.-D.; Xiong, H.; and Lai, J.-H. 2023. Homogcl: Rethinking homophily in graph contrastive learning. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, 1341–1352.

Liu, Y.; Jin, M.; Pan, S.; Zhou, C.; Zheng, Y.; Xia, F.; and Philip, S. Y. 2022. Graph self-supervised learning: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(6): 5879–5900.

Maddison, C. J.; Mnih, A.; and Teh, Y. W. 2016. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*.

Ning, Z.; Wang, P.; Qiao, Z.; Wang, P.; and Zhou, Y. 2025. Rethinking graph contrastive learning through relative similarity preservation. *arXiv preprint arXiv:2505.05533*.

Pei, H.; Wei, B.; Chang, K. C.-C.; Lei, Y.; and Yang, B. 2020. Geom-gcn: Geometric graph convolutional networks. *arXiv preprint arXiv:2002.05287*.

Ren, T.; Zhang, H.; Wang, Y.; Ju, W.; Liu, C.; Meng, F.; Yi, S.; and Luo, X. 2025. MHGC: Multi-scale hard sample mining for contrastive deep graph clustering. *Information Processing & Management*, 62(4): 104084.

Shchur, O.; Mumme, M.; Bojchevski, A.; and Günnemann, S. 2018. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*.

Suresh, S.; Li, P.; Hao, C.; and Neville, J. 2021. Adversarial graph augmentation to improve graph contrastive learning. In *Annual Conference on Neural Information Processing Systems*.

Thakoor, S.; Tallec, C.; Azar, M. G.; Munos, R.; Veličković, P.; and Valko, M. 2021. Bootstrapped representation learning on graphs. In *International Conference on Learning Representations Workshop*.

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

Veličković, P.; Fedus, W.; Hamilton, W. L.; Liò, P.; Bengio, Y.; and Hjelm, R. D. 2018b. Deep Graph Infomax. In *International Conference on Learning Representations*.

Wang, T.; and Isola, P. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, 9929–9939.

Wu, L.; Lin, H.; Tan, C.; Gao, Z.; and Li, S. Z. 2023. Self-supervised learning on graphs: Contrastive, generative, or predictive. *IEEE Transactions on Knowledge and Data Engineering*, 35(4): 4216–4235.

Xia, J.; Wu, L.; Chen, J.; Hu, B.; and Li, S. Z. 2022. Simgrace: A simple framework for graph contrastive learning without data augmentation. In *Proceedings of the Web Conference*, 1070–1079.

Xiao, T.; Zhu, H.; Chen, Z.; and Wang, S. 2023. Simple and asymmetric graph contrastive learning without augmentations. In *Advances in Neural Information Processing Systems*.

Xie, Y.; Xu, Z.; Zhang, J.; Wang, Z.; and Ji, S. 2022. Self-supervised learning of graph neural networks: A unified review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 2412–2429.

Xu, J.; Yang, Y.; Chen, J.; Jiang, X.; Wang, C.; Lu, J.; and Sun, Y. 2022. Unsupervised adversarially robust representation learning on graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 4290–4298.

Xu, K.; Chen, H.; Liu, S.; Chen, P.-Y.; Weng, T.-W.; Hong, M.; and Lin, X. 2019. Topology attack and defense for graph neural networks: an optimization perspective. In *International Joint Conference on Artificial Intelligence*.

Yang, L.; Li, Z.; Zhuo, J.; Liu, J.; Ma, Z.; Wang, C.; Wang, Z.; and Cao, X. 2025. Graph contrastive learning with joint spectral augmentation of attribute and topology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 21983–21991.

Yang, L.; Zhang, L.; and Yang, W. 2021. Graph adversarial self-supervised learning. In *Annual Conference on Neural Information Processing Systems*.

Zhang, S.; Chen, H.; Sun, X.; Li, Y.; and Xu, G. 2022. Unsupervised graph poisoning attack via contrastive loss backpropagation. In *Proceedings of the Web Conference*.

Zhang, Y.; Meng, S.; Chen, C.; Peng, M.; Gu, H.; and Huang, X. 2024. Linkthief: Combining generalized structure knowledge with node similarity for link stealing attack against gnn. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 4947–4956.

Zhao, Y.; Ji, F.; Zhao, K.; Li, X.; Kang, Q.; Liang, W.; Alkhatib, Y.; Jian, X.; and Tay, W. P. 2025. Simple Graph Contrastive Learning via Fractional-order Neural Diffusion Networks. *arXiv preprint arXiv:2504.16748*.

Zhu, W.; Peng, B.; and Chen, C. 2021. Self-supervised embedding for subspace clustering. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3687–3691.

Zhu, Y.; Xu, Y.; Yu, F.; Liu, Q.; and Wu, S. 2021. Graph contrastive learning with adaptive augmentation. In *Proceedings of the Web Conference*.

Zhu, Y.; Xu, Y.; Yu, F.; Liu, Q.; Wu, S.; and Wang, L. 2020. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131*.

Zügner, D.; Akbarnejad, A.; and Günnemann, S. 2018. Adversarial attacks on neural networks for graph data. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Zügner, D.; and Günnemann, S. 2019. Adversarial Attacks on Graph Neural Networks via Meta Learning. *arXiv preprint arXiv:1902.08412*.