CH-GCCIL: CLIQUE-HIERARCHICAL SPECTRAL INTERVENTION FOR GRAPH CAUSAL CONTRASTIVE INVARIANT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph contrastive learning has emerged as a dominant framework for graph representation learning by learning discriminative node representations via graph augmentation. Nevertheless, conventional stochastic augmentation schemes fail to differentiate causal factors from non-causal confounding factors, making the learned representations prone to capturing spurious correlational patterns rather than invariant causal structures. Current causal-invariant graph learning methods primarily rely on global spectral decomposition, which imposes a uniform frequency partition of causal and non-causal components on the entire graph. Such global paradigms overlook the inherent structural heterogeneity of graph cliques: nodes inside individual cliques exhibit consistent causal mechanisms, while cross-clique connections are dominated by non-causal noisy associations, thereby resulting in coarse and inaccurate causal separation. To address this limitation, this paper proposes a novel Graph Causal Contrastive Invariant Learning method with Clique-Hierarchical Spectral Intervention (CH-GCCIL) to achieve fine-grained causal disentanglement from a clique hierarchical perspective. Specifically, we first mine maximal cliques from the graph and categorize them into strongly causal cliques and noisy cliques based on a three-dimensional confidence metric. On this basis, we construct asymmetric dual augmented views and adopt Chebyshev approximation for efficient spectral filtering. Furthermore, we design a dual-scale contrastive loss function to optimize representation learning at both global graph and intra-clique levels. Extensive node classification experiments on five benchmark datasets validate the superiority of CH-GCCIL. It achieves Macro-F1 scores of 78.8%, 84.1%, and 44.1% on Wiki-CS, PubMed, and Flickr datasets, outperforming state-of-the-art baseline methods by 3.2%, 2.6%, and 4.1% respectively, which demonstrates the effectiveness and advancement of the proposed clique-hierarchical spectral intervention mechanism.

Keywords: Graph contrastive learning; Invariant learning; Causal inference; Spectral intervention; Clique mining

1 INTRODUCTION

Graph Neural Networks (GNNs) have made significant progress in tasks such as node classification, link prediction, and graph classification (Kipf & Welling, 2017; Veličković et al., 2018), but their training relies on a large amount of labeled data, and manual labeling is costly. To alleviate the problem of scarce labels, Graph Contrastive Learning (GCL) learns node representations by constructing augmented views (Veličković et al., 2019; Zhu et al., 2020). Its core paradigm is: random augmentation of the input graph to obtain two views, and the encoder maximizes the consistency of the representation of the same node in different views (Chen et al., 2020; Zhu et al., 2021). However, traditional random augmentation strategies are "blind" and cannot ensure that augmentation only perturbs non-causal information while preserving causal information (Arjovsky et al., 2019; Mitrović et al., 2021). From the perspective of Structural Causal Model (SCM) (Pearl, 2009; Pearl et al., 2016), node labels are determined by causal factors (such as the research topic of a paper), while non-causal factors (such as social connections in citation networks and author collaborations)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

constitute confounding variables. Random augmentation destroys causal information and preserves noise, leading to the learning of false associations in the representation (Wu et al., 2022; Fan et al., 2022).

GCIL (Graph Contrastive Invariant Learning from the Causal Perspective) (Mo et al., 2024) introduces causal intervention into graph contrastive learning, simulates intervention on non-causal factors through spectral filtering, and learns causal representations that are invariant across views. GCIL assumes that the low-frequency components of the graph signal correspond to causal information (smoothed signal) and the high-frequency components correspond to non-causal factors (noise), and achieves causal intervention by perturbing the high-frequency subspace. Despite its innovativeness, GCIL has three key limitations: it assumes that the entire graph shares a uniform causal/non-causal frequency partition, ignoring the heterogeneity of clique structures; its high computational complexity depends on the complete Laplacian eigenvalue decomposition, making it difficult to extend to large-scale graphs (Defferrard et al., 2016; Bruna et al., 2014); and the global loss constraint requires that both the invariant loss and the independent loss be calculated in the global dimension. (Zbontar et al., 2021; Mialon et al., 2022).

To address the above issues, this paper proposes CH-GCCIL (Clique-Hierarchical Spectral Intervention for Graph Causal Contrastive Invariant Learning), which comprises three core components: clique hierarchical construction, asymmetric hierarchical spectral intervention, and dual-scale causal constraints. First, maximal cliques are mined from the input graph, and the cliques are divided into strong causal cliques and noise cliques based on three-dimensional confidence. Then, high-frequency perturbations are applied to strong causal cliques to protect the low-frequency causal structure, and large perturbations are applied to noise cliques across the entire frequency range to destroy spurious associations. Finally, causal constraints are applied at both the global and intra-clique scales using hierarchical invariant loss and hierarchical independent loss. The main contributions of this paper are summarized as follows:

- **Introducing maximal cliques into graph causal invariant learning.** We introduce the maximal clique structure in graphs into causal contrastive learning, proposing a clique hierarchy from a causal perspective—nodes within a clique share similar causal mechanisms, while connections between cliques reflect more non-causal factors, providing a new structural prior for finely classifying causality.
- **Proposing the CH-GCCIL Framework.** We design three core components: clique hierarchy divides cliques into strongly causal cliques and noisy cliques; asymmetric hierarchical spectral intervention protects causal structures from non-causal structures; perturbs non-causal structures and dual-scale causal constraints are used to learn node representations.
- **Experiments validate the outstanding performance of the method.** On five benchmark datasets, CH-GCCIL outperforms existing baseline methods on multiple datasets. Ablation experiments confirm that all three dimensions of the three-dimensional confidence are indispensable. Chebyshev order analysis determines that $K = 3$ is the optimal choice, achieving a good balance between performance and efficiency.The implementation code is publicly available on : `https://anonymous.4open.science/r/CH-GCCIL_code`.

2 Related Work

Graph contrastive learning. GCL, as the mainstream paradigm of graph representation learning, learns node representations by constructing augmented views. Early methods such as DGI (Deep Graph Infomax) (Veličković et al., 2019) learn by maximizing the mutual information between node representations and graph-level representations. GRACE (Zhu et al., 2020) and GCA (Zhu et al., 2021) introduce random augmentation strategies such as edge deletion, feature masking, and node discarding, and maximize the consistency of the representation of the same node under different views through InfoNCE loss. MVGRL (Hassani & Khasahmadi, 2020) constructs multiple views using diffusion matrix and adjacency matrix. BGRL (Thakoor et al., 2022) adopts a bootstrapping method and does not require negative samples. In addition, GraphCL (You et al., 2020), CCA-SSG (Zhang et al., 2021), ProGCL (Xia et al., 2022), COSTA (Zhang et al., 2022) and other methods further improve the augmentation strategies and loss functions. These methods rely on random augmentation strategies and lack causal interpretability (Arjovsky et al., 2019; Mitrović et al., 2021). In

recent years, provably trainable graph contrastive learning methods (Yu et al., 2023) and contrastive learning based on condensed subgraph perception (Wu et al., 2024b) have further improved representation quality, while adaptive subgraph contrastive learning (Wang et al., 2025) has effectively alleviated the topological bias problem.

Causal inference for graph learning. Causal inference provides a theoretical framework for solving the problem of spurious associations (Pearl, 2009; Pearl et al., 2016). Invariant Risk Minimization (IRM) (Arjovsky et al., 2019) learns representations that are invariant across environments and has been extended to the field of graph learning. DIR-GNN (Wu et al., 2022) discovers the basis of reasoning by intervening in the non-causal part. DisC (Fan et al., 2022) decomposes the graph into causal subgraphs and bias subgraphs to alleviate the bias in the dataset. RGCL (Li et al., 2022) introduces invariance into self-supervised learning and preserves stable semantic information. CIGA (Chen et al., 2022) achieves OOD generalization across distribution shifts by capturing the invariance of the graph. GCIL (Mo et al., 2024) is the most relevant work, which introduces causal intervention into graph contrastive learning and simulates the intervention of non-causal factors through spectral filtering. The latest research has promoted the development of causal graph learning by achieving graph distribution out-generalization through causal intervention (Wu et al., 2024a) and by using spurious correlations to learn graph invariance representations (Yao et al., 2025).

3 METHODOLOGY

3.1 PROBLEM DEFINITION

Given an attribute graph $G = (V, E, X, Y)$, where $V = \{v_1, v_2, \ldots, v_n\}$ is the set of nodes, E is the set of edges, $X \in \mathbb{R}^{n \times d}$ is the node feature matrix, and Y is the set of node labels. An adjacency matrix $A \in \mathbb{R}^{n \times n}$ is given, where an edge exists between nodes v_i and v_j if and only if $A_{ij} = 1$. The goal is to learn a GNN encoder to obtain node representations invariant to distributional biases, such that H can effectively predict labels Y, independent of non-causal factors S. Two asymmetric views are constructed: the original graph $G_1 = (A, X)$ and its hierarchically spectrally intervened variant $G_2 = (\tilde{A}, X)$. A shared GNN encoder is then used to encode the two views, obtaining node representations H_1 and H_2. The model parameters are optimized by minimizing the total loss function Eq. (1):

$$\theta^* = \arg\min_{\theta} L_{\text{total}}(H_1, H_2, Y). \tag{1}$$

The explanations of the main symbols are available in the technical supplement.

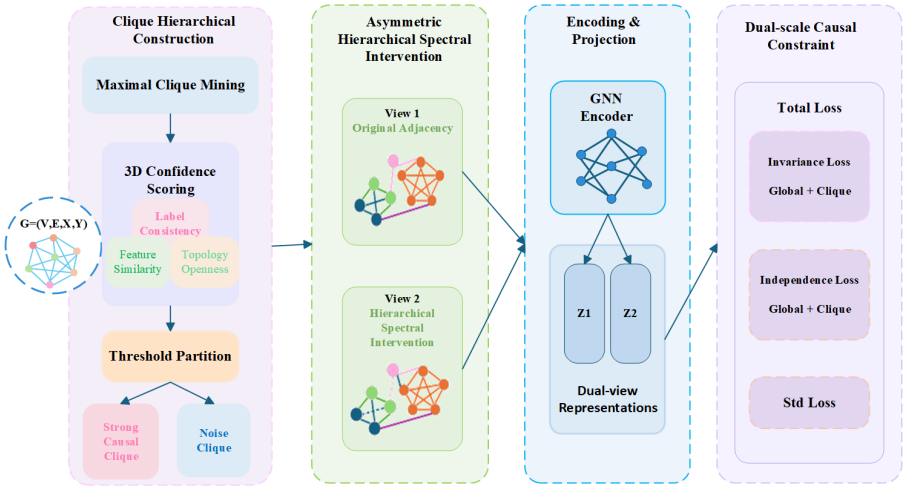

Figure 1: Overall framework of CH-GCCIL. The model consists of clique hierarchical construction, asymmetric hierarchical spectral intervention, and dual-scale causal constraints.

3.2 Overall Framework

CH-GCCIL contains three core components: (1) clique hierarchical construction; (2) asymmetric hierarchical spectral intervention; and (3) dual-scale causal constraints. Given an input graph G, a clique hierarchical structure is first constructed: maximal cliques are mined and the cliques are divided into strong causal cliques and noisy cliques based on three-dimensional confidence. Then, two views are constructed: view 1 keeps the original graph as the anchor point, and view 2 applies hierarchical spectral intervention to the adjacency matrix. The two views are encoded by a shared two-layer GNN encoder, and the resulting representations are constrained by dual-scale causal loss (invariance, independence, normalization). The overall framework is shown in Figure 1.

3.3 Clique Hierarchical Construction

3.3.1 Maximal Clique Mining

A clique is a subset of vertices in which any two vertices are connected. A clique that cannot have any more vertices added to is called a maximal clique. We use the Bron-Kerbosch algorithm to enumerate maximal cliques, and after size filtering and quantity control (if the limit is exceeded, we keep the largest half and add random sampling) for subsequent analysis.

3.3.2 Structural Causal Model

We use the structural causal model (SCM) to describe the generation process of the graph data. Each node is determined by both causal factors C and non-causal factors S, based on three assumptions: the graph can be decomposed into causal variables C and non-causal variables S; only C simultaneously affects the graph structure G and the downstream task Y, while S contains no valid information; the generation processes of C and S are independent of each other.The causal generation process for node classification is as follows: $C \rightarrow G \leftarrow S$, observed data is generated by two types of latent variables; $C \rightarrow Y$, C contains all the information needed for classification; $G \leftrightarrow Y$, not distinguishing between C and S will lead to the representation capturing spurious associations, and confounding factors, common causes between different causal variables, lead to redundant correlations.

Causal factor C corresponds to stable cross-environmental semantic signals (such as the research field of the paper), while non-causal factor S corresponds to spurious associations that change with the environment (such as the publication year of the paper). Since a true causal structure is formed when the causal factors of nodes within a cluster are consistent, and spurious associations are generated when they are inconsistent, differentiated interventions should be implemented at the cluster level: strong causal clusters retain the causal structure, while noisy clusters significantly perturb non-causal factors, such that:

$$P^{\text{do}(S=s_i)}(Y|C) = P^{\text{do}(S=s_j)}(Y|C), \tag{2}$$

Regardless of how the non-causal factors change, the conditional distribution of Y remains unchanged given C. Furthermore, to eliminate the influence of confounding factors, different causal variables must be independent of each other:

$$\forall c_i, c_j \in \mathcal{C}, \quad c_i \perp c_j. \tag{3}$$

3.3.3 Three-Dimensional Clique Confidence Scoring

For each discovered clique, a three-dimensional confidence score is calculated to measure the degree to which the clique represents a strong causal structure. The label consistency dimension in the three-dimensional confidence score is calculated using only the labels of the training set nodes; the labels of the test set nodes are not involved in the clique partitioning process, thus avoiding data leakage. Label information is only used to assist in the construction of clique hierarchies and does not directly participate in the contrastive learning training of the encoder.

Label Consistency. This measures the purity of labels within the clique, defined as the proportion of the majority label Eq. (4):

$$S_{\text{label}}(c) = \frac{\max_k |\{i \in c : y_i = k\}|}{|c|}. \tag{4}$$

The more concentrated the labels are, the higher the proportion of most labels, and the stronger the causal relationship, as nodes within the group share the same causal mechanism.

Feature similarity. The mean cosine similarity measures the features of nodes within a clique Eq. (5):

$$S_{\text{feat}}(c) = \frac{2}{|c|(|c|-1)} \sum_{i<j \in c} \cos(x_i, x_j). \tag{5}$$

The more similar the features, the more causal factors the nodes share in the feature space.

Topological openness. A measure of the closure of a clique Eq. (6):

$$S_{\text{topo}}(c) = 1 - \frac{\text{external_edges}}{|c| \cdot (n - |c|)}. \tag{6}$$

A more closed clique indicates fewer connections to the outside, lower topological openness, and higher confidence. Closed cliques are less affected by external non-causal factors.

Feature similarity and topological openness are both min-max normalized to $[0, 1]$, while label consistency itself falls within the range of $[0, 1]$. The overall confidence score is a three-dimensional weighted sum Eq. (7):

$$w_c = \alpha_{\text{label}} \cdot S_{\text{label}}(c) + \alpha_{\text{feat}} \cdot S_{\text{feat}}(c) + \alpha_{\text{topo}} \cdot S_{\text{topo}}(c). \tag{7}$$

The default weights $\alpha_{\text{label}} = 0.6$, $\alpha_{\text{feat}} = 0.25$, $\alpha_{\text{topo}} = 0.15$ reflect the relative importance of each dimension.

3.3.4 Clique Hierarchy Partitioning

Based on the confidence scores, the cliques are divided into two categories using the median threshold $\tau = \text{median}(w_c)$: cliques with $w_c \geq \tau$ are strong causal cliques (denoted as $\mathcal{C}_s$), and cliques with $w_c < \tau$ are noise cliques (denoted as $\mathcal{C}_n$), as given in Eq. (8):

$$\mathcal{C}_s = \{c \mid w_c \geq \tau\}, \quad \mathcal{C}_n = \{c \mid w_c < \tau\}. \tag{8}$$

As shown in Figure 2, the clique in the left figure has high label consistency and few external connections, thus the causal quality of the clique is high. In contrast, the clique in the right figure has low node label consistency and many external connections, resulting in low causal quality of the clique.

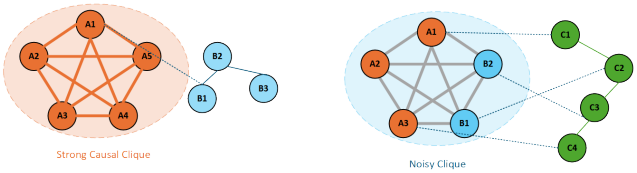

Figure 2: A conceptual comparison of strong causal groups and noisy groups: both are fully connected groups, the difference lies in label consistency and topological openness.

3.4 Asymmetric Hierarchical Spectral Intervention

Strong causal cliques represent stable causal structures in the graph and should be protected during intervention; noise cliques represent loose connections of non-causal factors and should be significantly perturbed during intervention to destroy spurious associations. The purpose of "stratification before intervention" is to apply differentiated spectral interventions to strong causal cliques and noise cliques, thereby more accurately preserving causal information and removing non-causal noise.

CH-GCCIL uses an asymmetric dual-view strategy, rather than applying spectral interventions to both views (Mo et al., 2024; Bo et al., 2023): View 1 is used as an anchor point to maintain the original causal structure, while View 2 is used as an intervention graph to perturb non-causal factors. The differences between the two views are entirely generated by the spectral intervention, allowing

the encoder to learn causal representations that are invariant to the intervention. The feature dropout rate of View 2 is halved, and since its structure has been significantly perturbed, there is no need for excessive feature perturbations to increase the view differences.

Graph Laplacian Matrix and Spectral Decomposition. The normalized graph Laplacian and its spectral decomposition provide the foundation for frequency-domain intervention. Detailed definitions of the normalized Laplacian, eigen decomposition, graph signal frequency domain, and spectral filtering are given in the technical supplement. The complete Laplace decomposition has a complexity of $O(n^3)$ and is not scalable (Defferrard et al., 2016; Bruna et al., 2014). For the adjacency matrix of the subgraph of each clique, a K-order Chebyshev polynomial is used to approximate a low-pass filter, as follows:

Scaling to $[-1, 1]$*:*

$$\hat{L} = \frac{2 \cdot L_{\text{norm}} - \lambda_{\max} \cdot I}{\lambda_{\max}}, \tag{9}$$

where $\lambda_{\max}$ is the upper bound of the spectral radius of the normalized Laplace, and $\hat{L}$ is clipped to $[-1, 1]$ to ensure numerical stability.

Chebyshev polynomial recurrence:

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x). \tag{10}$$

Low-pass filter coefficients:

$$c_0 = 1.0, \quad c_k = 2.0 \cdot \theta^k \quad (k \geq 1), \tag{11}$$

where θ is the low-pass cutoff frequency parameter.

Low-pass projection matrix:

$$H_{\text{low}} = \left(\sum_{k=0}^{K} c_k \cdot T_k(\hat{L}) \right) \cdot \text{scale}, \tag{12}$$

where scale $= 0.1$ is a scaling factor to prevent numerical overflow. The high-pass projection matrix is $H_{\text{high}} = I - H_{\text{low}}$.

The Chebyshev approximation has a complexity of $O(K \cdot |E_c|)$ on each subgraph, where $K = 3$, which is much lower than the $O(n^3)$ complexity of the complete eigenvalue decomposition, making the method scalable to large-scale graphs.

Based on the clique hierarchical results, differential spectral interventions are performed on the adjacency matrix of View 2 for each clique, and then cross-clique edges are processed globally. Minimal interventions are applied to strongly causal cliques to preserve their causal structure, while heavy interventions are applied to noisy cliques to destroy their spurious associations. This creates distributional differences between the two views caused by changes in non-causal factors, forcing the encoder to learn causal representations invariant to the interventions.

(a) Intervention of strong causal cliques: For the subgraph adjacency matrix of a strong causal clique, we first compute the Chebyshev low-pass and high-pass projections. Then we generate symmetric Gaussian noise $N \sim \mathcal{N}(0, \epsilon^2_{\text{strong}})$, where perturb_rate $=$ strong_perturb_rate $= 0.1$, injecting noise only into the high-frequency subspace Eq. (13):

$$\tilde{A}_c = A_c + H_{\text{high}} \cdot N \cdot H^T_{\text{high}}. \tag{13}$$

The results are nonnegated (trimmed to ≥ 0) and self-loops are removed. Low-frequency components (causal structure) are protected from noise contamination by H_{low}, while only high-frequency components (noise/mutations) are perturbed, thus preserving the causal structure of strong causal cliques.

(b) Intervention of noise cliques: For the adjacency matrix A_c of the subgraph of noise clique c, generate symmetric Gaussian noise $N \sim \mathcal{N}(0, \epsilon^2_{\text{noise}})$, where perturb_rate $=$ noise_perturb_rate $=$ 0.5, and directly add noise in the full frequency domain (without frequency domain separation), as in Eq. (14):

$$\tilde{A}_c = A_c + N. \tag{14}$$

The results are nonnegated (trimmed to ≥ 0) and self-loops are removed. The connections of the noise cliques are mainly driven by non-causal factors, so large perturbations are made across the entire frequency domain to destroy spurious correlations within the cliques.

(c) Global perturbation of cross-clique edges: We identify all cross-clique edges (edges whose endpoints are not in the same clique), delete them with a probability of cross_edge_drop_rate $= 0.3$, and add them at non-edge locations with a probability of cross_edge_add_rate $= 0.05$. Cross-clique edges carry social bridging and distribution shifts between cliques and are non-causal factors. Random deletion and modification simulate environmental changes (Wu et al., 2022; Fan et al., 2022). Clique mining and confidence scoring are performed only once before training, and the clique structure is fixed during training. Every 10 epochs, the clique information is reused to perform spectral intervention again, with different random noise and cross-clique edge deletions, so that View 2 changes dynamically during training, enhancing robustness.

3.5 Dual-Scale Causal Constraints

Before computing the losses, representations are first layer-normalized. Then we design three types of losses to impose causal constraints at both the global and intra-clique scales (Mialon et al., 2022; Zhang et al., 2021). Detailed component formulas are provided in the technical supplement.

Layered invariance loss. Invariance requires that the representations of the same node in the two views be as close as possible (Veličković et al., 2019; Mo et al., 2024). A global term aligns all nodes, while a confidence-weighted clique-level term enforces stronger consistency for nodes in high-confidence cliques. The total invariance loss combines both scales Eq. (15):

$$L_{\text{inv}} = \alpha_{\text{global}} \cdot L_{\text{inv}}^{\text{global}} + (1 - \alpha_{\text{global}}) \cdot L_{\text{inv}}^{\text{clique}}. \tag{15}$$

Layered independence loss. The independence requires low correlation between representation dimensions, penalizing the sum of squares of the off-diagonal elements of the covariance matrix at both global and clique scales. The total independence loss is Eq. (16):

$$L_{\text{ind}} = \gamma_{\text{global}} \cdot L_{\text{ind}}^{\text{global}} + \gamma_{\text{clique}} \cdot L_{\text{ind}}^{\text{clique}}. \tag{16}$$

Globally independent decorrelation across all dimensions ensures that node representations within a clique do not share redundant non-causal information. This loss only penalizes off-diagonal elements (inter-dimensional correlation) and does not constrain diagonal elements (dimensional variance), which are constrained separately by the standardized loss.

Standardization loss. To prevent the representation dimensions from collapsing to a constant, the standard deviation of each dimension should be close to 1 (Mo et al., 2024; Zbontar et al., 2021):

$$L_{\text{std}} = \frac{1}{d}\sum_{k=1}^{d} \text{ReLU}(1 - \sigma(z_k^{(1)})) + \frac{1}{d}\sum_{k=1}^{d} \text{ReLU}(1 - \sigma(z_k^{(2)})), \tag{17}$$

where $\sigma(z_k)$ is the standard deviation of the k-th dimension (add ϵ to prevent division by zero). A penalty is incurred when the standard deviation is < 1, and the loss is 0 when it is ≥ 1.

Total loss. The total loss combines all three components Eq. (18):

$$L_{\text{total}} = \lambda_{\text{inv}} \cdot L_{\text{inv}} + \lambda_{\text{ind}} \cdot L_{\text{ind}} + \lambda_{\text{std}} \cdot L_{\text{std}}. \tag{18}$$

4 Experiments

4.1 Experimental Setup

CH-GCCIL was evaluated on five widely used node classification benchmark datasets: Cora, CiteSeer, PubMed (Sen et al., 2008), Wiki-CS (Mernyei & Cangea, 2020), and Flickr. Detailed statistics of the datasets are shown in the technical supplement. We selected two types of methods as baselines: (1) semi-supervised methods: GCN (Kipf & Welling, 2017), GAT (Veličković et al., 2018); (2) self-supervised methods: DGI (Veličković et al., 2019), MVGRL (Hassani & Khasahmadi, 2020), GRACE (Zhu et al., 2020), GCA (Zhu et al., 2021), GraphCL (You et al., 2020), CCA-SSG (Zhang et al., 2021), COSTA (Zhang et al., 2022), ProGCL (Xia et al., 2022), GCIL (Mo et al., 2024). The detailed configuration items and parameters are provided in the technical supplement.

4.2 Main Results

Table 1: Node classification results (Macro-F1/Micro-F1, %) on five benchmark datasets. Best results are in **bold**.

	Cora		CiteSeer		PubMed		Wiki-CS		Flickr	
Method	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi
GCN	80.6±0.7	81.5±0.6	68.1±0.5	70.9±0.5	78.5±0.3	78.9±0.3	73.2±0.8	77.5±0.4	48.9±1.6	50.2±1.2
GAT	81.3±0.3	82.3±0.2	67.5±0.2	72.0±0.9	77.4±0.2	77.8±0.2	75.5±0.4	78.3±0.4	35.0±0.8	37.1±0.3
DGI	80.4±0.7	82.0±0.5	67.7±0.9	71.7±0.8	76.8±0.9	76.7±0.9	70.6±0.1	75.6±0.1	31.2±1.6	33.0±1.6
MVGRL	81.5±0.5	82.8±0.4	66.8±0.7	72.5±0.5	79.8±0.4	79.7±0.3	74.9±0.1	78.1±0.1	31.2±2.9	33.4±3.0
GRACE	79.2±1.0	80.0±1.0	65.1±1.2	68.7±1.1	80.0±0.7	79.9±0.7	74.8±0.2	78.2±0.1	35.7±1.3	37.3±1.0
GCA	79.9±1.1	81.1±1.0	62.8±1.3	65.9±1.0	80.8±0.6	81.4±0.6	74.9±0.0	78.3±0.0	41.2±0.5	42.2±0.6
GraphCL	80.7±0.9	82.3±0.9	67.8±1.0	71.9±0.9	77.0±0.4	76.8±0.5	–	–	32.1±1.1	34.5±0.9
CCA-SSG	82.9±0.8	83.6±0.9	67.9±1.0	73.1±0.7	80.7±0.6	81.0±0.6	73.7±0.2	78.2±0.1	37.0±1.1	39.3±0.9
COSTA	81.2±0.4	82.4±0.7	62.9±3.6	66.4±4.3	79.8±0.6	80.4±0.9	72.2±5.1	76.2±2.3	42.0±2.6	42.8±2.5
ProGCL	81.0±0.4	81.9±0.6	62.9±0.9	65.9±0.9	80.5±2.3	80.5±2.3	75.3±0.1	78.7±0.1	37.8±1.1	38.5±1.0
GCIL	83.8±0.5	84.4±0.7	**69.1±0.4**	**73.7±0.5**	81.5±0.5	81.6±0.7	75.6±0.1	78.6±0.1	40.0±0.8	43.0±0.4
CH-GCCIL	**83.8±0.7**	**85.2±0.6**	66.8±0.5	71.8±0.5	**84.1±0.2**	**84.5±0.2**	**78.8±0.2**	**81.3±0.2**	**44.1±0.3**	**46.1±0.4**

Table 1 shows that CH-GCCIL achieves good performance on multiple datasets. On PubMed, Wiki-CS, Flickr, and Cora datasets, CH-GCCIL outperforms baseline methods. However, due to the low quality of clique structure in CiteSeer (high proportion of triangular cliques, small average clique size, and low intra-clique label consistency), the 3D confidence score lacks discriminative power, making it difficult for clique-leveling strategies to effectively separate strong causal cliques from noisy cliques. Furthermore, the high sparsity of CiteSeer graphs (average degree only 2.74), the small number and size of cliques (Figure 3) limit the effectiveness of clique-leveling intervention. The clique-leveling strategy shows significant advantages on datasets with clear clique structures, but its performance is lower than global intervention on sparse graphs with low clique structure quality. Future research could consider using an adaptive backoff mechanism—automatically reverting to global spectral intervention when clique structure quality is low.

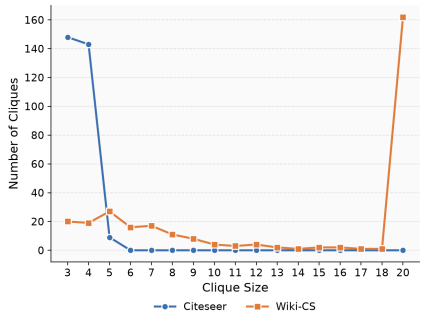

Figure 3: Clique size distribution comparison: CiteSeer (sparse, small cliques) vs. Wiki-CS (dense, larger cliques).

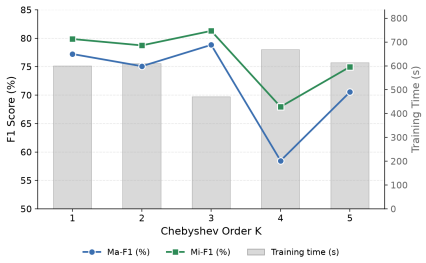

Figure 4: Chebyshev order sensitivity: performance and training time across $K \in \{1, 2, 3, 4, 5\}$.

4.3 Ablation Studies

4.3.1 Three-Dimensional Confidence Ablation

To verify the effectiveness of the three-dimensional confidence score, an ablation experiment was conducted on Wiki-CS, where each dimension was removed one by one and compared with the complete model and randomized partitions. The results are shown in Table 2. To verify the superiority of 3D confidence, a single-dimensional ablation experiment was also conducted on Wiki-CS, and the results are shown in Table 3.

Removing any dimension leads to a performance degradation, with significant differences in impact across dimensions: topological openness causes the largest performance drop (Ma-F1 score decrease

Table 2: 3D confidence ablation (Wiki-CS).

Models	Ma-F1	Mi-F1
CH-GCCIL	**78.8**	**81.3**
w/o label	77.1	80.0
w/o feature	65.3	71.9
w/o topology	50.2	61.4
Random	42.0	56.6

Table 3: 1D clique partitioning ablation (Wiki-CS).

Model	Ma-F1	Mi-F1
CH-GCCIL	**78.8**	**81.3**
Only topology	75.0	78.2
Only label	62.8	69.5
Only feature	57.8	67.8

Table 4: Confidence weight sensitivity (Wiki-CS).

Weights	Ma-F1	Mi-F1
(0.6,0.25,0.15)	**78.8**	**81.3**
Equal (1/3)	69.5	74.7

of 28.6%), followed by feature similarity, while label consistency has the smallest impact. On Wiki-CS, topological closure has the highest discriminative power for causal groups. Random grouping results in the worst performance, indicating that confidence-based group hierarchies are necessary.

Single-dimensional ablation experiments demonstrate the superiority of three-dimensional confidence scores. The three-dimensional synthesis method outperforms the optimal single-dimensional method by 3.8%, demonstrating the effectiveness of combining multiple dimensions. Among the dimensions, topological openness performs best, achieving a Ma-F1 score of 75.0% using only this dimension, indicating that the topological closure of the clique itself is a strong causal prior. Combining the two types of ablation experiments yields a consistent conclusion: all three dimensions are indispensable.

4.3.2 Weighted Sensitivity and Chebyshev Order Analysis

The sensitivity of confidence weights was analyzed on Wiki-CS, comparing the default weights $(0.6, 0.25, 0.15)$ with equal weights $(1/3, 1/3, 1/3)$ (Table 4). The impact of the Chebyshev polynomial order $K \in \{1, 2, 3, 4, 5\}$ on performance and efficiency was analyzed on Wiki-CS, as shown in Figure 4.

The default weights $(0.6, 0.25, 0.15)$ are significantly better than the equal weights $(1/3, 1/3, 1/3)$. This result indicates that reasonably allocating the three-dimensional weights in the cluster confidence score can improve the cluster stratification quality, with the configuration having higher label consistency weights showing the best effect.

Chebyshev order sensitivity analysis shows that $K = 3$ achieves the best performance and shortest training time, being optimal in terms of effectiveness, efficiency, and numerical stability; $K = 4$ shows a performance decline, and higher-order polynomials exhibit numerical instability; $K = 1$ offers reasonable performance but requires longer training time, as the first-order approximation has weaker expressive power and requires more epochs to converge. Third-order polynomials are sufficient to capture local topological structures.

5 Conclusion

This paper proposes CH-GCCIL, which achieves refined causal separation through clique hierarchical intervention. Inspired by the heterogeneity of clique structures in real-world graphs—where nodes within a clique share similar causal mechanisms and inter-clique connections reflect more non-causal factors—CH-GCCIL comprises three core components: clique hierarchical construction, asymmetric hierarchical intervention, and dual-scale causal constraints.

Experiments on five benchmark datasets validated the effectiveness of CH-GCCIL, particularly demonstrating significant advantages on datasets with prominent clique structures. Ablation experiments confirmed that all three dimensions of confidence are indispensable, with topological openness being the most discriminative dimension. Chebyshev order analysis determined that $K = 3$ is the optimal choice, achieving a good balance between performance and efficiency. Future work will explore adaptive backoff mechanisms to automatically revert to global spectrum intervention when the quality of the cluster structure is low, and extend this to more complex scenarios such as heterogeneous graphs and dynamic graphs.

REFERENCES

Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Deyu Bo, Xiao Wang, Yang Liu, Yuan Fang, Yawen Li, and Chuan Shi. A survey on spectral graph neural networks. *arXiv preprint arXiv:2302.05631*, 2023.

Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. In *International Conference on Learning Representations (ICLR)*, 2014.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, volume 119, pp. 1597–1607. PMLR, 2020.

Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, Kaili Ma, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. Learning causally invariant representations for out-of-distribution generalization on graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3844–3852, 2016.

Shaohua Fan, Xiang Wang, Yixin Mo, Chuan Shi, and Jie Tang. Debiasing graph neural networks via learning disentangled causal substructure. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 24934–24946, 2022.

Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In *International Conference on Machine Learning (ICML)*, volume 119, pp. 4116–4126. PMLR, 2020.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Siyuan Li, Xiang Wang, An Zhang, Yingxin Wu, Xiangnan He, and Tat-Seng Chua. Let invariant rationale discovery inspire graph contrastive learning. In *International Conference on Machine Learning (ICML)*, volume 162, pp. 13052–13065. PMLR, 2022.

Péter Mernyei and Cătălina Cangea. Wiki-CS: A Wikipedia-based benchmark for graph neural networks. *arXiv preprint arXiv:2007.02901*, 2020.

Grégoire Mialon, Randall Balestriero, and Yann LeCun. Variance covariance regularization enforces pairwise independence in self-supervised representations. *arXiv preprint arXiv:2209.14905*, 2022.

Jovana Mitrović, Brian McWilliams, Jacob C Walker, Lars Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. In *International Conference on Learning Representations (ICLR)*, 2021.

Yixin Mo, Xiang Wang, Shaohua Fan, and Chuan Shi. Graph contrastive invariant learning from the causal perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 8904–8912, 2024. doi: 10.1609/aaai.v38i8.28738.

Judea Pearl. *Causality: Models, reasoning, and inference*. Cambridge University Press, Cambridge, 2 edition, 2009.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. Wiley, Chichester, 2016.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008.

Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, et al. Large-scale representation learning on graphs via bootstrapping. In *International Conference on Learning Representations (ICLR)*, 2022.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. In *International Conference on Learning Representations (ICLR)*, 2019.

Yili Wang, Yaohua Liu, Ninghao Liu, Rui Miao, Ying Wang, and Xin Wang. AdaGCL+: An adaptive subgraph contrastive learning toward tackling topological bias. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):8073–8087, 2025. doi: 10.1109/TPAMI.2025.3574354.

Qitian Wu, Feiping Nie, Chundong Yang, Tong Bao, and Junchi Yan. Graph out-of-distribution generalization via causal intervention. In *Proceedings of the ACM Web Conference 2024 (WWW '24)*, pp. 850–860. ACM, 2024a. doi: 10.1145/3589334.3645604.

Yingxin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. Discovering invariant rationales for graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2022.

Yucheng Wu, Leye Wang, Xiao Han, and Han-Jia Ye. Graph contrastive learning with cohesive subgraph awareness. In *Proceedings of the ACM Web Conference 2024 (WWW '24)*, pp. 629–640. ACM, 2024b. doi: 10.1145/3589334.3645470.

Jun Xia, Lirong Wu, Ge Wang, Jiao Chen, and Stan Z Li. ProGCL: Rethinking hard negative mining in graph contrastive learning. In *International Conference on Machine Learning (ICML)*, volume 162, pp. 24332–24346. PMLR, 2022.

Tianjun Yao, Yongqiang Chen, Zhi Shen, et al. Learning graph invariance by harnessing spuriosity. In *International Conference on Learning Representations (ICLR)*, 2025.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5812–5823, 2020.

Yang Yu, Xiang Wang, Meng Zhang, Ninghao Liu, and Chuan Shi. Provable training for graph contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow Twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning (ICML)*, volume 139, pp. 12310–12320. PMLR, 2021.

Hengrui Zhang, Qitian Wu, Junchi Yan, David Wipf, and Philip S Yu. From canonical correlation analysis to self-supervised graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 76–89, 2021.

Yun Zhang, Hong Zhu, Zhen Song, Piotr Koniusz, and Irwin King. COSTA: Covariance-preserving feature augmentation for graph contrastive learning. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 2524–2534. ACM, 2022. doi: 10.1145/3534678.3539425.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131*, 2020.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Graph contrastive learning with adaptive augmentation. In *Proceedings of the Web Conference 2021 (WWW '21)*, pp. 2069–2080. ACM, 2021. doi: 10.1145/3442381.3449802.