Structure Matters: Principled Dual-View Augmentation for Unsupervised Graph OOD Detection

Wenping Zheng[1,*], Yuanyuan Guo[1], Yuzhi Wang[1], Yang Liu[1]
[1] School of Computer and Information Technology, Shanxi University, Taiyuan, China
wpzheng@sxu.edu.cn, guoyuanyuan08@163.com, wyzn98@gmail.com, yliu0522@163.com

***Abstract*—Unsupervised Graph Out-of-Distribution (OOD) detection is pivotal for ensuring model reliability in open-world scenarios, particularly absent label supervision. Existing approaches predominantly rely on random perturbations, such as feature masking and edge dropout, to construct augmented views. However, these heuristic strategies often disrupt semantic consistency—inadvertently introducing noise rather than informative variance—and fail to capture the intricate interplay between homophilic and heterophilic topologies. To address these limitations, we propose HGOOD, a novel framework featuring a principled dual-view augmentation strategy via hierarchical structural disentanglement. Specifically, HGOOD utilizes an invariant subgraph extractor to decompose the graph into distribution-stable invariant and noise-sensitive environmental subgraphs. A learnable edge separator then refines the invariant structure into homogeneous and heterogeneous views, which are strategically perturbed by the environmental subgraph to generate complementary, semantically faithful augmentations. Furthermore, we integrate a hierarchical contrastive learning scheme with differentiable Softmax-based clustering, enabling the adaptive optimization of group-level representations against dynamic feature distributions. Extensive experiments on multiple graph classification benchmarks demonstrate that HGOOD consistently outperforms state-of-the-art baselines, validating its effectiveness in robust OOD detection.**

***Index Terms*—Graph Neural Networks, Graph OOD Detection, Unsupervised Graph Learning, Structural Disentanglement, Graph Invariant Learning.**

I. Introduction

Real-world data distributions are often non-stationary, causing models to encounter test samples that deviate from the training distribution, known as Out-of-Distribution (OOD) data. GNNs tend to produce overconfident yet incorrect predictions, leading to significant performance degradation and reduced reliability [1]. Consequently, accurately identifying OOD samples has become a critical challenge for robust graph learning.

Most existing methods [2], [3] rely on annotated ID data, limiting their applicability in label-scarce settings. To overcome this limitation, recent studies train OOD-specific GNNs using only unlabeled ID data [4], [5]. These methods adopt unsupervised strategies, such as graph contrastive learning, to capture discriminative patterns. For example, GOOD-D [4] generates augmented views via random perturbations, while SEGO [5] constructs structure-aware views based on structural entropy. Both methods further employ multi-granular contrastive learning with group-level clustering to capture high-order semantics and enhance representation robustness under distribution shifts.

While these approaches mitigate label dependence, they still suffer from three limitations. First, random perturbations (e.g., feature masking and edge dropping) disrupt semantic consistency, introducing spurious correlations that obscure true structural patterns. Second, graph topology includes both homophilic and heterophilic connections. A unified augmentation strategy over-smooths node representations, diminishing feature distinctions and failing to preserve topological heterogeneity. Third, existing group-level contrastive methods use decoupled clustering, relying on node features while ignoring the interaction between topology and attributes.

To address these challenges, we propose HGOOD, an unsupervised framework that constructs invariant augmentations via structural decoupling. HGOOD decomposes the input graph into a distribution-stable invariant subgraph and a noise-sensitive environmental subgraph. We introduce a learnable edge separator to further disentangle the invariant structure into homogeneous and heterogeneous views. These views are perturbed by the environmental subgraph to enable hierarchical contrastive learning. HGOOD preserves essential semantics, enhancing robustness under distribution shifts.

The main contributions of this work are summarized as follows:

- **Semantic-Aware Structural Disentanglement:** We propose a framework that decouples graphs into invariant and environmental components, and separates the invariant structure into homogeneous and heterogeneous views.
- **Dynamic Global Contrast:** We design a group-level contrastive scheme that jointly optimizes node features and global semantics to enhance OOD detection.
- **Strong Empirical Performance:** Experiments on multiple benchmarks show that HGOOD outperforms existing methods.

II. Related Work

1) OOD Detection on Graphs: OOD detection ensures model reliability under distribution shifts. While widely stud-

Wenping Zheng is the corresponding author. This work is supported by the National Natural Science Foundation of China (Nos. 62072292).

ied in computer vision and NLP [6], [7], it remains challenging for graphs due to complex topological dependencies. Existing methods are either supervised, relying on labeled in-distribution data [2], [3], or unsupervised, leveraging unlabeled data to capture in-distribution patterns [4], [5]. HGOOD follows the unsupervised paradigm with a focus on principled structural modeling for graph OOD detection.

2) Graph Contrastive Learning: GCL is widely used for unsupervised graph representation learning [5], [8]–[10], typically by contrasting augmented views of the same graph. However, common augmentations fail to capture diverse topological patterns. HGOOD learns to generate homogeneous and heterogeneous views with environmental perturbations for consistent contrastive learning.

3) Graph Invariant Learning: GIL aims to capture distribution-stable patterns while filtering spurious correlations [11]–[14], making it suitable for OOD scenarios. Existing methods typically treat invariant structures as a whole [14], overlooking their internal diversity. HGOOD further disentangles invariant structures into homogeneous and heterogeneous views for more fine-grained modeling.

III. Preliminaries and Problem Formulation

A. Notations

Let $G = (\mathscr{V}, \mathscr{E}, \mathbf{X})$ denote a graph, where $\mathscr{V} = \{v_1, \ldots, v_n\}$ is the set of n nodes and $\mathscr{E} \subseteq \mathscr{V} \times \mathscr{V}$ is the set of edges. The node attributes are represented by a feature matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, where d is the feature dimension. The topological structure is described by an adjacency matrix $\mathbf{A} \in \{0,1\}^{n \times n}$, where $\mathbf{A}_{ij} = 1$ if $(v_i, v_j) \in \mathscr{E}$, and 0 otherwise. Consequently, a graph can be concisely represented as $G = (\mathbf{A}, \mathbf{X})$.

B. Problem Formulation

In this work, we focus on the problem of Unsupervised Graph-level Out-of-Distribution (OOD) Detection. We consider a training dataset $\mathscr{D}_{\text{train}} = \{G_1, \ldots, G_M\}$ consisting solely of unlabeled graphs sampled from an In-Distribution (ID) probability distribution $\mathbb{P}_{\text{ID}}$. During the testing phase, the model encounters samples from a mixed dataset $\mathscr{D}_{\text{test}} = \mathscr{D}_{\text{test}}^{\text{ID}} \cup \mathscr{D}_{\text{test}}^{\text{OOD}}$, where $\mathscr{D}_{\text{test}}^{\text{ID}}$ is drawn from $\mathbb{P}_{\text{ID}}$ (disjoint from $\mathscr{D}_{\text{train}}$) and $\mathscr{D}_{\text{test}}^{\text{OOD}}$ is drawn from a distinct distribution $\mathbb{P}_{\text{OOD}}$ (i.e., $\mathbb{P}_{\text{ID}} \neq \mathbb{P}_{\text{OOD}}$).

The objective is to learn a scoring function $f : \mathscr{G} \to \mathbb{R}$ that assigns an anomaly score $s = f(G)$ to an input graph G. The score should enable the distinction between ID and OOD samples, such that:

$$f(G_{\text{ood}}) > f(G_{\text{id}}), \quad \forall G_{\text{ood}} \in \mathscr{D}_{\text{test}}^{\text{OOD}},\ G_{\text{id}} \in \mathscr{D}_{\text{test}}^{\text{ID}}. \tag{1}$$

Crucially, the training process is strictly unsupervised, meaning no ground-truth labels or OOD samples are accessible during training.

C. Graph Contrastive Learning

Our framework builds upon the paradigm of Graph Contrastive Learning (GCL). A typical GCL framework generates two augmented views, $\tilde{G}_i^a$ and $\tilde{G}_i^b$, for an input graph G_i via data augmentation strategies. A GNN encoder $g(\cdot)$ and a projection head $h(\cdot)$ then map these views into latent representations $\mathbf{z}_i^a, \mathbf{z}_i^b \in \mathbb{R}^{d'}$.

The training objective aims to maximize the agreement between positive pairs (congruent views of the same graph) while minimizing it for negative pairs (views of different graphs). This is commonly achieved by minimizing the InfoNCE loss [9], [15]. Formally, for a batch of N graphs, the pairwise contrastive loss for the i-th graph is defined as:

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

where $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity function, and τ is the temperature parameter. The total loss $\mathscr{L}$ is the arithmetic mean over all positive pairs in the batch:

$$\mathscr{L} = \frac{1}{2N} \sum_{i=1}^{N} \left[\ell(\mathbf{z}_i^a, \mathbf{z}_i^b) + \ell(\mathbf{z}_i^b, \mathbf{z}_i^a) \right]. \tag{3}$$

IV. Methodology

In this section, we introduce the framework (Figure 1), termed HGOOD. It consists of three components: (1) an **Invariant Structure Extractor** that decomposes graphs into invariant subgraphs and environmental noise; (2) a **Learnable Edge Separator** that further splits the invariant structure into homogeneous and heterogeneous views; and (3) a **Differentiable Hierarchical-Level Contrastive** scheme that learns global semantic representations via adaptive clustering. These components enable HGOOD to learn structure-aware representations for OOD detection.

A. Invariant Structure Extraction

To mitigate the impact of distribution shifts, our framework commences with an Invariant Structure Extractor designed to distill distribution-stable substructures from the input graph. Formally, we posit that a graph G is decomposable into two distinct components: an *invariant subgraph* G^I, comprising core structural patterns, and an *environmental subgraph* G^E, representing noise or spurious correlations [12], [14].

Theoretically, an optimal extractor $\mathscr{T}(\cdot)$ identifies the invariant subgraph $G^I = \mathscr{T}(G)$ such that the predictive distribution of the target label Y remains consistent across diverse environments $e \in \mathscr{E}_{nv}$:

$$P(Y \mid G^I, e_1) = P(Y \mid G^I, e_2), \quad \forall e_1, e_2 \in \mathscr{E}_{nv}. \tag{4}$$

In the unsupervised setting where ground-truth labels and environmental indices are unavailable, we assume that intrinsic structural patterns exhibit higher stability than environment-induced noise. Accordingly, we approximate the invariant objective by learning a parameterized extractor $\mathscr{T}_\phi(\cdot)$ that filters out noise-sensitive connections.

To enhance structural stability modeling, we introduce a self-supervised pretraining stage based on masked link prediction.Given a graph $G = (\mathbf{A}, \mathbf{X})$, a subset of edges is randomly removed to obtain a masked graph. A GNN encoder $g_{pre}(\cdot)$ is trained to extract node representations $\mathbf{H} = g_{pre}(G_{mask})$,and a bilinear decoder predicts the existence of masked edges:

$$\hat{p}_{uv} = \sigma\left(\mathbf{h}_u^\top \mathbf{W}_{\text{lp}} \mathbf{h}_v\right), \tag{5}$$

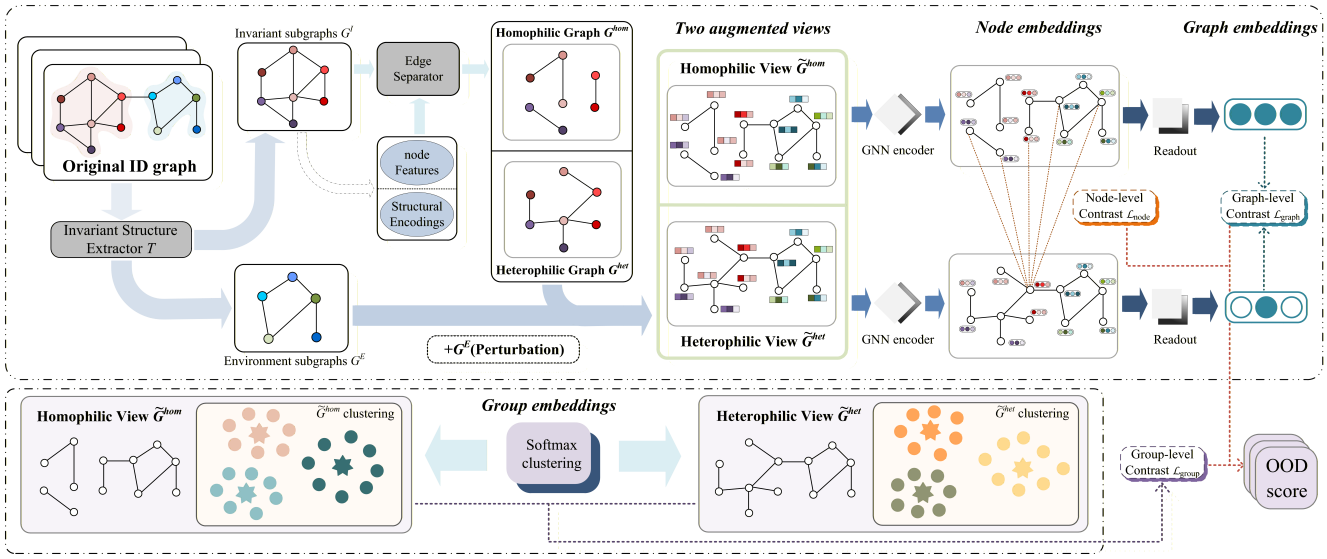

Fig. 1: Overview of HGOOD with three components: invariant structure extractor, learnable edge separator, and hierarchical contrastive learning.

where $\mathbf{W}_{\text{lp}}$ is learnable. This pretraining encourages the encoder to capture reproducible structural dependencies under stochastic perturbations.

Parameterized Probability Generation. After pretraining, the encoder is reused to initialize the invariant extractor $\mathscr{T}_\phi$. Given node representations $\mathbf{H} = \text{GNN}(G)$, an MLP estimates the invariance probability p_{uv} for each edge $(u,v) \in \mathscr{E}$:

$$p_{uv} = \text{MLP}(\mathbf{h}_u \,\|\, \mathbf{h}_v), \tag{6}$$

where $\|$ denotes the concatenation operation.

Differentiable Sampling via Gumbel-Softmax. To bridge the gap between discrete edge sampling and end-to-end backpropagation, we utilize the Gumbel-Softmax [16] relaxation. This technique allows for the differentiable sampling of a soft adjacency mask $\hat{\mathbf{A}}^I$. Specifically, for each edge (u,v), we generate a soft inclusion decision $\hat{y}_{uv} \in (0,1)$:

$$\hat{y}_{uv} = \frac{e^{(\log p_{uv} + g_1)/\tau'}}{e^{(\log p_{uv} + g_1)/\tau'} + e^{(\log(1-p_{uv}) + g_0)/\tau'}}, \tag{7}$$

where $g_0, g_1 \sim \text{Gumbel}(0,1)$ are i.i.d. samples drawn from the standard Gumbel distribution, and τ' is the temperature hyperparameter controlling the sharpness of the categorical distribution.

Based on these learned probabilities, the **Invariant Subgraph** G^I is defined by the weighted adjacency matrix $\mathbf{A}^I$, where $\mathbf{A}^I_{uv} = \hat{y}_{uv}$. Conversely, the **Environmental Subgraph** G^E, which captures the complementary noise patterns, is defined by the matrix $\mathbf{A}^E_{uv} = 1 - \hat{y}_{uv}$.

To preclude the trivial solution where the entire graph is retained as invariant, we impose a sparsity constraint by setting a maximum edge retention ratio r. In practice, r is fixed across datasets and treated as a general sparsity prior rather than a dataset-specific tuning parameter, encouraging the extractor to prioritize the most stable and discriminative structures.

B. Learnable Edge Separator

While the extracted invariant subgraph G^I effectively isolates stable patterns from noise, treating it as a monolithic entity is insufficient. Real-world graphs exhibit complex topological dependencies where *homophilic* (similarity-based) and *heterophilic* (complementarity-based) edges coexist. To capture these fine-grained semantic distinctions, we introduce a Learnable Edge Separator to decompose G^I into semantically distinct homogeneous and heterogeneous views.

Structure-Aware Homophily Estimation. Although GNNs inherently integrate structural and feature information, their unified aggregation mechanism tends to homogenize node features, thereby diluting the discriminative signals necessary to identify heterophilic connections. To mitigate this, we explicitly incorporate topological roles via structural encoding. We employ a random walk diffusion process where the transition matrix is denoted as $\mathbf{T} = \mathbf{A}\mathbf{D}^{-1}$. For each node v_i, a d_s-dimensional structural encoding $\mathbf{z}_i$ is computed to capture its d_s-hop topological context:

$$\mathbf{z}_i = \left[\mathbf{T}_{ii}, (\mathbf{T}^2)_{ii}, \cdots, (\mathbf{T}^{d_s})_{ii}\right]^\top \in \mathbb{R}^{d_s}. \tag{8}$$

We then construct an edge separator to estimate the homophily strength of each edge. By concatenating the raw features $\mathbf{x}_i$ with structural encodings $\mathbf{z}_i$, we form a comprehensive node representation. A symmetric MLP scoring function then estimates the homophily logit ω_{ij} for each edge $(i,j) \in G^I$:

$$\omega_{ij} = \frac{1}{2}\left(\text{MLP}(\mathbf{h}_i \,\|\, \mathbf{h}_j) + \text{MLP}(\mathbf{h}_j \,\|\, \mathbf{h}_i)\right), \tag{9}$$

where $\mathbf{h}_i = \text{MLP}_{\text{enc}}([\mathbf{x}_i \,\|\, \mathbf{z}_i])$ denotes the fused node embedding. A higher ω_{ij} indicates a stronger tendency towards homophily.

Differentiable Partition via Gumbel-Sigmoid. To decouple the invariant structure while maintaining end-to-end differentiability, we employ the Gumbel-Sigmoid [17] relaxation. Based on the estimated logits, we sample a continuous homophily indicator $\hat{m}_{ij} \in (0,1)$ for each edge:

$$\hat{m}_{ij} = \sigma\left(\frac{\omega_{ij} + \log\delta - \log(1-\delta)}{\tau''}\right), \tag{10}$$

where $\sigma(\cdot)$ is the sigmoid function, $\delta \sim \text{Uniform}(0,1)$ represents auxiliary noise, and τ'' is the temperature parameter. Consequently, we obtain two complementary views: the *Homophilic Graph* G^{hom} (emphasizing similarity) and the *Heterophilic Graph* G^{het} (emphasizing structural complementarity). Their weighted adjacency matrices are defined as:

$$\mathbf{A}_{ij}^{hom} = \hat{m}_{ij} \odot \mathbf{A}_{ij}^{I}, \quad \mathbf{A}_{ij}^{het} = (1 - \hat{m}_{ij}) \odot \mathbf{A}_{ij}^{I}. \tag{11}$$

Controllable Dual-View Augmentation. Mere disentanglement effectively reduces the graph to simpler substructures. To construct robust augmented views for contrastive learning, we propose a principled environmental injection strategy. Instead of random perturbation, we strategically inject edges from the previously extracted environmental subgraph G^E back into the separated views as"informative noise." Formally, we obtain the augmented views $\tilde{G}^{hom}$ and $\tilde{G}^{het}$ by superimposing a subset of environmental edges. The injection budget is controlled by a ratio coefficient r_{add}. For a target view $G^* \in \{G^{hom}, G^{het}\}$, the number of injected edges is determined by $n_{add} = \lfloor r_{add}(|E_{G^*}| + |E_{G^E}|)\rfloor$. $\mathbf{A}^{bridge}$ denotes the set of injected bridging edges. This process acts as a structural regularization:

$$\tilde{\mathbf{A}}^{v} = \mathbf{A}^{v} \oplus \mathbf{A}^{E} \oplus \mathbf{A}^{bridge,v}, \quad v \in \{hom, het\}. \tag{12}$$

By explicitly reintroducing the environmental context, this strategy prevents the model from overfitting to sparse invariant structures and enhances robustness against OOD perturbations.

C. Hierarchical-Level Contrastive

To fully exploit the disentangled structural semantics, we propose a hierarchical contrastive learning framework. This module aligns the homogeneous view $\tilde{G}^{hom}$ and the heterogeneous view $\tilde{G}^{het}$ across node, graph, and group levels, ensuring that the learned representations capture both local nuances and global distribution patterns.

Dual-View Representation Learning. We employ two parallel, weight-independent GNN encoders to extract features from the augmented views. Adopting GIN as the backbone for its superior expressiveness, the node embedding update at the l-th layer for a view $v \in \{hom, het\}$ is formulated as: $\mathbf{h}_i^{(v,l)} = \text{MLP}^{(v,l)}\left(\mathbf{h}_i^{(v,l-1)} + \sum_{u \in \mathcal{N}(i)} \mathbf{h}_u^{(v,l-1)}\right)$. After L layers, we obtain the final node embeddings $\mathbf{H}^{(v)}$. A readout function then aggregates these into a graph-level representation: $\mathbf{h}_G^{(v)} = \sum_{i \in \mathcal{V}} \mathbf{h}_i^{(v,L)}$.

Node-Level Contrast. This objective maximizes the consistency of node embeddings across views, ensuring local semantic stability. We project node features into a shared latent space via an MLP to obtain $\mathbf{z}_i^{(v)}$. The loss considers both intra-view and inter-view correlations:

$$\mathscr{L}_{\text{node}} = \frac{1}{|B|} \sum_{G_k \in B} \frac{1}{2|\mathcal{V}_k|} \sum_{i \in \mathcal{V}_k} \left[\ell(\mathbf{z}_i^{hom}, \mathbf{z}_i^{het}) + \ell(\mathbf{z}_i^{het}, \mathbf{z}_i^{hom})\right], \tag{13}$$

where B denotes the batch size, and $\ell(\cdot,\cdot)$ is the InfoNCE loss defined in Eq. (2).

Graph-Level Contrast. To capture coarse-grained topological properties, we align the global representations $\mathbf{h}_G^{hom}$ and $\mathbf{h}_G^{het}$ after projection into $\mathbf{z}_G^{hom}, \mathbf{z}_G^{het}$:

$$\mathscr{L}_{\text{graph}} = \frac{1}{2|B|} \sum_{k=1}^{|B|} \left[\ell(\mathbf{z}_{G_k}^{hom}, \mathbf{z}_{G_k}^{het}) + \ell(\mathbf{z}_{G_k}^{het}, \mathbf{z}_{G_k}^{hom})\right]. \tag{14}$$

Group-Level Contrast. Distinct from instance-level comparisons, group-level contrast aims to discover high-order semantic clusters (prototypes) that persist across views. We apply a differentiable clustering mechanism. For each view $v \in \{hom, het\}$, we compute a soft cluster assignment matrix $\mathbf{P}^{(v)} \in \mathbb{R}^{|B| \times K}$ via a learnable projection $\mathbf{W}_c$:

$$\mathbf{P}^{(v)} = \text{Softmax}(\mathbf{z}_G^{(v)} \mathbf{W}_c), \tag{15}$$

where K is the number of clusters. The cluster centroids (prototypes) $\mathbf{C}^{(v)} \in \mathbb{R}^{K \times d'}$ are then computed as the probability-weighted sum of graph embeddings: $\mathbf{C}_k^{(v)} = \sum_{i=1}^{|B|} \mathbf{P}_{ik}^{(v)} \mathbf{z}_{G_i}^{(v)}$. The group-level loss enforces consistency between the discovered prototypes:

$$\mathscr{L}_{\text{group}} = \frac{1}{2K} \sum_{k=1}^{K} \left[\ell(\mathbf{C}_k^{hom}, \mathbf{C}_k^{het}) + \ell(\mathbf{C}_k^{het}, \mathbf{C}_k^{hom})\right]. \tag{16}$$

Adaptive Loss Balancing. To optimize the three-level contrastive objectives, we use a weighted loss. Since losses at different levels may vary in scale, we adopt an adaptive weighting strategy: $\mathscr{L} = (\sigma_{\text{n}})^{\alpha}\mathscr{L}_{\text{node}} + (\sigma_{\text{g}})^{\alpha}\mathscr{L}_{\text{graph}} + (\sigma_{\text{p}})^{\alpha}\mathscr{L}_{\text{group}}$, where $\sigma_{\text{n}}, \sigma_{\text{g}}, \sigma_{\text{p}}$ are the running standard deviations of the losses, and $\alpha > 0$ controls the weighting strength. This enables adaptive balancing across different levels.

OOD Inference Score. During inference, we use the contrastive loss as an anomaly score. For a test batch $\mathscr{B}$, losses are computed using intra-batch negatives. We then apply Z-score normalization based on training statistics: $Score(G) = \sum_{key \in \{n,g,p\}} \frac{\ell_{key}(G) - \mu_{key}}{\sigma_{key}}$, where μ_{key} and σ_{key} are the mean and standard deviation on the training set. Higher scores indicate larger deviations from ID patterns.

V. Experiments

In this section, we empirically evaluate the proposed **HGOOD** framework through four research questions: **RQ1:** How effective is HGOOD compared with competitive baselines on identifying OOD graphs? **RQ2:** How transferable is HGOOD to anomaly detection? **RQ3:** How do hierarchical contrast and homogeneous–heterogeneous disentanglement affect HGOOD's performance? **RQ4:** How sensitive is HGOOD to key hyperparameters?

A. Experimental Setup

Datasets. Following the standard protocol in GOOD-D [4], we employ 10 pairs of datasets from two mainstream benchmarks, TUDataset [18] and OGB [19], to simulate OOD scenarios. Additionally, to evaluate generalization capability (RQ2), we conduct experiments on 5 datasets from TUDataset under a standard anomaly detection setting. In this setting, samples from the minority class are treated as anomalies, while the majority class constitutes the normal (ID) data.

Baselines. We benchmark HGOOD against 15 competitive methods, categorized into: (1) 7 Graph Contrastive Learning (GCL) approaches [4], [5], [20]; (2) 6 Graph Kernel-based methods [21], [22]; and (3) 2 end-to-end Graph Anomaly Detection models [23], [24].

TABLE I: OOD detection results in terms of AUC (%, mean ± std). The best and runner-up results are highlighted with **bold** and underline, respectively. A.A. is short for average AUC. A.R. denotes average rank.

ID dataset	BZR	PTC-MR	AIDS	ENZYMES	IMDB-B	Tox21	FreeSolv	BBBP	ClinTox	Esol	**A.A.**	**A.R.**
OOD dataset	COX2	MUTAG	DHFR	PROTEIN	IMDB-M	SIDER	ToxCast	BACE	LIPO	MUV		
PK-LOF	42.22±8.39	51.04±6.04	50.15±3.29	50.47±2.87	48.03±2.53	51.33±1.81	49.16±3.70	53.10±2.07	50.00±2.17	50.82±1.48	49.63	13.9
PK-OCSVM	42.55±8.26	49.71±6.58	50.17±3.30	50.46±2.78	48.07±2.41	51.33±1.81	48.82±3.29	53.05±2.10	50.06±2.19	51.00±1.33	49.52	13.8
PK-iF	51.46±1.62	54.29±4.33	51.10±1.43	51.67±2.69	50.67±2.47	49.87±0.82	52.28±1.87	51.47±1.33	50.81±1.10	50.85±3.51	51.45	12.1
WL-LOF	48.99±6.20	53.31±8.98	50.77±2.87	52.66±2.47	52.28±4.50	51.92±1.58	51.47±4.23	52.80±1.91	51.29±3.40	51.26±1.31	51.68	11.3
WL-OCSVM	49.16±4.51	53.31±7.57	50.98±2.71	51.77±2.21	51.38±2.39	51.08±1.46	50.38±3.81	52.85±2.00	50.77±3.69	50.97±1.65	51.27	12.0
WL-iF	50.24±2.49	51.43±2.02	50.10±0.44	51.17±2.01	51.07±2.25	50.25±0.96	52.60±2.38	50.78±0.75	50.41±2.17	50.61±1.96	50.87	13.3
OGGIN	76.66±4.17	80.38±6.84	86.01±6.59	57.65±2.96	67.93±3.86	46.09±1.66	59.60±4.78	61.21±8.12	49.13±4.13	54.04±5.50	63.87	8.9
GLocalKD	75.75±5.99	70.63±3.54	93.67±1.24	57.18±2.03	78.25±4.35	66.28±0.98	64.82±3.31	73.15±1.26	55.71±3.81	86.83±2.35	72.23	6.1
InfoGraph-iF	63.17±9.74	51.43±5.19	93.10±1.35	60.00±1.83	58.73±1.96	56.28±0.81	56.92±1.69	53.68±2.90	48.51±1.87	54.16±5.14	59.60	9.4
InfoGraph-MD	86.14±6.77	50.79±8.49	69.02±11.67	55.25±3.51	**81.38±1.14**	59.97±2.06	58.05±5.46	70.49±4.63	48.12±5.72	77.57±1.69	65.68	8.4
GraphCL-iF	60.00±3.81	50.86±4.30	92.90±1.21	61.33±2.27	59.67±1.65	56.81±0.97	55.55±2.71	59.41±3.58	47.84±0.92	62.12±4.01	60.65	9.7
GraphCL-MD	83.64±6.00	73.03±2.38	93.75±2.13	52.87±6.11	79.09±2.73	58.30±1.52	60.31±5.24	75.72±1.54	51.58±3.64	78.73±1.40	70.70	6.3
GOOD-D*simp*	93.00±3.20	78.43±2.67	98.91±0.41	61.89±2.51	79.71±1.19	65.30±1.27	70.48±2.75	81.56±1.97	66.13±2.98	91.39±0.46	78.68	4.2
GOOD-D	93.52±2.49	80.28±4.02	98.99±0.57	61.74±1.79	79.49±0.95	64.54±1.34	80.95±4.86	83.07±3.25	70.69±4.65	91.41±1.51	80.47	3.2
SEGO	95.81±1.26	85.09±2.34	99.23±0.16	63.59±2.46	78.29±1.53	66.70±0.76	88.35±2.07	86.89±0.25	78.90±2.99	**95.01±0.63**	83.79	2.0
HGOOD	**97.99±0.71**	**86.06±2.69**	**99.79±0.07**	**66.63±3.20**	81.02±1.31	**67.96±1.03**	**92.56±0.93**	**88.63±1.06**	**80.98±3.01**	93.63±1.11	85.52	1.2

Implementation and Evaluation. We adopt AUC as the evaluation metric, where higher values indicate better detection performance. For fair comparison, all baselines are re-implemented under a unified setting, including the same GNN backbone (GIN), data splits, training epochs, batch size, and hardware environment (NVIDIA A100 GPU, 40GB). Results are reported as the mean and standard deviation over 5 runs with fixed random seeds.

B. Performance on OOD Detection (RQ1)

To address RQ1, we compare HGOOD with baselines on the constructed OOD benchmarks, with results shown in Table I. HGOOD achieves the best results on 8 out of 10 dataset pairs, with an average rank of 1.2, and obtains an average AUC of 85.52%, outperforming SEGO by 1.73% and GOOD-D by 5.05%, demonstrating the effectiveness of the proposed structural disentanglement. It ranks second on IMDB-M/IMDB-B and Esol/MUV, possibly due to the dense structures of social graphs and the complex properties of molecular graphs, which make homophily–heterophily separation more challenging. Overall, HGOOD maintains strong performance across different domains.

TABLE II: Anomaly detection results in terms of AUC (% mean ± std).

Method	ENZYMES	DHFR	BZR	NCI1	IMDB-B
PK-OCSVM	53.67±2.66	47.91±3.76	46.85±5.31	49.90±1.18	50.75±3.10
PK-iF	51.30±2.01	52.11±3.96	55.32±6.18	50.58±1.38	50.80±3.17
WL-OCSVM	55.24±2.66	50.24±3.13	50.56±5.87	50.63±1.22	54.08±5.19
WL-iF	51.60±3.81	50.29±2.77	52.46±3.30	50.74±1.70	50.20±0.40
GraphCL-iF	53.60±4.88	51.10±2.35	60.24±5.37	49.88±0.53	56.50±4.90
OCGIN	58.75±5.98	49.23±3.05	65.91±1.47	**71.98±1.21**	60.19±8.90
GLocalKD	61.39±8.81	56.71±3.57	69.42±7.78	68.48±2.39	52.09±3.41
GOOD-D_{simp}	61.23±4.58	62.71±3.38	74.48±4.91	59.56±1.62	65.49±1.06
GOOD-D	63.90±3.69	62.67±3.11	75.16±5.15	61.12±2.21	65.88±0.75
SEGO	76.62±7.35	65.31±2.98	89.21±4.51	70.34±1.31	66.48±0.38
HGOOD	**78.03±6.21**	**66.72±3.01**	**90.12±3.51**	71.03±2.01	**66.67±0.92**

C. Performance on Anomaly Detection (RQ2)

To investigate the transferability of HGOOD (RQ2), we evaluate it on standard anomaly detection tasks with only normal data for training [23], [24], as shown in Table II. HGOOD outperforms competitive baselines, including specialized anomaly detection methods such as GLocalKD and SEGO. This indicates that HGOOD effectively captures the intrinsic distribution of normal data and learns discriminative in-distribution patterns for anomaly detection.

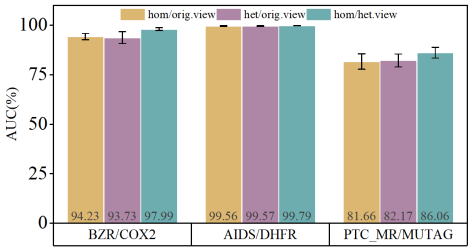

Fig. 2: Ablation on view combinations.

D. Ablation Study (RQ3)

Impact of Hierarchical Contrastive Losses. We evaluate the contributions of node ($\mathscr{L}_{\text{node}}$), graph ($\mathscr{L}_{\text{graph}}$), and group ($\mathscr{L}_{\text{group}}$) levels. Results in Table III show: (1) The full model achieves the best performance on 8 out of 10 datasets, indicating the benefit of multi-scale information. (2) Removing $\mathscr{L}_{\text{graph}}$ causes the largest performance drop, highlighting the importance of graph-level representations for capturing distribution shifts. (3) On complex datasets (e.g., ENZYMES, FreeSolv), improvements are observed only when all three losses are combined, suggesting that different levels provide complementary information.

Effectiveness of Structural Disentanglement. We compare three settings: (1) G^{hom} vs. G^{orig}, (2) G^{het} vs. G^{orig}, and (3) G^{hom} vs. G^{het}. As shown in Figure 2, while single-view settings achieve reasonable performance, G^{hom} vs. G^{het} consistently performs best. This indicates that separating homogeneous and heterogeneous structures provides complementary information and improves discrimination under distribution shifts.

Visualization. We analyze the learned embeddings on the AIDS/DHFR dataset pair. The results show that graph-level representations exhibit the strongest separability, while node- and group-level embeddings provide complementary discriminative information. This demonstrates the effectiveness of HGOOD in capturing multi-scale structural patterns for OOD detection.

E. Parameter Sensitivity (RQ4)

Maximum Edge Ratio r. The parameter $r \in [0.4, 0.9]$ controls the sparsity of the extracted invariant subgraph. As shown in Figure 3, moderate values of r yield the best performance,

TABLE III: Ablation study of HGOOD components. Metric: AUC (%, mean ± std).

$\mathscr{L}_{node}$	$\mathscr{L}_{graph}$	$\mathscr{L}_{group}$	BZR COX2	PTC-MR MUTAG	AIDS DHFR	ENZYMES PROTEIN	IMDB-M IMDB-B	Tox21 SIDER	FreeSolv ToxCast	BBBP BACE	ClinTox LIPO	Esol MUV
✓	-	-	79.31±3.14	70.44±3.02	95.04±2.22	62.23±2.54	78.95±3.81	63.47±0.54	60.88±1.21	68.93±2.88	55.98±2.72	85.64±1.20
-	✓	-	85.64±4.32	79.54±4.71	96.66±1.98	58.94±1.43	76.62±1.53	66.47±2.32	79.50±5.49	77.94±0.99	69.11±2.76	89.57±2.97
-	-	✓	78.81±4.22	75.38±6.64	92.09±0.95	55.83±4.22	74.02±2.15	55.24±3.44	59.36±5.93	60.55±5.98	59.99±5.61	85.98±0.71
✓	✓	-	95.43±2.15	79.43±0.65	98.99±0.11	63.32±1.16	79.98±0.22	**67.99±1.43**	73.76±3.54	82.68±3.54	67.04±3.66	**93.78±2.16**
✓	-	✓	87.31±2.84	78.44±2.70	97.97±0.65	61.43±2.07	78.14±1.98	64.90±1.88	60.34±6.67	73.79±3.10	61.14±5.79	88.48±2.02
-	✓	✓	89.88±6.01	80.83±4.66	98.00±1.65	57.67±3.90	80.47±1.01	66.70±0.93	79.89±4.32	81.00±3.74	74.84±0.93	90.56±2.99
✓	✓	✓	**97.99±0.71**	**86.06±2.69**	**99.79±0.07**	**66.63±3.20**	**81.02±1.31**	67.96±1.03	**92.56±0.93**	**88.63±1.06**	**80.98±3.01**	93.63±1.11

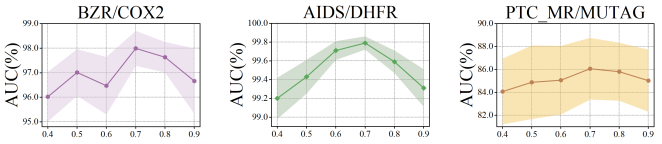

Fig. 3: Sensitivity analysis of the maximum edge ratio r.

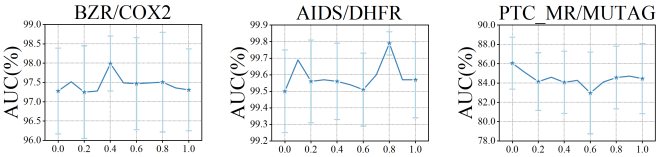

Fig. 4: Sensitivity analysis of the adaptation strength α.

while too small r results in overly sparse subgraphs that lose core structural patterns, and too large r introduces noise that weakens the distinction between invariant and environmental structures.

Adaptation Strength α. We analyze the sensitivity of $\alpha \in [0,1]$ in the adaptive loss weighting mechanism. As depicted in Figure 4, the optimal α varies across datasets, highlighting the need for adaptive balancing to adjust the contribution of each task.

VI. Conclusion

We propose **HGOOD**, an unsupervised framework for graph OOD detection that disentangles invariant structures into homogeneous and heterogeneous views and generates semantically faithful augmentations. A hierarchical contrastive learning scheme captures dependencies from nodes to global semantic prototypes. Experiments on multiple benchmarks show that HGOOD consistently outperforms state-of-the-art baselines, demonstrating the effectiveness of structural disentanglement. Future work includes extending HGOOD to dynamic graph streams and exploring the interpretability of the disentangled views.

References

[1] H. Jo, S. Park, and S.-I. Cho, "A survey of unsupervised learning-based out-of-distribution detection," *2024 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, pp. 1–4, 2024.

[2] L. Wang, D. He, H. Zhang, Y. Liu, W. Wang, S. Pan, D. Jin, and T. Chua, "GOODAT: towards test-time graph out-of-distribution detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 14, 2024, pp. 15 537–15 545.

[3] Y. Guo, C. Yang, Y. Chen, J. Liu, C. Shi, and J. Du, "A data-centric framework to endow graph neural networks with out-of-distribution detection ability," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 638–648.

[4] Y. Liu, K. Ding, H. Liu, and S. Pan, "GOOD-D: on unsupervised graph out-of-distribution detection," in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 2023, pp. 339–347.

[5] Y. Hou, H. Zhu, R. Liu, Y. Su, J. Xia, J. Wu, and K. Xu, "Structural entropy guided unsupervised graph out-of-distribution detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 14, 2025, pp. 17 258–17 266.

[6] V. Sehwag, M. Chiang, and P. Mittal, "SSD: A unified framework for self-supervised outlier detection," in *ICLR*, 2021.

[7] W. Zhou, F. Liu, and M. Chen, "Contrastive out-of-distribution detection for pretrained transformers," in *EMNLP*. ACL, 2021, pp. 1100–1111.

[8] Y. Zheng, Y. Zheng, X. Zhou, C. Gong, V. C. S. Lee, and S. Pan, "Unifying graph contrastive learning with flexible contextual scopes," in *2022 IEEE International Conference on Data Mining*, 2022, pp. 793–802.

[9] Y. Zhu, Y. Xu, F. Yu, Q. Liu, S. Wu, and L. Wang, "Graph contrastive learning with adaptive augmentation," in *Proceedings of the Web Conference*. ACM / IW3C2, 2021, pp. 2069–2080.

[10] K. Ding, Y. Wang, Y. Yang, and H. Liu, "Eliciting structural and semantic global knowledge in unsupervised graph contrastive learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 6, 2023, pp. 7378–7386.

[11] C. Cai and Y. Wang, "Convergence of invariant graph networks," in *International Conference on Machine Learning*, vol. 162. PMLR, 2022, pp. 2457–2484.

[12] H. Li, Z. Zhang, X. Wang, and W. Zhu, "Learning invariant graph representations for out-of-distribution generalization," in *Neural Information Processing Systems*, vol. 35, 2022, pp. 11 828–11 841.

[13] D. Xia, X. Wang, N. Liu, and C. Shi, "Learning invariant representations of graph neural networks via cluster generalization," in *Neural Information Processing Systems*, vol. 36, 2023, pp. 45 602–45 613.

[14] T. Jia, H. Li, C. Yang, T. Tao, and C. Shi, "Graph invariant learning with subgraph co-mixup for out-of-distribution generalization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 8, 2024, pp. 8562–8570.

[15] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, vol. 119. PMLR, 2020, pp. 1597–1607.

[16] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *ICLR*, 2017.

[17] C. J. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," in *ICLR*, 2017.

[18] C. Morris, N. M. Kriege, F. Bause, K. Kersting, P. Mutzel, and M. Neumann, "Tudataset: A collection of benchmark datasets for learning with graphs," *in ICML*, 2020.

[19] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec, "Open graph benchmark: Datasets for machine learning on graphs," in *Neural Information Processing Systems*, vol. 33, 2020, pp. 22 118–22 133.

[20] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, "Graph contrastive learning with augmentations," 2020.

[21] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt, "Graph kernels," *in The Journal of Machine Learning Research*, vol. 11, pp. 1201–1242, 2010.

[22] N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn, and K. M. Borgwardt, "Weisfeiler-lehman graph kernels," *in The Journal of Machine Learning Research*, vol. 12, no. 9, pp. 2539–2561, 2011.

[23] L. Zhao and L. Akoglu, "On using classification datasets to evaluate graph outlier detection: Peculiar observations and new insights," *Big Data*, vol. 11, no. 3, pp. 151–180, 2023.

[24] R. Ma, G. Pang, L. Chen, and A. van den Hengel, "Deep graph-level anomaly detection by glocal knowledge distillation," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 2022, pp. 704–714.