InfoIGL: Invariant Graph Learning Driven by Information Theory

Anonymous authors
Paper under double-blind review

Abstract

Graph-based tasks often violate the i.i.d. assumption as data collection scenarios change, having attracted significant attention to graph out-of-distribution (OOD) generalization. While extracting invariant features is a popular solution, existing methods are limited by the complexity of graphs with distribution shifts in both attributes and structures. Moreover, identifying invariance on graphs is challenging due to lack of prior knowledge for the invariant features. To address these problems, we propose a novel framework, InfoIGL, which leverages information theory to extract invariant graph representations. The framework treats mutual information as the invariance of graphs by exploiting rich *semantic* relations among different distributions. Specifically, InfoIGL decomposes the process of extracting invariant features for graphs into two tasks: **reducing redundant information** and **maximizing mutual information**. To reduce redundancy, InfoIGL leverages attention mechanism to reduce the entropy of graph representations through optimizing their probability distribution. Then InfoIGL integrates semantic-wise and instance-wise contrastive learning to maximize mutual information through joint optimization. Additionally, instance constraint and hard negative mining are utilized to avoid the collapse of contrastive learning. Experiments on both synthetic and real-world datasets demonstrate that our method achieves state-of-the-art performance under OOD generalization for graph classification tasks. The source code is available at https://anonymous.4open.science/r/InfoIGL-268D.

1 Introduction

Graphs are ubiquitous in the real world, appearing as chemical molecules (Fout et al., 2017), social networks (Wu et al., 2020), and knowledge graph (Hamaguchi et al., 2017), to name a few examples. In recent years, graph neural networks (GNNs) (Kipf & Welling, 2017; Velickovic et al., 2017; Xu et al., 2018) have emerged as a potent representation learning technique for analyzing and making predictions on the graph structures. Despite significant advancements, most existing GNN approaches rely heavily on the i.i.d. assumption that the distribution of test data is independently and identically distributed to the training data. Such an assumption, however, seldom holds in practice due to the spatial-temporal heterogeneity nature of graph, thus having attracted significant attention in addressing the out-of-distribution (OOD) problem.

Identifying graph features that remain invariant across distribution shifts (Koyama & Yamaguchi, 2020; Wang & Veitch, 2022; Wu et al., 2022; Liu et al., 2022b) is paramount in overcoming the graph OOD problem. Recent studies focused mainly on developing methods in the principle of invariance, which can be roughly categorized into two groups. On one hand, *graph manipulation approaches* (Miao et al., 2022; Rong et al., 2019; Wang et al., 2021; Han et al., 2022) typically first generate diverse augmented data (*e.g.,* adding or removing nodes and edges) to increase the coverage of data distribution, then learn representations consistent for all manipulated versions. However, accurately performing data augmentations can be non-trivial, particularly when dealing with complex graphs, of which the failure may destroy the invariant features. In contrast, *causal disentanglement methods* aim to extract the underlying causal subgraphs responsible for generating consistent output across different distributions, utilizing techniques such as stable learning (Fan et al., 2021; Li et al., 2021) and causal intervention (Wu et al., 2022; Sui et al., 2022; Liu et al., 2022a). Nevertheless, it requires prior knowledge regarding causal subgraphs, which may be difficult to obtain in practice. As shown in Figure 1, the performance improvements of representative methods belonging to the two groups are faint as compared with traditional ERM (Montanari & Saeed, 2022). To circumvent the aforementioned limitations, in this work, we exploit information theory to effectively identify invariant features within complex graph data. Specifically, on the concept of mutual information(Latham & Roudi, 2009), we first quantify the shared information among the samples within the same class, then isolate the common patterns that determine invariance, and extract the invariant features. However, two major challenges need to be addressed:

- Quantifying mutual information for intricate graphs poses great challenges due to high-dimensional continuous variables and unknown data distributions, which might significantly decrease the model efficiency.
- Directly maximizing the mutual information of the extracted features may be misled by redundant information, thereby leading to inaccurate estimation and exacerbating distribution shift issues.

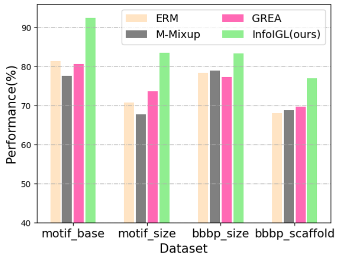


Figure 1: Performance comparison. M-Mixup (Wang et al., 2021) and GREA (Liu et al., 2022a)are representative methods in data manipulation and causal disentanglement approaches respectively.

To bridge the gap, we propose a novel framework **InfoIGL** for graph OOD generalization, which decomposes the invariant feature extraction process into two key tasks driven by information theory: **redundant information reduction** to compress the redundancy of the extracted graph features, and **mutual information maximization** to preserve useful invariant features, both tasks complement each other mutually to achieve effective invariant feature extraction from graphs. Specifically, inspired by the data compression theory (Al-Shaykh & Mersereau, 1998; Hosaka & Kabashima, 2006), InfoIGL first reduces redundancy present in the input graph via an attention mechanism to prevent erroneous identification of redundant information as invariant features. Then InfoIGL utilizes semantic-instance (Zhang et al., 2022b; Yue et al., 2021) and instance-instance (Wang et al., 2022) contrastive learning to maximize the mutual information of positive instances supervised by class labels. To prevent the target collapse issue of contrastive loss (Zhang et al., 2022b; Yao et al., 2022), InfoIGL strengthens instance-wise contrastive learning with instance constraint and hard negative mining techniques. We provide thorough theoretical analyses for these modules from the perspective of information theory and demonstrate the effectiveness of InfoIGL with extensive experiments. Our main contributions are summarized as follows.

- To address graph OOD problem, we propose a novel framework called InfoIGL driven by information theory, which views mutual information as invariant features through realizing two tasks.
- We probe the relationship between semantic-wise and instance-wise contrastive learning to maximize mutual information with adequate theoretical evidence.
- We conduct extensive experiments on diverse benchmark datasets to demonstrate the effectiveness of our proposed framework InfoIGL.

2 Problem Formulation and Preliminary

2.1 Problem Formulation

Let $\mathbb{G}$ and $\mathbb{Y}$ be the sample space and label space, respectively. We denote a graph by $G_i \in \mathbb{G}$ whose adjacent matrix and node feature matrix are $\mathbf{A}$ and $\mathbf{X}$, respectively. A graph predictor $f = \theta \circ \Phi : \mathbb{G} \rightarrow \mathbb{Y}$ maps the input graph G_i to the corresponding label $\mathbf{y}_i \in \mathbb{Y}$, which can be decomposed into a graph encoder Φ and a classifier θ. Let $\mathcal{D}^{tr} = \{(G_i, \mathbf{y}_i)\}_{i=1}^{N^{tr}}$ and $\mathcal{D}^{te} = \{(G_i, \mathbf{y}_i)\}_{i=1}^{N^{te}}$ be the training and testing dataset under distribution shift, *i.e.,* $P^{tr}(G_i, \mathbf{y}_i) \neq P^{te}(G_i, \mathbf{y}_i)$, where $\mathcal{D}^{te}$ remains unobserved during the training stage. Consequently, the generalization for graph OOD can be formulated as follows:

$$\min_{\Phi:G\rightarrow H, \theta:H\rightarrow Y} \mathbb{E}_{G,\mathbf{y}\in\mathcal{D}^{te}}[\mathcal{L}(\mathbf{y}, \theta \circ \Phi(G)] \tag{1}$$

To overcome the distribution shifts between $\mathcal{D}^{te}$ and $\mathcal{D}^{tr}$, invariant representation learning aims to optimize the encoder $\Phi(\cdot)$ to capture representations that keep invariant across all distributions when the classifier θ is deterministic and well performed, *i.e.,* $P^{tr}(Y|\Phi(G)) = P^{te}(Y|\Phi(G))$. Towards this end, we scrutinize the task of invariant feature learning through the lens of information theory. Our fundamental insight is that **the entropy of invariant features is equivalent to the mutual information across graphs of the same class**. By maximizing the mutual information between extracted graph representations, we can approach the ideal mutual information and ultimately obtain representations of invariant features. However, the quantization of mutual information for the extracted graph representations is vulnerable to interference from redundant information present in the graphs. To overcome this challenge, we propose to decompose this task into two subtasks (*cf.* Figure 2): the compression of redundant information, followed by the maximization of mutual information.

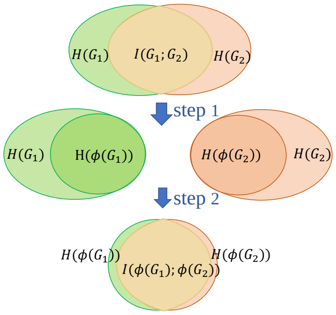


Figure 2: Task decomposition. Step1: Compressing $H(\Phi(G))$ to reduce redundancy; Step2: Maximizing $I(\Phi(G_1);\Phi(G_2))$ to its upper bound $I(G_1, G_2)$

2.2 Task I: Redundant Information Reduction

In this work, we use the term "*redundant information*" to refer to the irrelevant yet potentially confounding features present in graph data. According to Proposition 1, directly quantifying the mutual information without first isolating these extraneous factors may be misled by redundant information (Chuang et al., 2022), resulting in inaccurate estimation.

Proposition 1. *Given variables X_1, X_2 and nosie S_1, S_2, X is independent of S, we have: $I(X_1, S_1; X_2, S_2) - I(X_1; X_2) = I(S_1; S_2) \geq 0$. Directly calculating the mutual information between variables with noise may lead to capturing spurious correlations (Chuang et al., 2022).*

To solve this problem, we take inspiration from the concept of "redundancy" in information theory (*cf.* Theorem 1), which guides us to reduce redundant information via data compression.

Theorem 1. *In information theory (Shannon, 1948), the redundancy γ of variable X is defined as follows: $\gamma = 1 - \eta = 1 - \frac{\mathrm{H}_\infty(\mathrm{X})}{\mathrm{H}_\mathrm{m}(\mathrm{X})}$, where $H_\infty(X)$ is the limiting entropy, $H_m(X)$ is the m-order Markov average entropy.*

In the context of information theory, $H_\infty(X)$ and $H_m(X)$ refer to the average entropy per symbol (*i.e.,* X) in a sequence of infinite or m symbols respectively. To extend the concept of redundancy to graph learning, we posit that the entropy of graph representations generated by an ideal graph encoder through infinite training data (*i.e.,* $H(\Phi_\infty(G))$) can be viewed as the limiting entropy, while its counterpart obtained from the training set (*i.e.,* $H(\Phi(G))$) can be viewed as m-order Markov average entropy. Thereby, the goal of task I is to reduce $H(\Phi(G))$ close to $H(\Phi^*(G))$.

2.3 Task II: Mutual Information Maximization

Mutual information assesses the level of shared information between variables, which can serve as a measure of invariance across different graphs. Ideally, if the mutual information between graph representations generated by graph encoder Φ reaches its upper bound $I(G_1; G_2)$, it can be considered as the optimal graph encoder Φ^* for extracting invariant features (Shwartz-Ziv & Tishby, 2017; Tian et al., 2020), that is, $\Phi^* = \underset{\Phi}{\operatorname{argmax}} I(\Phi(G_1); \Phi(G_2))$. However, directly maximizing mutual information poses significant challenges due to the high dimensionality of continuous variables and the unknown nature of data distributions. To overcome these obstacles, it is possible to establish tractable objectives that approaches the boundary of mutual information.

Theorem 2. *(Mutual information boundary (Oord et al., 2018; Poole et al., 2019)) Given two variables X, Y, the bound of mutual information between X and Y can be defined as follows:*

$$I(X;Y) \geq E[\frac{1}{K}\sum_{i=1}^{K}\log\frac{\exp(f(x_i, y_i)))}{\frac{1}{K}\sum_{j=1}^{K}\exp(f(x_i, y_j))}] \triangleq I_{NCE} \tag{2}$$

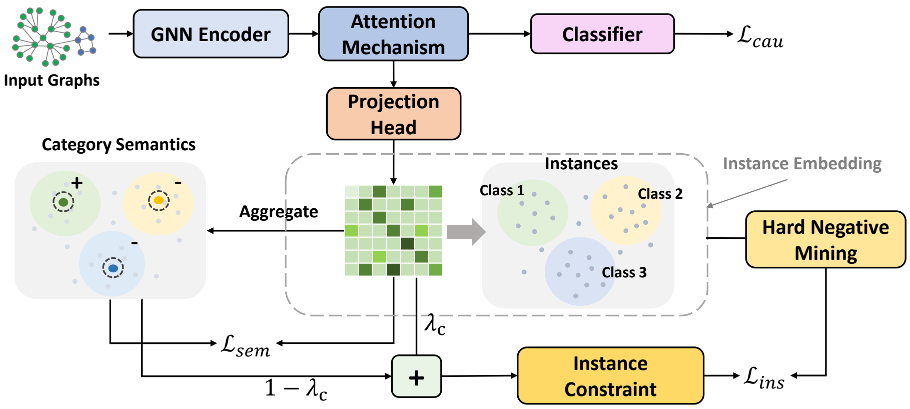


Figure 3: The proposed InfoIGL framework.

where x_i and y_i denote a positive pair sampled from the joint distribution $P(X,Y)$, while x_i and y_j form a negative pair sampled from the product of marginal distributions $P(X)P(Y)$.

As defined in Equation equation 2, maximizing the mutual information equals to minimizing the InfoNCE loss (Oord et al., 2018), which inspires us to leverage contrastive learning to accomplish Task II.

3 Method and Theoretical Analysis

In this section, we introduce our framework InfoIGL that simultaneously addresses two tasks to extract invariant graph representations, as illustrated in Figure 3. Specifically, (1) For task I, we utilize attention mechanism to reduce redundancy. (2) For task II, we jointly optimize the semantic-instance and instance-instance contrastive learning to maximize the mutual information. (3) Finally, we transfer the invariant representation to the downstream task — graph classification. Task I and Task II are complementary to each other. Reducing redundancy can prevent capturing spurious correlations, while maximizing mutual information can avoid discarding useful information.

3.1 Towards Redundant Information Reduction

To reduce redundancy, we hire attention mechanism (Sui et al., 2022; Brody et al., 2022; Thekumparampil et al., 2018; Kim & Oh, 2022) to assign weights to nodes and edges according to the Theorem 3.

Theorem 3. *The entropy of variable X (i.e., $H(X) = -\sum p(x_i)log_2 p(x_i)$) is a convex function, which can be reduced through optimization of the probability distribution $P(X)$.*

In the context of graphs, we can reduce the entropy of graph representations by optimizing the probability distribution of features, which can be realized through attention mechanism (Sui et al., 2022; Brody et al., 2022). First, we obtain the representations for graph nodes with GNNs. Taking GIN (Xu et al., 2018) as an example, the node update module is defined as follows:

$$\mathbf{h}_v^{(k)} = \mathrm{MLP}^k((1+\epsilon^{(k)})\cdot \mathbf{h}_v^{(k-1)} + \sum_{u\in N(v)} \mathbf{h}_u^{(k-1)}) \tag{3}$$

where $\mathrm{MLP}(\cdot)$ stands for multi-layer perceptron, ϵ is a learnable parameter, $\mathbf{h}_v$ and $\mathbf{h}_u$ separately denote the representations of nodes v and u, $N(v)$ denotes the neighbour nodes of v.

The attention score of node v and edge (u,v) can be obtained as follows:

$$\alpha_v = \mathrm{softmax}(\frac{Q_v K_v^\top}{\sqrt{d_k}})V_v, \quad \alpha_{uv} = \mathrm{softmax}(\mathrm{LeakyReLU}(\mathrm{MLP}(\mathbf{h}_u || \mathbf{h}_v))) \tag{4}$$

$Q_v = \mathbf{h}_v \cdot \mathbf{W}^Q, K_v = \mathbf{h}_v \cdot \mathbf{W}^K, V_v = \mathbf{h}_v \cdot \mathbf{W}^V$, $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$ are trainable parameter matrices, d_k is K_v's dimension, $||$ is the concatenation operation.

After assigning weights for nodes and edges, we encode graphs into graph-level representations $\mathbf{h}_G$:

$$\mathbf{h}'_v = \alpha_v \cdot \mathbf{h}_v, \mathbf{h}'_{vu} = \alpha_{uv} \cdot \mathbf{h}_{uv}, \quad \mathbf{h}_G = \mathrm{READOUT}(\mathbf{H}_v, \mathbf{H}_{uv}) \tag{5}$$

where $\mathbf{h}_v, \mathbf{h}_{uv}$ are node and edge embedding obtained from GIN, $\mathbf{H}_v = [..., \mathbf{h}'_v, ...]^\top_{v \in \mathcal{V}}, \mathbf{H}_{uv} = [..., \mathbf{h}'_{uv}, ...]^\top_{u,v \in \mathcal{V}}$, $\mathcal{V}$ denotes the node set. Till now, we have optimized the probability distribution of graph features and reduced the entropy of graph representations.

3.2 Towards Mutual Information Maximization

According to Theorem 2, we can leverage contrastive learning to maximize the mutual information between representations. However, traditional contrastive-based methods may not effectively work since complex instance-instance relationships may impede model generalization(Yao et al., 2022; Xu et al., 2021; Zhang et al., 2022c). To fully unleashing the advantages of contrastive learning, we jointly optimize semantic-wise and instance-wise contrastive losses after projecting the embedding to another space with projection head, which can promote both inter-class separation and intra-class compactness.

3.2.1 Projection Head

A projection head is a small network that maps the embedding to another space where further contrastive learning is applied (Yao et al., 2022; Gupta et al., 2022), which can prevent the conflicts between reducing representation entropy and strengthening contrastive uniformity according to the Proposition 2.

Proposition 2. *Contrastive learning can promote embedding uniformity (Wang & Isola, 2020), which may increase entropy (Ma et al., 2023) and violate task I. Using a projection head as a springboard can avoid potential entropy increase in the original embedding* $\mathbf{h}_G$.

We employ a two-layer MLP $f_{\theta_p}(\cdot)$ as the projection head for $\mathbf{h}_G$:

$$\mathbf{z}_G = f_{\theta_p}(\mathbf{h}_G). \tag{6}$$

Then InfoIGL applies contrastive learning and encourages uniformity in new space $\mathbf{z}_G$.

3.2.2 Semantic-wise Contrastive Learning

Semantic-instance contrastive learning is more robust to complex instances by neutralizing the noise in multiple instances. To realize semantic-wise contrastive learning, we first introduce the cluster center of each class as the corresponding category semantic. Formally, we initialize the semantic $\mathbf{w}_c$ as the average semantic representation over examples belonging to class c: $\mathbf{w}_c = \frac{1}{N_c}\sum_{i=1}^{N_c}(\mathbf{z}_{G_i})$, N_c is the number of $\mathbf{z}_G$ with label c in a batch. Then we update the current round $\mathbf{w}_c^{(r)}$ by calculating the similarity between each instance embedding $\mathbf{z}_{G_i}$ and the semantic of last round $\mathbf{w}_c^{(r-1)}$:

$$m_i^{(r)} = \mathrm{softmax}(\mathrm{cosine}(\mathbf{z}_{G_i}, \mathbf{w}_c^{(r-1)})), \quad \mathbf{w}_c^{(r)} = \textstyle\sum_{i=1}^{N_c} m_i^{(r)} \cdot \mathbf{z}_{G_i} \tag{7}$$

where $\mathrm{cosine}(\cdot)$ denotes the cosine similarity. Then we define the semantic-instance contrastive loss:

$$\mathcal{L}_{\text{sem}} = -\frac{1}{N^{tr}}\sum_{i=1}^{N^{tr}} \log \frac{\exp(\mathbf{z}_{G_i}^\top \mathbf{w}_c/\tau)}{\exp(\mathbf{z}_{G_i}^\top \mathbf{w}_c/\tau) + \sum_{k=1,k\neq c}^{C-1}\exp(\mathbf{z}_{G_i}^\top \mathbf{w}_k/\tau)} \tag{8}$$

where N^{tr} denotes the number of graphs in a batch, $\mathbf{w}_c$ denotes the target category semantic of $\mathbf{z}_{G_i}$, C denotes the number of classes, τ is the scale factor.

3.2.3 Instance-wise Contrastive Learning

However, merely leveraging semantic-wise contastive learning may sacrifice some potential useful semantic relations (Yao et al., 2022) according to the Proposition 3.

Proposition 3. *The mutual information between semantic $\mathbf{w}_c$ and instance $\mathbf{z}_G$ is lower than the mutual information between instance $\mathbf{z}_G$: $I(\sum_{i=1}^{N_c}(m_i \cdot \mathbf{z}_{G_i}); \mathbf{z}_{G_1}; ..; \mathbf{z}_{G_{N_c}}) \leq I(\mathbf{z}_{G_1}; ...; \mathbf{z}_{G_i}; ...; \mathbf{z}_{G_{N_c}})$.*

To compensate for the missing mutual information, we introduce instance-instance contrastive learning. The loss function for instance-instance contrastive learning can be defined as follows:

$$\mathcal{L}_{\text{ins}} = -\frac{1}{N^{tr}} \sum_{i=1}^{N^{tr}} \log \frac{\exp(\mathbf{z}_{G_i}^{\top} \mathbf{z}_{G_+} / \tau')}{\exp(\mathbf{z}_{G_i}^{\top} \mathbf{z}_{G_+} / \tau) + \sum_{k=1}^{K} \exp(\mathbf{z}_{G_i}^{\top} \mathbf{z}_{G_k} / \tau')} \tag{9}$$

Positive sample $\mathbf{z}_{G_+}$ are randomly selected from graphs that belong to the same class as $\mathbf{z}_{G_i}$, K denotes the number of negative samples for each graph instance.

3.2.4 Discussion for Instance-instance Contrastive Learning

For instance-instance contrastive learning, model collapse (Jing et al., 2021)(*i.e.,* samples are mapped to the same point) is not uncommon due to excessive alignment. So we apply instance constraint (Yao et al.) and hard negative mining (Xuan et al., 2020; Robinson et al., 2020) to prevent getting stuck in trivial solutions (Yao et al., 2022).

Instance constraint. Enhancing the uniformity (Wang & Isola, 2020) of graph embeddings $\mathbf{z}_G$ can prevent model collapse from excessive alignment, which can be assisted by the uniformity of semantics $\mathbf{w}_c$. The semantics $\mathbf{w}_c$ are ensured to be uniformly distributed by semantic-wise contrastive learning. Thus we utilize instance constraint:

$$\mathbf{z}'_G = \lambda_c \cdot \mathbf{z}_G + (1 - \lambda_c) \cdot \mathbf{w}_c \tag{10}$$

where $\mathbf{w}_c$ refers to the corresponding semantic belonging to the same class as $\mathbf{z}_G$.

Hard negative mining. Hard negative pair (Xuan et al., 2020; Robinson et al., 2020) plays an important role in contrastive loss, which can help the network learn a better decision boundary. We identify hard negative graph samples according to the Definition 1.

Definition 1. *For sample $\mathbf{z}'_{G_i}$, there are two principles to mine hard negative samples: 1) $y(\mathbf{z}'_{hard}) \neq y(\mathbf{z}'_{G_i})$, y means the graph label; 2). $||\Phi(\mathbf{z}'_{hard}) - \Phi(\mathbf{z}'_{G_i})||^2 \leq \delta$, δ is a very small value.*

To identify hard negative samples for instances $\mathbf{z}'_G$ belonging to class c, we calculate the distance between semantic $\mathbf{w}_c$ and samples from other classes within a batch, then we choose the K nearest ones as hard negative samples $\{\mathbf{z}'_{\text{hard}_\text{k}}\}_{k=1}^{K}$. Then the loss for instance-wise contrastive learning can be modified as follows:

$$\mathcal{L}_{\text{ins}} = -\frac{1}{N^{tr}} \sum_{i=1}^{N^{tr}} \log \frac{\exp(\mathbf{z}'^{\top}_{G_i} \mathbf{z}'_{G_+} / \tau')}{\exp(\mathbf{z}'^{\top}_{G_i} \mathbf{z}'_{G_+} / \tau') + \sum_{k=1}^{K} \exp(\mathbf{z}'^{\top}_{G_i} \mathbf{z}'_{hard_k} / \tau')} \tag{11}$$

3.3 Downstream Task Transfer and Overall Framework

To make the invariant graph representation applicable for graph classification, we define the loss function for prediction as follows:

$$\mathcal{L}_{\text{pred}} = -\frac{1}{N^{tr}} \sum_{i=1}^{N^{tr}} \mathbf{y}_i^{\top} \log(\theta(\mathbf{h}_{G_i})) \tag{12}$$

where $\mathbf{y}_i$ is the label of G_i, θ is the classifier for $\mathbf{h}_{G_i}$.

The overall framework is illustrated in Figure 3 and the final loss can be given by:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_s \mathcal{L}_{\text{sem}} + \lambda_i \mathcal{L}_{\text{ins}} \tag{13}$$

where the hyperparameters λ_s, λ_i are scaling weights for each loss, which can adjust the impact of different modules on the model's results. The algorithm of training stage is listed in the Appendix.

4 Experiments

In this section, we conduct extensive experiments on multiple datasets to answer the following questions: Q1: How effective is InfoIGL compared to prior methods for OOD generalization in graphs? Q2: How do the two tasks make sense respectively? Q3: How do the different contrastive modes impact InfoIGL's performance respectively? Q4: How sensitive the model is to the hyperparameters?

4.1 Datasets and Baselines

We conduct experiments on one synthetic (*i.e.,* Motif) and three real-world (*i.e.,* HIV, Molbbbp, and CMNIST) datasets which are designed for graph OOD (Gui et al., 2022; Hu et al., 2020), where Motif and CMNIST are evaluated with ROC-AUC while HIV and Molbbbp are evaluated with ACC. We compare our InfoIGL against diverse graph OOD generalization baselines: **Optimization methods:** ERM, IRM (Arjovsky et al., 2019), VREx (Krueger et al., 2021), GroupDRO (Sagawa et al., 2020), FLAG (Kong et al., 2022); **Causal learning:** DIR (Wu et al., 2022), CAL (Sui et al., 2022), GREA (Liu et al., 2022a), CIGA (Chen et al.), Disc (Fan et al., 2022); **Stable learning:** StableGNN (Fan et al., 2021), OOD-GNN (Li et al., 2021); **Data manipulation:** GSAT (Miao et al., 2022), DropEdge (Rong et al., 2019), M-Mixup (Wang et al., 2021), G-Mixup (Han et al., 2022). Since the practical implementation of InfoIGL involves utilizing graph contrastive learning to maximize mutual information, we also incorporate **classical graph contrastive learning** methods as benchmarks, including CNC (Zhang et al., 2022a), GMI (Peng et al., 2020), Infograph (Sun et al., 2019), GraphCL (Hafidi et al.). Detailed explanations for datasets and baselines are provided in the Appendix.

4.2 Overall Results(Q1)

Table 1: Performance of different methods on synthetic (Motif) and real-world (HIV, Molbbbp, CMNIST) datasets. The best results are in **bold**, and the runner-up results are $\underline{\text{underlined}}$.

methods	Motif		HIV		Molbbbp		CMNIST
	size	base	size	scaffold	size	scaffold	color
ERM	$70.75_{\pm 0.56}$	$81.44_{\pm 0.45}$	$63.26_{\pm 2.47}$	$72.33_{\pm 1.04}$	$78.29_{\pm 3.76}$	$68.10_{\pm 1.68}$	$28.60_{\pm 1.87}$
IRM	$69.77_{\pm 0.88}$	$80.71_{\pm 0.46}$	$59.90_{\pm 3.15}$	$72.59_{\pm 0.45}$	$77.56_{\pm 2.48}$	$67.22_{\pm 1.15}$	$27.83_{\pm 2.13}$
GroupDRO	$69.98_{\pm 0.86}$	$81.43_{\pm 0.70}$	$61.37_{\pm 2.79}$	$\mathbf{73.64}_{\pm \mathbf{0.86}}$	$79.27_{\pm 2.43}$	$66.47_{\pm 2.39}$	$29.07_{\pm 3.14}$
VREx	$70.24_{\pm 0.72}$	$81.56_{\pm 0.35}$	$60.23_{\pm 1.70}$	$72.60_{\pm 0.82}$	$78.76_{\pm 2.37}$	$68.74_{\pm 1.03}$	$28.48_{\pm 2.87}$
FLAG	$56.26_{\pm 3.98}$	$72.29_{\pm 1.31}$	$66.44_{\pm 2.32}$	$70.45_{\pm 1.55}$	$79.26_{\pm 2.26}$	$67.69_{\pm 2.36}$	$32.30_{\pm 2.69}$
DIR	$54.96_{\pm 9.32}$	$82.96_{\pm 4.47}$	$72.61_{\pm 2.03}$	$69.05_{\pm 0.92}$	$76.40_{\pm 4.43}$	$66.86_{\pm 2.25}$	$33.20_{\pm 6.17}$
CAL	$66.64_{\pm 2.74}$	$81.94\ _{\pm 1.20}$	$\underline{83.33_{\pm 2.84}}$	$73.05_{\pm 1.86}$	$79.20\ _{\pm 3.81}$	$67.37_{\pm 3.61}$	$27.99_{\pm 3.24}$
GREA	$\underline{73.31_{\pm 1.85}}$	$80.60_{\pm 2.49}$	$66.48_{\pm 4.13}$	$70.96_{\pm 3.16}$	$77.34_{\pm 3.52}$	$\underline{69.72_{\pm 1.66}}$	$29.02_{\pm 3.26}$
CIGA	$70.65_{\pm 4.81}$	$75.01_{\pm 3.56}$	$65.98_{\pm 3.31}$	$64.92_{\pm 2.09}$	$76.08_{\pm 1.21}$	$66.43_{\pm 1.99}$	$23.36_{\pm 9.32}$
DisC	$53.34_{\pm 13.71}$	$76.70_{\pm 0.47}$	$56.59_{\pm 10.09}$	$67.12_{\pm 2.11}$	$75.68_{\pm 3.16}$	$60.72_{\pm 0.89}$	$24.99_{\pm 1.78}$
GSAT	$64.16_{\pm 3.35}$	$\underline{83.71_{\pm 2.30}}$	$65.63_{\pm 0.88}$	$68.88_{\pm 1.96}$	$75.63_{\pm 3.83}$	$66.78_{\pm 1.45}$	$28.17_{\pm 1.26}$
DropEdge	$55.27_{\pm 5.93}$	$70.84_{\pm 6.81}$	$54.92_{\pm 1.73}$	$66.78_{\pm 2.68}$	$78.32_{\pm 3.44}$	$66.49_{\pm 1.55}$	$22.65_{\pm 2.90}$
M-Mixup	$67.81_{\pm 1.13}$	$77.63_{\pm 0.57}$	$64.87_{\pm 1.77}$	$72.03_{\pm 0.53}$	$78.92_{\pm 2.43}$	$68.75_{\pm 1.03}$	$26.47_{\pm 3.45}$
G-Mixup	$59.92_{\pm 2.10}$	$74.66_{\pm 1.89}$	$70.53_{\pm 2.02}$	$71.69_{\pm 1.74}$	$78.55_{\pm 4.16}$	$67.44_{\pm 1.62}$	$31.85_{\pm 5.82}$
OOD-GNN	$68.62_{\pm 2.98}$	$74.62_{\pm 2.66}$	$57.49_{\pm 1.08}$	$70.45_{\pm 2.02}$	$79.48_{\pm 4.19}$	$66.72_{\pm 1.23}$	$26.49_{\pm 2.94}$
StableGNN	$59.83_{\pm 3.40}$	$73.04_{\pm 2.78}$	$58.33_{\pm 4.69}$	$68.23_{\pm 2.44}$	$77.47_{\pm 4.69}$	$66.74_{\pm 1.30}$	$28.38_{\pm 3.49}$
CNC	$66.52_{\pm 3.12}$	$82.51_{\pm 1.26}$	$70.68_{\pm 2.15}$	$66.53_{\pm 2.19}$	$76.19_{\pm 3.52}$	$68.16_{\pm 1.25}$	$32.41_{\pm 1.28}$
GMI	$67.90_{\pm 1.46}$	$79.52_{\pm 0.45}$	$74.34_{\pm 0.55}$	$\underline{73.44_{\pm 0.35}}$	$77.67_{\pm 0.30}$	$69.38_{\pm 1.02}$	$30.24_{\pm 5.98}$
InfoGraph	$67.49_{\pm 2.54}$	$75.57_{\pm 0.88}$	$74.63_{\pm 0.80}$	$71.41_{\pm 0.82}$	$\underline{80.82_{\pm 0.49}}$	$70.39_{\pm 1.34}$	$\underline{33.84_{\pm 1.52}}$
GraphCL	$66.90_{\pm 2.80}$	$74.40_{\pm 0.90}$	$77.13_{\pm 0.17}$	$72.94_{\pm 0.68}$	$80.64_{\pm 0.78}$	$69.36_{\pm 1.32}$	$32.81_{\pm 1.71}$
InfoIGL(ours)	$\mathbf{85.53}_{\pm \mathbf{2.37}}$	$\mathbf{92.51}\ _{\pm \mathbf{0.16}}$	$\mathbf{93.15}_{\pm \mathbf{0.77}}$	$72.37_{\pm 1.63}$	$\mathbf{83.39}\ _{\pm \mathbf{2.76}}$	$\mathbf{77.05}\ _{\pm \mathbf{2.24}}$	$\mathbf{38.93}_{\pm \mathbf{1.11}}$
improvemrnt	$\uparrow 12.22\%$	$\uparrow 8.80\%$	$\uparrow 9.82\%$	$\downarrow 1.07\%$	$\uparrow 2.57\%$	$\uparrow 7.33\%$	$\uparrow 5.09\%$

We train and evaluate our proposed InfoIGL, together with all the baselines, 10 times to obtain the average performance (mean ± standard deviation). Details of hyperparameters are listed in the

Appendix. It can be observed from Table 1 that optimization methods exhibit stable performance with moderate accuracy and low variance, while causal learning baselines show unstable performance with undulating accuracy and high variance. Besides, stable learning and data manipulation baselines perform relatively poorly compared to other baselines. Additionally, traditional graph contrastive learning methods can partially combat distributional shifts, but their effectiveness is not as strong as InfoIGL since they were not designed specially to extract invariant features. These observations indicate that almost all of the baselines have their limitations for graph OOD generalization. Our proposed framework, InfoIGL, achieves state-of-the-art performance on diverse datasets with low variance, outperforming the strongest baseline by 9.82% on HIV (size) and 12.22% on Motif (size). These results demonstrate the effectiveness of InfoIGL in extracting stable and invariant graph representations for graph classification tasks.

4.3 Ablation Study for Q2

Table 2: Results of ablation experiments on two tasks.

methods	Motif		HIV		Molbbbp		CMNIST
	size	base	size	scaffold	size	scaffold	color
ERM	70.75$\pm$0.56	81.44$\pm$0.45	63.26$\pm$2.47	72.33$\pm$1.04	78.29$\pm$3.76	68.10$\pm$1.68	28.60$\pm$1.87
InfoIGL-A	69.69$\pm$6.24 $\downarrow$	87.14$\pm$0.88 $\uparrow$	76.99$\pm$2.55$\uparrow$	71.56$\pm$1.96 $\downarrow$	79.72$\pm$3.50 $\uparrow$	74.48$\pm$1.00$\uparrow$	34.54$\pm$2.11$\uparrow$
InfoIGL-C	68.01$\pm$2.09$\downarrow$	86.63$\pm$1.33$\uparrow$	72.81$\pm$2.92$\uparrow$	68.02$\pm$2.28$\downarrow$	75.32$\pm$1.38$\downarrow$	65.62$\pm$1.07$\downarrow$	33.40$\pm$2.10$\uparrow$
InfoIGL	85.53$\pm$2.37$\uparrow$	92.51$\pm$0.16$\uparrow$	93.15$\pm$0.77 $\uparrow$	72.37$\pm$1.63$\downarrow$	83.39$\pm$2.76 $\uparrow$	77.05$\pm$2.24$\uparrow$	38.93$\pm$1.11$\uparrow$

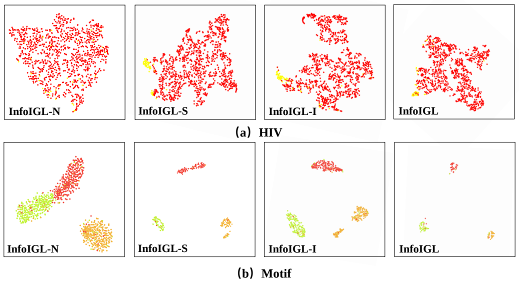
To validate the significance of each task individually, we conduct separate ablation studies on the attention mechanism and contrastive learning. Specifically, we compare InfoIGL with two variations: (1) InfoIGL-A: which includes only the attention mechanism in Task I, and (2) InfoIGL-C, which focuses solely on the contrastive learning in Task II. The results are reported in Table 2. InfoIGL-A outperforms ERM but falls short of InfoIGL, underscoring the significance of contrastive learning and the necessity of Task II. In contrast, InfoIGL-C yields poorer results than ERM and InfoIGL, shedding light on the impact of Task I.

4.4 Ablation Study for Q3

Table 3: Results of ablation experiments on semantic-wise and instance-wise contrastive learning.

methods	Motif		HIV		Molbbbp		CMNIST
	size	base	size	scaffold	size	scaffold	color
InfoIGL-N	69.69$\pm$6.24	87.14$\pm$0.88	76.99$\pm$2.55	71.56$\pm$1.96	79.72$\pm$3.50	74.48$\pm$1.00	34.54$\pm$2.11
InfoIGL-S	84.77$\pm$2.10$\uparrow$	89.93$\pm$0.93 $\uparrow$	87.30$\pm$1.21 $\uparrow$	72.12$\pm$1.87$\uparrow$	81.97$\pm$1.79$\uparrow$	76.76$\pm$3.66$\uparrow$	37.31$\pm$1.50$\uparrow$
InfoIGL-I	80.05$\pm$2.99$\uparrow$	90.36$\pm$1.54$\uparrow$	91.38$\pm$2.38$\uparrow$	63.70$\pm$5.45$\downarrow$	74.91$\pm$2.07$\downarrow$	70.11$\pm$2.02$\downarrow$	35.67$\pm$1.19$\uparrow$
InfoIGL	**85.53$\pm$2.37**$\uparrow$	**92.51$\pm$0.16**$\uparrow$	**93.15$\pm$0.77**$\uparrow$	**72.37$\pm$1.63**$\uparrow$	**83.39$\pm$2.76**$\uparrow$	**77.05$\pm$2.24**$\uparrow$	**38.93$\pm$1.11**$\uparrow$

We perform an ablation study to analyze the impacts of semantic-wise and instance-wise contrastive learning respectively. Specifically, we compare InfoIGL with three variations: (1) InfoIGL-N, which does not use contrastive learning; (2) InfoIGL-S, which employs semantic-wise contrastive learning only; and (3) InfoIGL-I, which applies instance-wise contrastive learning only. We report the results in Table 3. Our observations are as follows: 1) Merely applying instance-instance contrastive learning may cause performance degradation, which confirms the Proposition 1 that directly maximizing mutual information between instances may have a negative impact from an experimental perspective. 2) Applying semantic-instance contrastive learning can achieve improvement coherently on diverse



(a) HIV

(b) Motif

Figure 4: The t-SNE (Van der Maaten & Hinton, 2008) visualizations.

datasets, which demonstrates the robustness of semantic-wise contrastive learning. 3) InfoIGL with both semantic-wise and instance-wise contrastive learning outperforms all of the three variations across diverse datasets, proving that jointly optimizing the two contrastive losses can inspire their individual potentials. Semantic-instance and instance-instance contrastive learning can promote each other and cooperate complementarily.

Additionally, we employ the t-SNE (Van der Maaten & Hinton, 2008) technique to visualize the embedding of graph instances on the HIV (Gui et al., 2022) and Motif (Gui et al., 2022) datasets (*cf.* Figure 4), where four variations (InfoIGL-N, InfoIGL-S, InfoIGL-I, InfoIGL) are compared. The results reveal that compared to InfoIGL-N, the embeddings obtained by InfoIGL-S and InfoIGL-I exhibit a more compact clustering pattern, reflecting the efficacy of semantic-wise and instance-wise contrastive learning in aligning shared information. Furthermore, it is evident that InfoIGL exhibits the best convergence effect, as its embeddings are more tightly clustered than those produced by the other variations.

4.5 Sensitive Analysis(Q3)

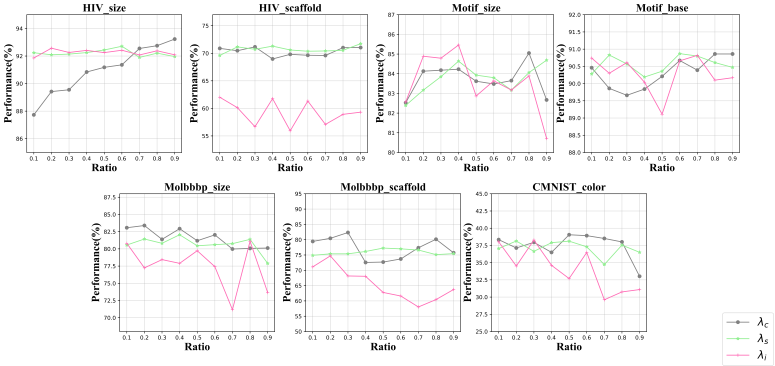


Figure 5: Sensitivity analysis of hyperparameters $\lambda_c, \lambda_s, \lambda_i$

To assess the sensitivity of InfoIGL to its hyperparameters, namely λ_c for instance constraint and λ_s and λ_i for contrastive loss respectively, we conduct sensitivity analysis experiments by tuning these hyperparameters within the range of $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ under the controlled experimental setting. Specifically, when adjusting a specific hyperparameter, we fix the remaining hyperparameters at the values that yield the best performance. The results are presented in Figure 5. It is noteworthy that InfoIGL is relatively insensitive to λ_c and λ_s (gray and green), as adjusting their values does not cause significant fluctuations in InfoIGL's performance. While λ_i (pink) has a significant impact on the InfoIGL and requires fine-tuning. Furthermore, the results demonstrate the negative impact of excessively large values for λ_i in InfoIGL. For instance, the performance of InfoIGL on Motif (size) and CMNIST (color) drops significantly when λ_i exceeds 0.6. By comparing the performance of InfoIGL across different datasets under different hyperparameter settings, we can identify the optimal hyperparameters for each dataset. For example, λ_s ranging from 0.3 to 0.7 and λ_c ranging from 0.5 to 0.8 are more suitable hyperparameter values.

5 Conclusion

In this paper, we propose a novel approach InfoIGL to extract invariant representation for graph OOD generalization from the perspective of information theory. We view the mutual information between graphs as invariance and decompose the process of extracting invariant representation into two tasks: reducing redundant information and maximizing mutual information. The two tasks work together to ensure that the useful invariant features are fully isolated. Specifically, we utilize an attention mechanism to reduce redundancy. Then we jointly optimize the semantic-instance and instance-instance contrastive learning to maximize the mutual information, while also introducing instance constraint and hard negative mining to prevent model collapse. Theoretical analysis together with extensive experiments demonstrate the superiority of InfoIGL, highlighting its potential for real-world applications.

REFERENCES

Osama K Al-Shaykh and Russell M Mersereau. Lossy compression of noisy images. *IEEE Transactions on Image Processing*, 7(12):1641–1652, 1998.

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

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *ICLR*, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, MA KAILI, 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*.

Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, MA Kaili, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. Learning causally invariant representations for out-of-distribution generalization on graphs. *Advances in Neural Information Processing Systems*, 35:22131–22148, 2022.

Ching-Yao Chuang, R Devon Hjelm, Xin Wang, Vibhav Vineet, Neel Joshi, Antonio Torralba, Stefanie Jegelka, and Yale Song. Robust contrastive learning against noisy views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16670–16681, 2022.

Shaohua Fan, Xiao Wang, Chuan Shi, Peng Cui, and Bai Wang. Generalizing graph neural networks on out-of-distribution graphs. *arXiv preprint arXiv:2111.10657*, 2021.

Shaohua Fan, Xiao Wang, Yanhu Mo, Chuan Shi, and Jian Tang. Debiasing graph neural networks via learning disentangled causal substructure. *arXiv preprint arXiv:2209.14107*, 2022.

Alex Fout, Jonathon Byrd, Basir Shariat, and Asa Ben-Hur. Protein interface prediction using graph convolutional networks. *Advances in neural information processing systems*, 30, 2017.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. Good: A graph out-of-distribution benchmark. *arXiv preprint arXiv:2206.08452*, 2022.

Kartik Gupta, Thalaiyasingam Ajanthan, Anton van den Hengel, and Stephen Gould. Understanding and improving the role of projection head in self-supervised learning. *arXiv preprint arXiv:2212.11491*, 2022.

H Hafidi, M Ghogho, P Ciblat, and A Swami. Graphcl: Contrastive self-supervised learning of graph representations. arxiv 2020. *arXiv preprint arXiv:2007.08025*.

Takuo Hamaguchi, Hidekazu Oiwa, Masashi Shimbo, and Yuji Matsumoto. Knowledge transfer for out-of-knowledge-base entities: A graph neural network approach. *arXiv preprint arXiv:1706.05674*, 2017.

Xiaotian Han, Zhimeng Jiang, Ninghao Liu, and Xia Hu. G-mixup: Graph data augmentation for graph classification. *arXiv preprint arXiv:2202.07179*, 2022.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Tadaaki Hosaka and Yoshiyuki Kabashima. Statistical mechanical approach to lossy data compression: Theory and practice. *Physica A: Statistical Mechanics and its Applications*, 365(1):113–119, 2006.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.

Weiran Huang, Mingyang Yi, and Xuyang Zhao. Towards the generalization of contrastive self-supervised learning. *arXiv preprint arXiv:2111.00743*, 2021.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Dongkwan Kim and Alice Oh. How to find your friendly neighborhood: Graph attention design with self-supervision. *arXiv preprint arXiv:2204.04879*, 2022.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

Kezhi Kong, Guohao Li, Mucong Ding, Zuxuan Wu, Chen Zhu, Bernard Ghanem, Gavin Taylor, and Tom Goldstein. Robust optimization as data augmentation for large-scale graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 60–69, 2022.

Masanori Koyama and Shoichiro Yamaguchi. When is invariance useful in an out-of-distribution generalization problem? *arXiv preprint arXiv:2008.01883*, 2020.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.

Peter E Latham and Yasser Roudi. Mutual information. *Scholarpedia*, 4(1):1658, 2009.

Haoyang Li, Xin Wang, Ziwei Zhang, and Wenwu Zhu. Ood-gnn: Out-of-distribution generalized graph neural network. *arXiv preprint arXiv:2112.03806*, 2021.

Gang Liu, Tong Zhao, Jiaxin Xu, Tengfei Luo, and Meng Jiang. Graph rationalization with environment-based augmentations. *arXiv preprint arXiv:2206.02886*, 2022a.

Jiashuo Liu, Jiayun Wu, Jie Peng, Zheyan Shen, Bo Li, and Peng Cui. Distributionally invariant learning: Rationalization and practical algorithms. *arXiv preprint arXiv:2206.02990*, 2022b.

Jiajun Ma, Tianyang Hu, and Wenjia Wang. Deciphering the projection head: Representation evaluation self-supervised learning. *arXiv preprint arXiv:2301.12189*, 2023.

Siqi Miao, Mia Liu, and Pan Li. Interpretable and generalizable graph learning via stochastic attention mechanism. In *International Conference on Machine Learning*, pp. 15524–15543. PMLR, 2022.

Andrea Montanari and Basil N Saeed. Universality of empirical risk minimization. In *Conference on Learning Theory*, pp. 4310–4312. PMLR, 2022.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Zhen Peng, Wenbing Huang, Minnan Luo, Qinghua Zheng, Yu Rong, Tingyang Xu, and Junzhou Huang. Graph representation learning via graphical mutual information maximization. In *Proceedings of The Web Conference 2020*, pp. 259–270, 2020.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pp. 5171–5180. PMLR, 2019.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. Dropedge: Towards deep graph convolutional networks on node classification. *arXiv preprint arXiv:1907.10903*, 2019.

Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *ICLR*, 2020.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020.

Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27 (3):379–423, 1948.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Yongduo Sui, Xiang Wang, Jiancan Wu, Min Lin, Xiangnan He, and Tat-Seng Chua. Causal attention for interpretable and generalizable graph classification. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1696–1705, 2022.

Fan-Yun Sun, Jordan Hoffmann, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. *arXiv preprint arXiv:1908.01000*, 2019.

Kiran K Thekumparampil, Chong Wang, Sewoong Oh, and Li-Jia Li. Attention-based graph neural network for semi-supervised learning. *arXiv preprint arXiv:1803.03735*, 2018.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.

Rui Wang, Zuxuan Wu, Zejia Weng, Jingjing Chen, Guo-Jun Qi, and Yu-Gang Jiang. Cross-domain contrastive learning for unsupervised domain adaptation. *IEEE Transactions on Multimedia*, 2022.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Yiwei Wang, Wei Wang, Yuxuan Liang, Yujun Cai, and Bryan Hooi. Mixup for node and graph classification. In *Proceedings of the Web Conference 2021*, pp. 3663–3674, 2021.

Zihao Wang and Victor Veitch. A unified causal view of domain invariant representation learning. *arXiv preprint arXiv:2208.06987*, 2022.

Yingxin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. Discovering invariant rationales for graph neural networks. In *ICLR*, 2022.

Yongji Wu, Defu Lian, Yiheng Xu, Le Wu, and Enhong Chen. Graph convolutional networks with markov random field reasoning for social spammer detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 1054–1061, 2020.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Minghao Xu, Hang Wang, Bingbing Ni, Hongyu Guo, and Jian Tang. Self-supervised graph-level representation learning with local and global structure. In *International Conference on Machine Learning*, pp. 11548–11558. PMLR, 2021.

Hong Xuan, Abby Stylianou, Xiaotong Liu, and Robert Pless. Hard negative examples are hard, but useful. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pp. 126–142. Springer, 2020.

Shanmin Yang, Keren Fu, Xiao Yang, Ye Lin, Jianwei Zhang, and Cheng Peng. Learning domain-invariant discriminative features for heterogeneous face recognition. *IEEE Access*, 8:209790–209801, 2020.

Huaxiu Yao, Xinyu Yang, Allan Zhou, and Chelsea Finn. Multi-domain long-tailed learning by augmenting disentangled representations. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*.

Xufeng Yao, Yang Bai, Xinyun Zhang, Yuechen Zhang, Qi Sun, Ran Chen, Ruiyu Li, and Bei Yu. Pcl: Proxy-based contrastive learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7097–7107, 2022.

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. In *NeurIPS*, pp. 9240–9251, 2019.

Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto Sangiovanni Vincentelli. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13834–13844, 2021.

Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022a.

Shaofeng Zhang, Lyn Qiu, Feng Zhu, Junchi Yan, Hengrui Zhang, Rui Zhao, Hongyang Li, and Xiaokang Yang. Align representations with base: A new approach to self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16600–16609, 2022b.

Tong Zhang, Congpei Qiu, Wei Ke, Sabine Süsstrunk, and Mathieu Salzmann. Leverage your local and global representations: A new self-supervised learning strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16580–16589, 2022c.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pp. 7523–7532. PMLR, 2019.

A Related Work

Invariant learning for graph OOD. Invariant learning (Zhao et al., 2019; Rosenfeld et al., 2020; Yang et al., 2020; Wang & Veitch, 2022; Chen et al., 2022) has gained increasing attention as a powerful tool for developing robust representations capable of withstanding distribution shifts. In the context of graph data, invariant learning has been extensively explored to improve generalization performance in OOD scenarios. Recent works have demonstrated the effectiveness of diverse approaches in identifying invariant graph features, including optimization methods (Arjovsky et al., 2019; Krueger et al., 2021; Sagawa et al., 2020; Kong et al., 2022), causal learning (Wu et al., 2022; Sui et al., 2022; Liu et al., 2022a; Fan et al., 2022), stable learning (Fan et al., 2021; Li et al., 2021), and data manipulation (Miao et al., 2022; Rong et al., 2019; Wang et al., 2021; Han et al., 2022).

Contrastive learning and information theory. Contrastive learning (He et al., 2020; Chen et al., 2020a;b; Khosla et al., 2020) has achieved success in aligning representations by pulling together positive pairs and pushing apart negative pairs, which is tightly related to information theory, particularly mutual information. The contrastive loss can maximize the mutual information between positive pairs from the same class, promoting the intra-class compactness and inter-class discrimination (Wang et al., 2022). Some approaches (Huang et al., 2021; Zhang et al., 2022b; Yao et al., 2022) have attempted to leverage contrastive learning in domain generalization tasks, demonstrating the potential of contrastive learning for handling OOD issues.

B Theoretical Proofs and Discussions

B.1 Proofs.

Proof for Theorem 3. For a random variable X, its entropy function is defined as $H(X) = -\sum_i p(x_i)log_2 p(x_i)$, where $p(x_i)$ denotes the probability of X taking on the value x_i. Given a probability distribution $p(x_1), p(x_2), ..., p(x_n), \sum_{i=1}^n p(x_i) = 1$, we take the derivative of the entropy function with respect to the variable x_i:

$$p(x_j) = 1 - \sum_{i=1}^{n-1} p(x_i) \tag{14}$$

$$\frac{\partial H}{\partial p(x_i)} = -[1 + log_2(p(x_i) - 1 - log_2(p(x_j))] = -log(\frac{p(x_i)}{1 - p(x_i) - \sum_k^{k \neq i,j} p(x_k))}) \tag{15}$$

$$\frac{\partial^2 H}{\partial p(x_i)^2} = -\frac{1}{p(x_i)} - \frac{1}{1 - p(x_i) - \sum_k^{k \neq i,j} p(x_k)} < 0 \tag{16}$$

Equationequation 15 is the first derivative of entropy, and equation equation 16 is the second derivative. Since the second derivative is negative and $\underset{p(x_i)\to 0}{lim} \frac{\partial H}{\partial p(x_i)} \to +\infty$, $\underset{p(x_i)\to 1}{lim} \frac{\partial H}{\partial p(x_i)} \to -\infty$, we conclude that the entropy function is a strictly convex function. To obtain the maximum point of the convex function, we introduce Lagrange multipliers λ and construct the Lagrangian function as follows:

$$L(p(x_1), p(x_2), ..., p(x_n), \lambda) = -\sum_{i=1}^{n} p(x_i)log_2 p(x_i) + \lambda(\sum_{i=1}^{n} p(x_i) - 1) \tag{17}$$

Taking the first partial derivative of the Lagrangian function and setting it to zero yield:

$$\frac{\partial L}{\partial p(x_i)} = \frac{\partial}{\partial p(x_i)}[-\sum_{i=1}^{n} p(x_i)log_2 p(x_i) + \lambda(\sum_{i=1}^{n} p(x_i) - 1)] \tag{18}$$

$$\Rightarrow \lambda = log_2 p(x_i) + 1, p(x_i) = 2^{\lambda - 1} \tag{19}$$

$$p(x_1) = p(x_2) = ... = p(x_n) = \frac{1}{n} \tag{20}$$

This means that given probability distribution, the entropy achieves its maximum when the distribution is uniform distribution. Conversely, if significant differences exist among the value $p(x_i)$, the entropy will be lower.

The convexity of the entropy function is an important property in information theory, as it gives us insight into how to optimize $H(X)$ based on its probability distribution. Then if we assign weights to change the probability distribution of graph features, the entropy of the graph representation $H(\Phi(G))$ can be optimized to a smaller value.

Proof for Proposition 1. Given variables X_1, X_2 and noise S_1, S_2, X is independent of S, we define the mutual information between the composition of X and S as $I(X_1, S_1; X_2, S_2)$. According to the chain rule for information, the mutual information $I(X_1, S_1; X_2, S_2)$ can be transformed as follows:

$$I(X_1, S_1; X_2, S_2) = I(S_1; X_2, S_2|X_1) + I(X_1; X_2, S_2|S_1) \tag{21}$$

Since X is independent of S, the mutual information between X and S equals zero, that means: $I(X_1; S_2) = I(X_2; S_1) = I(X_1; S_1) = I(X_2; S_2) = 0$. So $I(S_1; X_2, S_2|X_1) = I(S_1; S_2|X_1)$, $I(X_1; X_2, S_2|S_1) = I(X_1; X_2|S_1)$, then we can simplify $I(X_1, S_1; X_2, S_2)$ as follows:

$$\begin{aligned} I(X_1, S_1; X_2, S_2) &= I(S_1; S_2|X_1) + I(X_1; X_2|S_1) \\ &= I(S_1; S_2) + I(X_1; X_2) \geq I(X_1; X_2) \end{aligned} \tag{22}$$

Since $I(X_1, S_1; X_2, S_2) = I(X_1, X_2) + (S_1; S_2)$, we can conclude that directly calculating the mutual information between variables with noise $I(X_1, S_1; X_2, S_2)$ may capture spurious correlations $I(S_1; S_2)$.

B.2 Discussions.

Explanation for Proposition 2. Under L2 normalization, the Euclidean distance is equivalent to the negative of cosine similarity:

$$x^\top y = 1 - \frac{1}{2}||x - y||_2^2 \tag{23}$$

We define $\{\mathbf{x}_j\}_{j=1}^K$ as the negative samples for instance $\mathbf{x}_i$. Then the contrastive loss for negative samples can be defined as:

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N} -\log\frac{1}{\sum_j \exp(\mathbf{x}_i^\top \mathbf{x}_j/\tau)} &= \frac{1}{N}\sum_{i=1}^{N}\log e^{1/\tau}(\sum_{j=1}^{K}\exp(-||\mathbf{x}_i - \mathbf{x}_j||_2^2/2\tau) \\ &= \frac{1}{N}\sum_{i=1}^{N}\log(\sum_{j=1}^{K}\exp(-||\mathbf{x}_i - \mathbf{x}_j||_2^2/2\tau) + \alpha_0 \end{aligned} \tag{24}$$

The loss for negative samples aims to encourage uniformity (Wang & Isola, 2020) and will encourage the distribution of samples close to uniform distribution (Ma et al., 2023). According to Theorem 3, the uniformity of probability distribution will cause entropy to increase, which violates our task I. Thus we apply contrastive learning in another space $\mathbf{z}_G$ that is mapped by a projection head (*i.e.,* $\mathbf{z}_G = \sigma(\text{MLP}(\mathbf{h}_G))$). Due to the non-linear mapping property of MLP, the space $\mathbf{z}_G$ and the space $\mathbf{h}_G$ are in the different unit balls. Since uniformity is encouraged in the output of the projection head (*i.e.,* $\mathbf{z}_G$), the problem of entropy increase in original representation $\mathbf{h}_G$ can be alleviated.

Explanation for Proposition 3. For class c, given the instance embeddings $\{\mathbf{z}_{G_i}\}_{i=i}^{N_c}$ and the corresponding semantic $\mathbf{w}_c = \sum_{i=1}^{N_c}(m_i \cdot \mathbf{z}_{G_i})$. Based on the properties of multivariate mutual information, the mutual information $I(\sum_{i=1}^{N_c}(m_i \cdot \mathbf{z}_{G_i}); \mathbf{z}_{G_1}; ...; \mathbf{z}_{G_{N_c}})$ can be transformed as follows:

$$I(\sum_{i=1}^{N_c}(m_i \cdot \mathbf{z}_{G_i}); \mathbf{z}_{G_1}; ...; \mathbf{z}_{G_{N_c}}) = I(\mathbf{z}_{G_1}; ...; \mathbf{z}_{G_{N_c}}) - I(\mathbf{z}_{G_1}; ...; \mathbf{z}_{G_{N_c}} | \sum_{i=1}^{N_c}(m_i \cdot \mathbf{z}_{G_i})) \tag{25}$$

Since $\sum_{i=1}^{N_c}(m_i \cdot \mathbf{z}_{G_i})$ contains all the variables $\mathbf{z}_{G_i}$ and provides some knowledge about their joint distribution, so the conditional mutual information $I(\mathbf{z}_{G_1}; ...; \mathbf{z}_{G_{N_c}} | \sum_{i=1}^{N_c}(m_i \cdot \mathbf{z}_{G_i}))$ is non-negative. Then we can conclude:

$$I(\sum_{i=1}^{N_c}(m_i \cdot \mathbf{z}_{G_i}); \mathbf{z}_{G_1}; \mathbf{z}_{G_2}; ...; \mathbf{z}_{G_{N_c}}) \leq I(\mathbf{z}_{G_1}; \mathbf{z}_{G_2}; ...; \mathbf{z}_{G_N}) \tag{26}$$

Therefore, the mutual information between the semantic $\mathbf{w}_c = \sum_{i=1}^{N_c}(m_i \cdot \mathbf{z}_{G_i})$ and all graph embedding $\mathbf{z}_{G_i}$ is always less than or equal to the mutual information between $\mathbf{z}_{G_i}$: $I(\mathbf{w}_c; \mathbf{z}_{G_1}; \mathbf{z}_{G_2}; ...; \mathbf{z}_{G_{N_c}}) \leq I(\mathbf{z}_{G_1}; \mathbf{z}_{G_2}; ...; \mathbf{z}_{G_{N_c}})$.

C Training Procedure

C.1 The Training Algorithm

The overall training procedure of the proposed InfoIGL is summarized in Algorithm 1.

Algorithm 1: InfoIGL's learning algorithm.

```
Input: Training graph dataset $\mathcal{D}^{tr}$, hyperparameters $\lambda_s, \lambda_i, \lambda_c$, number of epochs $e$, batch size $b$.
Output: graph encoder $\Phi(\cdot)$ (GNN encoder and Attention mechanism), classifier $\theta(\cdot)$.
Initialize graph encoder $\Phi(\cdot)$, Projection head $f_{\theta_p}(\cdot)$, classifier$\theta(\cdot)$.
for epoch in $1, 2, ...e$ do
    Sample data batches $\mathcal{B} = \mathcal{D}^1, \mathcal{D}^2, ..., \mathcal{D}^K$ with batch size $b$ from $\mathcal{D}^{tr}$.
    for $D^i \leftarrow \{(G_k, \mathbf{y}_k)\}_{k=1}^{b} \subset \mathcal{B}, i \in 1...K$ do
        # Reduce redundant information
        Calculate $\mathbf{h}_{G_k} \leftarrow \Phi(G_k)$ .
        # Maximize mutual information
        Calculate $\mathbf{z}_{G_k} = f_{\theta_p}(\mathbf{h}_{G_k})$.                    # Projection head
        Aggregate $\mathbf{z}_{G_k}$ to $\mathbf{w}_c$.                                # Get category semantics
        Calculate the semantic-wise contrastive loss $\mathcal{L}_{\text{sem}}$.
        Calculate $\mathbf{z}'_{G_k} = \lambda_c \mathbf{z}_{G_k} + (1 - \lambda_c)\mathbf{w}_c$.     # Constrain instance embedding
        Obtain hard negative samples $\mathbf{z}'_{hard}$.                          #hard negative mining
        Calculate the instance-wise contrastive loss $\mathcal{L}_{\text{ins}}$.
        # Transfer to downstream tasks
        Calculate the prediction loss $\mathcal{L}_{\text{pred}}$.
        $\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_s \mathcal{L}_{\text{sem}} + \lambda_i \mathcal{L}_{\text{ins}}$.         # Total loss
    end for
end for
Update all the trainable parameters to minimize $\mathcal{L}$
```

D Experimental Details

D.1 Dataset Description

We adopt one synthetic dataset (Motif) and three real-world datasets (HIV, Molbbbp, CMNIST) to evaluate our model. Here we provide the introduction and statistical details for datasets.

- **Motif (Gui et al., 2022)** is a synthetic dataset motivated by Spurious-Motif (Ying et al., 2019), graphs of which are generated by connecting a base graph (wheel, tree, ladder, star, and path) and a motif (house, cycle, and crane). The motifs are invariant for prediction while the base graphs may cause distribution shifts. Here, we use the concept shift of "base" and the covariate shift of "size" to create testing datasets.
- **HIV (Gui et al., 2022) and Molbbbp (Hu et al., 2020)** are small-scale molecular datasets adapted from MoleculeNet (Wu et al., 2018) in the real world, where atoms serve as nodes and chemical bonds serve as edges. The data splits of the two datasets are designed based on two domain selections: "scaffold" and "size". "Scaffold" selection is based on the Bemis-Murcko scaffold, which represents the two-dimensional structural foundation of a molecule, while "size" selection involves the number of nodes in a molecular graph - an inherent structural feature of a graph. As neither feature should determine the label, both can contribute to significant distribution shifts. We select OOD testing datasets based on the concept shift of "size" and "scaffold".
- **CMNIST (Gui et al., 2022)** is a real-world dataset created by applying superpixel techniques to handwritten digits. The dataset is divided into two parts: covariate shift split and concept shift split. In the covariate shift split, digits are colored with seven different colors, with the first five colors, the sixth color, and the seventh color assigned to the training, validation, and test sets, respectively. In the concept shift split, digits are colored with 10 different colors, with each color being highly correlated with one digit label in the training set. However, colors have weak or no correlations

with labels in the validation and test sets. Here we use the covariate shift of "color" as the OOD testing dataset.

Statistics of these datasets are presented in Table 4

Table 4: Statistics of multiple datasets.

Dataset	**#Train**	**#Val**	**#Test**	**#Classes**	**#Metrics**
Motif-size	18000	3000	3000	3	ACC
Motif-base	12600	6000	6000	3	ACC
HIV-size	14454	9956	10525	2	ROC-AUC
HIV-scaffold	15209	9365	10037	2	ROC-AUC
Molbbbp-size	1631	204	204	2	ROC-AUC
Molbbbp-scaffold	1631	204	204	2	ROC-AUC
CMNIST-color	42000	14000	14000	10	ACC

D.2 Compared Baselines

We compare our method against several state-of-the-art methods that are designed for graph OOD generalization. These baselines can be classified into four categories:

- **Optimization methods** aim to design optimization objectives to enhance the robustness of the model across different environments, including ERM, IRM (Arjovsky et al., 2019), VREx (Krueger et al., 2021), GroupDRO (Sagawa et al., 2020), FLAG (Kong et al., 2022). Empirical Risk Minimization(ERM) is a principle used in machine learning for selecting models. It involves selecting the model that minimizes the empirical risk, which is the average loss over the training data. IRM (Arjovsky et al., 2019) aims to find data representations that have consistent performance across all environments by discouraging feature distributions that lead to different optimal linear classifiers for each environment. VREx (Krueger et al., 2021) focuses on both covariate robustness and invariant prediction, and reduces risk variance in test environments by minimizing the variances of risks in training environments. GroupDRO (Sagawa et al., 2020) uses fair optimization to address the issue of insufficient training for minority distributions. Additionally, it employs risk interpolation, which involves minimizing the loss in the worst training environment explicitly. FLAG (Kong et al., 2022) is achieved by applying gradient-based adversarial perturbations to the node features, which helps the model maintain robustness to fluctuations in input data and improve prediction performance.
- **Causal learning** utilizes causal theory to extract causal features that play a key role in model prediction and ignore non-causal features, including DIR (Wu et al., 2022), CAL (Sui et al., 2022), GREA (Liu et al., 2022a), CIGA (Chen et al.), Disc (Fan et al., 2022); DIR (Wu et al., 2022) chooses a subset of graph representations as causal rationales, and performs interventional data augmentation to generate multiple distributions. CAL (Sui et al., 2022), GREA (Liu et al., 2022a) and DisC (Fan et al., 2022) are designed based on the Structural Causal Models (SCM) to disentangle causal subgraphs. While CIGA (Chen et al.) maximizes the mutual information between causal graphs and labels to extract causal-invariant subgraphs.
- **Stable learning** is committed to independently extracting stable features across different environments by means of sample reweighting, such as StableGNN (Fan et al., 2021), OOD-GNN (Li et al., 2021); StableGNN (Fan et al., 2021) proposes a novel causal variable-distinguishing regularizer which can eliminate correlations between each pair of high-level variables through a set of sample weights while OOD-GNN (Li et al., 2021) can eliminate statistical dependencies between variant and invariant representations by using randomized Fourier features and sample reweighting in a nonlinear graph representation.
- **Data manipulation** is dedicated to generating diverse augmented data to increase the diversity of data distribution, such as GSAT (Miao et al., 2022), DropEdge (Rong et al., 2019),

M-Mixup (Wang et al., 2021), G-Mixup (Han et al., 2022). GSAT (Miao et al., 2022) injects randomness to prevent unrelated information from interfering with the labels, and utilizes the reduction in randomness to select subgraphs relevant to the labels. DropEdge (Rong et al., 2019), M-Mixup (Wang et al., 2021) and G-Mixup (Han et al., 2022) apply various transformations to augment the graph data which is beneficial to realize graph generalization.

For ERM, IRM (Arjovsky et al., 2019), VREx (Krueger et al., 2021), GroupDRO (Sagawa et al., 2020), DIR (Wu et al., 2022), and M-Mixup (Wang et al., 2021), we obtain their performance on diverse datasets following the prior research (Gui et al., 2022). While for other baseline implementations, we use the codes released by the authors and evaluate their performance on a variety of datasets.

Additionally, **classic graph contrastive learning methods** which incorporate information theory, including CNC (Zhang et al., 2022a), GMI (Peng et al., 2020), Infograph (Sun et al., 2019), GraphCL (Hafidi et al.), are discussed as benchmarks. CNC (Zhang et al., 2022a) utilized ERM to identify samples with different spurious features and then leveraged contrastive learning to learn similar representations for samples of the same class. GMI (Peng et al., 2020) directly derived MI by comparing the input (i.e., the sub-graph consisting of the input neighborhood) and the output (i.e., the hidden representation of each node) of the encoder. Infograph (Sun et al., 2019) enhanced the alignment of graph representations by maximizing the mutual information between representations learned by InfoGraph and that acquired by existing supervised methods. GraphCL (Hafidi et al.) learned node embeddings by maximizing the similarity between representations of two randomly perturbed versions of the same node.

D.3 Details of Hyperparameters for InfoIGL

Our code is implemented based on PyTorch Geometric. For all the experiments, we use the Adam optimizer, where the initial learning rate and the minimum learning rate are searched within $\{0.01, 0.001, 0.0001\}$ and $\{0.001, 0.00001, 0.000001\}$, respectively. We select embedding dimensions from $\{32, 64, 128, 300\}$ and choose batch sizes from $\{64, 128, 256, 512, 1024\}$. The dropout ratio is searched within $\{0.1, 0.3, 0.5\}$ while $\lambda_c, \lambda_s, \lambda_i$ are searched within $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. We adopt grid search to tune the hyperparameters and list the details of hyperparameters for InfoIGL in Table 5.

Table 5: The hyperparameters for InfoIGL on different datasets

hyperparameters	Motif		HIV		Molbbbp		CMNIST
	size	base	size	scaffold	size	scaffold	color
encoder layers	3	3	3	3	2	2	5
embedding dim	64	128	300	128	128	300	32
max epoch	200	200	100	200	100	100	150
pretrain	40	40	80	40	20	80	60
batch size	1024	128	256	1024	64	1024	256
ini-lr	0.001	0.001	0.001	0.01	0.01	0.0001	0.001
min-lr	1e-3	1e-6	1e-6	1e-6	1e-6	1e-6	1e-3
weight decay	0	1e-1	1e-2	1e-5	1e-5	0	0
drop ratio	0.5	0.5	0.3	0.3	0.3	0.3	0.5
λ_c	0.7	0.7	0.7	0.7	0.2	0.7	0.7
λ_s	0.8	0.5	0.5	0.5	0.2	0.2	0.5
λ_i	0.2	0.5	0.5	0.1	0.2	0.2	0.1

D.4 Results of Sensitivity Analysis

We conduct experiments to evaluate how sensitive is InfoIGL to the choice of graph neural network architectures (GCN, GIN, and GAT). The results are listed in Table 6. As shown in Table 6, InfoIGN-GCN, InfoIGL-GIN, and InfoIGL-GAT are competent on Motif and CMNIST datasets while they far surpass the baseline ERM. The results demonstrate the effectiveness of our method, irrespective of the choice of GNN backbones.

Table 6: Results of experiments with different backbones.

methods	Motif		CMNIST
	size	base	color
ERM	$70.75_{\pm 0.56}$	$81.44_{\pm 0.45}$	$28.60_{\pm 1.87}$
InfoIGL-GCN	$86.53_{\pm 2.15}$	$91.56_{\pm 0.91}$	$38.30_{\pm 0.76}$
InfoIGL-GIN	$85.53_{\pm 2.37}$	$92.51_{\pm 0.16}$	$38.93_{\pm 1.11}$
InfoIGL-GAT	$84.66_{\pm 1.23}$	$90.32_{\pm 1.45}$	$37.51_{\pm 2.07}$

Additionally, we perform experiments to access the sensitivity of InfoIGL to its hyperparameters (*i.e.,* $\lambda_c, \lambda_s, \lambda_i$). In the main body of the paper, we only presented a line graph for sensitivity analysis, here we list the detailed results of sensitivity analysis on $\lambda_c, \lambda_s, \lambda_i$ in Table 7, Table 8, Table 9, respectively.

Each column of the table represents the experimental results of the model with varying ratios on different datasets. From the data in each column of the table, it can be observed that the model results vary slightly across different ratios, indicating that the model is not sensitive to these hyperparameters.

Table 7: Performance of InfoIGL with different constraint ratio λ_c.

ratio	Motif		HIV		Molbbbp		CMNIST
	size	base	size	base	scaffold	scaffold	color
0.1	$82.53_{\pm 2.75}$	$90.46_{\pm 1.13}$	$87.73_{\pm 2.40}$	$70.89_{\pm 2.12}$	$83.07_{\pm 1.24}$	$79.44_{\pm 3.24}$	$38.34_{\pm 1.78}$
0.2	$84.13_{\pm 1.79}$	$89.86_{\pm 0.70}$	$89.42_{\pm 2.60}$	$70.47_{\pm 2.12}$	$83.39_{\pm 2.76}$	$80.45_{\pm 3.22}$	$37.13_{\pm 3.37}$
0.3	$84.18_{\pm 2.05}$	$89.66_{\pm 1.10}$	$89.55_{\pm 2.98}$	$71.16_{\pm 1.41}$	$81.38_{\pm 1.65}$	$82.34_{\pm 1.04}$	$37.95_{\pm 2.78}$
0.4	$84.23_{\pm 1.65}$	$89.84_{\pm 0.76}$	$90.84_{\pm 2.31}$	$68.97_{\pm 3.40}$	$82.93_{\pm 1.71}$	$72.56_{\pm 0.51}$	$36.49_{\pm 2.22}$
0.5	$83.62_{\pm 2.38}$	$89.21_{\pm 1.68}$	$91.18_{\pm 1.38}$	$69.81_{\pm 2.50}$	$81.18_{\pm 2.86}$	$72.69_{\pm 2.96}$	$39.05_{\pm 2.05}$
0.6	$83.48_{\pm 1.80}$	$90.67_{\pm 0.69}$	$91.36_{\pm 1.50}$	$69.64_{\pm 2.64}$	$82.03_{\pm 1.37}$	$73.70_{\pm 2.02}$	$38.92_{\pm 1.52}$
0.7	$83.65_{\pm 1.05}$	$90.39_{\pm 0.83}$	$92.55_{\pm 1.02}$	$69.60_{\pm 2.27}$	$79.98_{\pm 3.01}$	$77.36_{\pm 2.02}$	$38.51_{\pm 3.22}$
0.8	$85.05_{\pm 3.30}$	$90.86_{\pm 0.87}$	$92.75_{\pm 0.84}$	$71.02_{\pm 1.81}$	$80.07_{\pm 2.91}$	$80.14_{\pm 1.24}$	$38.01_{\pm 1.70}$
0.9	$82.67_{\pm 2.44}$	$90.86_{\pm 0.77}$	$93.23_{\pm 1.01}$	$71.04_{\pm 2.54}$	$80.12_{\pm 3.25}$	$75.71_{\pm 1.27}$	$33.02_{\pm 2.85}$

Table 8: Performance of InfoIGLL with different semantic loss ratio λ_s

ratio	Motif		HIV		Molbbbp		CMNIST
	size	based	size	scaffold	size	scaffold	color
0.1	$81.39_{\pm 2.84}$	$90.28_{\pm 0.57}$	$92.24_{\pm 1.26}$	$69.61_{\pm 1.16}$	$80.56_{\pm 2.51}$	$74.89_{\pm 2.30}$	$37.05_{\pm 2.15}$
0.2	$81.17_{\pm 3.40}$	$90.83_{\pm 1.00}$	$92.09_{\pm 1.07}$	$71.16_{\pm 1.77}$	$81.43_{\pm 1.98}$	$75.38_{\pm 0.89}$	$38.14_{\pm 2.43}$
0.3	$83.85_{\pm 3.10}$	$90.57_{\pm 0.81}$	$92.14_{\pm 0.65}$	$70.72_{\pm 0.81}$	$80.81_{\pm 2.41}$	$75.38_{\pm 2.45}$	$36.64_{\pm 1.87}$
0.4	$84.64_{\pm 1.91}$	$90.19_{\pm 0.73}$	$92.25_{\pm 0.89}$	$71.30_{\pm 0.96}$	$82.05_{\pm 3.13}$	$76.15_{\pm 2.56}$	$37.90_{\pm 0.64}$
0.5	$81.93_{\pm 2.98}$	$90.36_{\pm 0.36}$	$92.44_{\pm 1.16}$	$70.60_{\pm 0.92}$	$80.45_{\pm 2.12}$	$77.26_{\pm 2.11}$	$38.12_{\pm 0.49}$
0.6	$92.71_{\pm 0.95}$	$70.37_{\pm 0.82}$	$83.80_{\pm 3.73}$	$87.87_{\pm 3.28}$	$80.61_{\pm 4.34}$	$77.01_{\pm 1.54}$	$37.30_{\pm 2.82}$
0.7	$83.17_{\pm 3.54}$	$90.80_{\pm 0.42}$	$91.90_{\pm 0.51}$	$70.42_{\pm 0.54}$	$80.76_{\pm 2.64}$	$76.62_{\pm 0.60}$	$34.72_{\pm 3.41}$
0.8	$84.06_{\pm 2.44}$	$90.61_{\pm 1.15}$	$92.21_{\pm 0.81}$	$70.54_{\pm 1.84}$	$81.38_{\pm 2.11}$	$75.12_{\pm 1.64}$	$37.55_{\pm 0.46}$
0.9	$84.69_{\pm 4.12}$	$90.47_{\pm 0.27}$	$91.96_{\pm 0.73}$	$71.74_{\pm 1.67}$	$77.89_{\pm 3.99}$	$75.44_{\pm 2.20}$	$36.50_{\pm 2.77}$

Table 9: Performance of InfoIGL with different instance loss ratio λ_i.

ratio	Motif		HIV		Molbbbp		CMNIST
	size	base	size	scaffold	size	scaffold	color
0.1	82.57±0.57	90.74±0.76	91.86±0.69	62.03±4.63	83.81±1.75	71.12±1.79	38.02±0.61
0.2	84.89±1.41	90.30±0.77	92.57±0.76	60.13±3.68	77.24±1.78	74.71±2.17	34.51±1.53
0.3	84.79±2.08	90.61±0.44	92.26±1.23	56.67±1.13	78.44±3.11	68.15±2.88	38.28±1.75
0.4	85.46±1.63	90.04±1.58	92.41±0.68	61.77±2.81	77.90±2.08	68.04±5.93	34.61±0.75
0.5	82.87±2.60	90.11±0.42	92.25±1.21	55.93±3.79	79.73±3.74	62.76±2.78	32.69±3.57
0.6	83.64±1.39	90.66±1.03	92.42±0.88	61.33±3.45	77.42±2.22	60.60±4.97	36.48±1.80
0.7	83.17±1.30	90.82±0.76	92.08±1.09	57.09±2.29	71.17±6.02	57.02±2.42	29.61±5.93
0.8	83.89±1.44	91.10±0.77	92.38±0.61	58.92±2.68	76.11±2.48	60.43±3.95	30.75±1.41
0.9	80.71±1.60	91.17±0.37	92.09±0.45	59.31±2.57	73.66±4.99	63.73±4.49	27.09±5.08

E Time and Space Complexity Analysis

Let N denote the number of graphs, n denote the average node number per graph, l_G, l_A, l_P and d_G, d_A, d_P denote the numbers of layers and the embedding dimensions in the GNN backbone, attention mechanism and projection head, respectively, C denote the number of class and K denote the number of hard negative samples per instance. The time complexity of the GNN backbone is $O(Nnl_Gd_G)$. For the attention mechanism, the time complexity is $O(Nnl_Ad_A)$. For the projection head, since it turns from node level to graph level, the time complexity is $O(Nl_Pd_P)$. For semantic-wise contrastive learning, the time complexity is $O(NC)$. For instance-wise contrastive learning, the time complexity is $O(NK)$. Therefore, the time complexity of the whole model is $O(N(nl_Gd_G + nl_Ad_A + nl_Pd_P) + C + K)$, and the order of magnitude is $O(Nn)$. Similarly, the space complexity is also approximately $O(Nn)$ which is about the same as baselines.

F Limitations and Potential Negative Social Impacts

F.1 Limitations

Informal hard negative mining. There are several techniques for generating hard negative samples in machine learning. One approach is to choose negative samples that closely resemble positive examples by sampling from a pool of negatives. These selected samples can be relabeled as "hard negatives" and included in the training process. Another method involves the use of sophisticated algorithms like online hard example mining (OHEM), which identifies challenging negative samples based on their loss values during training. However, instead of these methods, we select hard negative samples by computing the distance between the negative samples and the semantic center that corresponds to the positive sample. While this informal hard negative mining technique may conserve computational resources, it could also introduce a certain degree of error.

Lack of testing the applicability to other tasks. Contrastive learning is typically used to extract useful features that can be transferred to downstream tasks. Therefore, the InfoIGL framework, which employs contrastive learning techniques, has the potential to improve the performance of other downstream tasks such as node classification. The applicability of InfoIGL to other tasks is what we leave for future work.

F.2 Potential Negative Social Impacts

As our framework can extract the invariant representations with information theory and realize reliable GNNs which can alleviate graph OOD problems, we are confident that the benefits of our work outweigh any potential negative impacts. However, the trustworthy GNN may cause over-reliance and be widely applied in the real world, leading to decreased human ability and consequent unemployment. Additionally, excessive trust in this technology may cause significant losses due to potential error overlooking. Therefore, it is crucial that people use the framework prudently and increase supervision when applying it.