Negative as Positive: Enhancing Out-of-distribution Generalization for Graph Contrastive Learning

Zixu Wang
CAS Key Laboratory of AI Safety,
Institute of Computing Technology,
Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
wangzixu22s@ict.ac.cn

Bingbing Xu
CAS Key Laboratory of AI Safety,
Institute of Computing Technology,
Chinese Academy of Sciences
Beijing, China
xubingbing@ict.ac.cn

Yige Yuan
CAS Key Laboratory of AI Safety,
Institute of Computing Technology,
Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
yuanyige20z@ict.ac.cn

Huawei Shen
CAS Key Laboratory of AI Safety,
Institute of Computing Technology,
Chinese Academy of Sciences
Beijing, China
shenhuawei@ict.ac.cn

Xueqi Cheng
CAS Key Laboratory of AI Safety,
Institute of Computing Technology,
Chinese Academy of Sciences
Beijing, China
cxq@ict.ac.cn

ABSTRACT

Graph contrastive learning (GCL), standing as the dominant paradigm in the realm of graph pre-training, has yielded considerable progress. Nonetheless, its capacity for out-of-distribution (OOD) generalization has been relatively underexplored. In this work, we point out that the traditional optimization of InfoNCE in GCL restricts the cross-domain pairs only to be negative samples, which inevitably enlarges the distribution gap between different domains. This violates the requirement of domain invariance under OOD scenario and consequently impairs the model's OOD generalization performance. To address this issue, we propose a novel strategy "Negative as Positive", where the most semantically similar cross-domain negative pairs are treated as positive during GCL. Our experimental results, spanning a wide array of datasets, confirm that this method substantially improves the OOD generalization performance of GCL.

CCS CONCEPTS

• **Computing methodologies → Machine learning**.

KEYWORDS

Graph Representation Learning; Graph OOD Generalization; Graph Contrastive Learning

ACM Reference Format:
Zixu Wang, Bingbing Xu, Yige Yuan, Huawei Shen, and Xueqi Cheng. 2024. Negative as Positive: Enhancing Out-of-distribution Generalization for Graph Contrastive Learning. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3626772.3657927

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07.
https://doi.org/10.1145/3626772.3657927

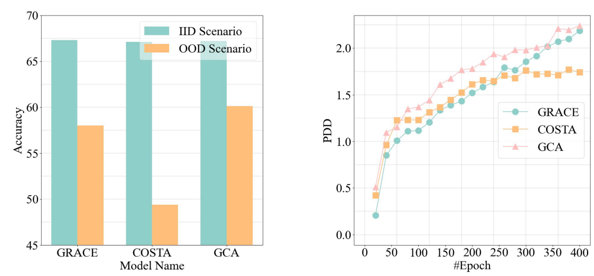

Figure 1: Left: Traditional GCLs perform badly under OOD scenario compared to IID one. Right: Pairwize-Domain-Discrepancy grows during GCL.

1 INTRODUCTION

Graph Contrastive Learning (GCL) with supervised fine-tuning has emerged as the dominant paradigm for graph pre-training, exhibiting remarkable performance across diverse downstream tasks while requiring only a limited amount of labeled data[7, 11, 15, 19, 20, 25, 26, 31, 32, 36, 37]. Generally, GCL aims at training a graph encoder that maximizes the mutual information between instances with similar semantic information via augmentation.

Most existing works assume the pre-text graph and downstream graph are independent and identically distributed (IID)[36, 37]. However, the graph in the downstream task often exhibits an out-of-distribution (OOD) pattern compared to that encountered in pre-text task[3, 4, 13, 16, 28, 29, 33, 35]. Furthermore, we find that current methods perform poorly on the OOD downstream graph than IID ones, as shown on the left side of Fig. 1.

To delve into the phenomenon mentioned above, we utilize pairwise domain discrepancy (PDD), which is widely used in prior works[10, 14, 17, 21] to measure the model's OOD generalization capability. PDD describes the average distance between domain centers in the embedding space. As shown on the right side of Fig. 1, PDD gradually increases during GCL training, aligning with the declined performance under the OOD scenario. Through in-depth analysis (details in Sec. 3.1), we argue that the model's reduced generalization capability stems from treating cross-domain pair as a negative sample solely in the traditional GCL paradigm. By aiming to reduce negative sample similarity in InfoNCE[18], domains are pushed further apart, resulting in increased PDD and poor OOD generalization performance.

Motivated by the above analysis, we propose **N**egative **as** **P**ositive, namely **NaP**, to enhance the OOD generalization of GCL. Specifically, considering that the embedding of nodes represents its semantics, NaP dynamically transfers a subset of cross-domain negative samples as positive samples based on the embedding similarity, and reduces the distance of positive samples. Therefore, NaP can narrow the distribution gap among embedding from different domains, further preserving domain-shared knowledge and enhancing OOD generalization. Extensive experiments on various datasets and tasks demonstrate the improved domain generalization capability of the proposed method compared to the SOTA GCL methods.

2 PRELIMINARIES

2.1 Task Formulation of OOD in GCL

Let $\mathcal{G} = (\mathbf{X}, \mathbf{A})$ denote a graph, where $\mathbf{X} \in \mathbb{R}^{N\times F}$ denotes the nodes' feature map, and $\mathbf{x}_i$ is the feature of node v_i. $\mathbf{A} \in \mathbb{R}^{N\times N}$ denotes the adjacency matrix, where $\mathbf{A}_{ij} = 1$ means v_i and v_j are connected. As Eq. 1 shows, GCL aims at training a GNN encoder[12, 23, 27, 30] $g_\theta(\mathcal{G})$ by maximizing the mutual information between instances with similar semantic information via augmentation. The augmented graph is noted as $\mathcal{G}_\psi$, where ψ represents one kind of augmentation method such as used in [5, 9, 36, 37],

$$\theta^* = \max_\theta \mathcal{I}\left(g_\theta(\mathcal{G}_\alpha), g_\theta(\mathcal{G}_\beta)\right) \tag{1}$$

The formulation of OOD in GCL is as follows: θ^* in Eq. 1 is optimized on data $\{(G^i)|_{i=1}^S\}$, and leveraged to infer G^T, with $P(G^T) \neq P((G^i)|_{i=1}^S)$, where S is the number of domains in pretraining. In contrast, within IID scenarios, $P(G^T) = P((G^i)|_{i=1}^S)$. Fig.1 shows the test accuracy for OOD and IID scenarios of a representative benchmark GOOD-Twitch, where each graph G^i is a gamer network and different domains represent the different languages used in the network. All three GCL methods[34, 36, 37] exhibit significant performance degradation in the presence of OOD, emphasizing the critical importance of investigating this phenomenon.

2.2 Pairwise Domain Discrepancy

Pairwise domain discrepancy(PDD) is widely used to measure the model's OOD generalization capability in prior works[10, 14, 17, 21]. It's the average distance among all pairs of the domains' centers. Denote the center embedding of domain d as $\bar{h^d} = \frac{1}{N_d}\sum_{i=1}^{N_d} \mathbf{H}_i^d$,

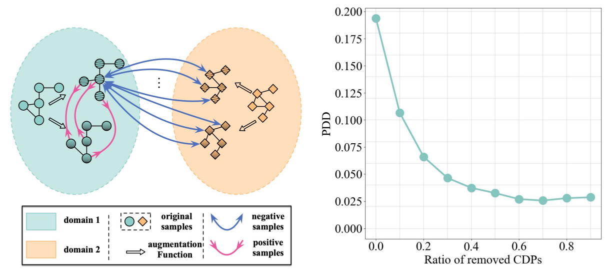

Figure 2: Left: All CDPs are negative samples. Right: PDD decreases while more CDPs are removed.

and PDD is as follows:

$$PDD = \frac{1}{\binom{P}{2}} \sum_{p,q|_{1\le p<q\le P}} ||\bar{h^p} - \bar{h^q}||, \tag{2}$$

where P denotes the number of domains, $\mathbf{H}_i^d$ denotes the embedding of i-th node in domain d and N_d denotes the number of nodes in domain d.

3 PROPOSED METHOD

In this section, we first show the motivation of NaP and then introduce each part of NaP in detail.

3.1 Motivation

The phenomenon of OOD is highly prevalent in GCL, which underscores the need to address OOD issues. Taking one common scenario as an example: in social networks, GCL may be trained on highly influential communities but applied to low-influence users [2]. This phenomenon is also common in areas such as financial risk prediction[1] (high-market-value companies VS medium-sized ones) and fraudulent accounts detection (old fraudulent style VS new ones). Such commonality highlights the critical need to address OOD in GCL. However, as shown in Fig. 1, the traditional GCLs perform poorly on OOD scenarios, and the PDD of all domains continues to increase during the training of GCLs. The increasing PDD indicates that GCL will widen the gap in domain distribution and push domains further apart, violating an ideal OOD generalization, which should capture the shared knowledge among different domains and facilitate the seamless transfer to unseen target domains.

Let **C**ross-**D**omain **P**air (CDP) represent two nodes from different domains. We argue that the principal constituents of negative samples for optimizing Eq. 1 are CDPs, being a significant factor in the poor OOD generalization capability. Specifically, as shown on the left side of Fig.2, CDPs can only be negative samples, and the traditional contrastive loss will decrease the similarity of negative samples, leading to the pushing-apart effect between the nodes in CDP. Furthermore, as shown on the right side of Fig. 2, the PDD of node embedding of GCL decreases as the ratio of removed CDP increases which proves that CDPs are harmful to GCL's OOD generalization. Therefore, the CDPs in traditional GCL tend to push the representations of samples from different domains apart, resulting in a higher PDD and a poor OOD generalization ability.

3.2 NaP: Negative as Positive

Based on the above motivation, we propose NaP, which transfers a subset of the most semantically similar negative samples as positive ones. Fig.3 illustrates the overall framework of NaP, including the encoding module and the objective module. Note that our NaP framework can be adapted to existing GCL methods that use InfoNCE as loss function, e.g., GRACE[36], GCA[37], and GraphCL[7].

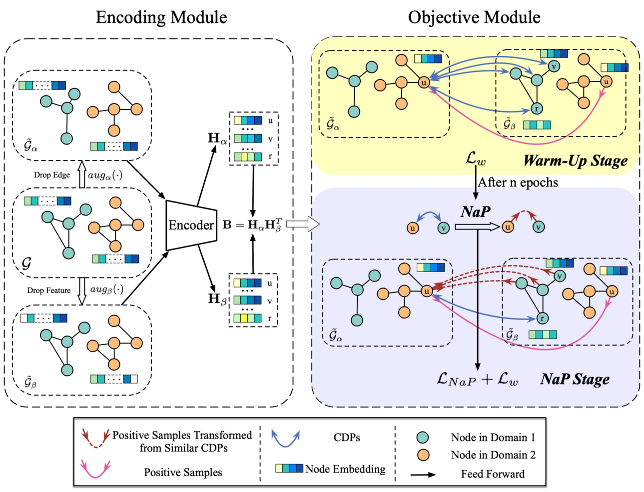

Figure 3: The overall framework of NaP consists of two modules: the encoding module and the objective module. The objective module comprises two stages: the warm-up stage and the NaP stage.

3.2.1 Encoding Module. The objective of this module is to obtain the embedding of each node. We first generate different views of $\mathcal{G}$ as $\tilde{\mathcal{G}_\alpha}$, $\tilde{\mathcal{G}_\beta}$ using graph augmentations. And input the augmented graphs into a shared GCN[12] encoder to get the embedding $\mathbf{H}_\alpha$, $\mathbf{H}_\beta$. The propagation of the l-th layer of GCN is represented as:

$$\mathbf{H}^{l+1} = \sigma(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^l\mathbf{W}^l), \tag{3}$$

where $\sigma(\cdot)$ is the activation function, $\tilde{\mathbf{A}}$ is the adjacency matrix with self-loop, $\tilde{\mathbf{D}}$ is the corresponding degree matrix and $\mathbf{W}$ is the parameter matrix.

3.2.2 Objective Module. Considering that the representations obtained from randomly initialized models may not accurately reflect the semantic information of the samples, we have to train the GCL in the traditional way for several epochs. Therefore, there are two stages in this module: Warm-up stage and NaP stage.

(1) Warm-Up Stage: Firstly, we use the traditional InfoNCE loss to train the GCL as the warm-up for the NaP stage. The InfoNCE loss for each positive pair $(v_{\alpha i}, v_{\beta i})$ in warm-up stage is:

$$\mathcal{L}_w = -\log \frac{\exp(\frac{\theta(v_{\alpha i}, v_{\beta i})}{\tau})}{\exp(\frac{\theta(v_{\alpha i}, v_{\beta i})}{\tau}) + \sum_{j\neq i}\exp(\frac{\theta(v_{\alpha i}, v_{\beta j})}{\tau}) + \sum_{j\neq i}\exp(\frac{\theta(v_{\alpha i}, v_{\alpha j})}{\tau})} \tag{4}$$

The $\theta(v_{\alpha i}, v_{\beta j})$ means cosine similarity between $\mathbf{H}_{\alpha i}$, $\mathbf{H}_{\beta j}$.

(2) NaP Stage: After n epochs warm-up, we enter the NaP stage where a subset of CDPs is chosen to transform into positive samples to mitigate the domain discrepancies introduced by CDPs. We select the most similar CDPs based on the between-view embedding similarity in the current epoch and transform the chosen CDPs into positive samples by adding a new loss item. Firstly, we compute the between-view-similarity matrix:

$$\mathbf{B} = \mathbf{H}_\alpha \mathbf{H}_\beta^T \tag{5}$$

We focus our attention on cross-domain samples, so we update $\mathbf{B}$ as follows:

$$\mathbf{B}_{ij} = 0 \text{ if } d_i = d_j \tag{6}$$

The d_i means the domain index of $v_i, i \in \{1, 2, ..., N\}$. After sorting the elements in $\mathbf{B}$, we can select the top r of most similar samples and their indices idx as follows:

$$\text{idx} = \arg \max_{I\subset\mathbb{R}^{N\times N}:|I|=r} \sum_{(i,j)\in I} \mathbf{B}_{ij} \tag{7}$$

To obtain the transformed CDPs, we set the mask matrix:

$$mask_{ij} = 1 \text{ if } (i, j) \in \text{idx else } 0 \tag{8}$$

Up to this point, only the top r most similar CDPs are retained in the mask. We add a new loss item to transform these CDPs into positive samples, namely $\mathcal{L}_{NaP}$:

$$\mathcal{L}_{NaP} = -\log \frac{\sum_{j\neq i} mask_{ij}\{\exp(\frac{\theta(v_{\alpha i}, v_{\beta j})}{\tau}) + \exp(\frac{\theta(v_{\alpha i}, v_{\alpha j})}{\tau})\}}{\exp(\frac{\theta(v_{\alpha i}, v_{\beta i})}{\tau}) + \sum_{j\neq i}\exp(\frac{\theta(v_{\alpha i}, v_{\beta j})}{\tau}) + \sum_{j\neq i}\exp(\frac{\theta(v_{\alpha i}, v_{\alpha j})}{\tau})} \tag{9}$$

Finally, for each positive pair $(v_{\alpha i}, v_{\beta i})$, the loss in NaP stage is written as below:

$$\mathcal{L} = \mathcal{L}_{NaP} + \mathcal{L}_w$$
$$= -\log \frac{\exp(\frac{\theta(v_{\alpha i}, v_{\beta i})}{\tau}) + \sum_{j\neq i} mask_{ij}\{\exp(\frac{\theta(v_{\alpha i}, v_{\beta j})}{\tau}) + \exp(\frac{\theta(v_{\alpha i}, v_{\alpha j})}{\tau})\}}{\exp(\frac{\theta(v_{\alpha i}, v_{\beta i})}{\tau}) + \sum_{j\neq i}\exp(\frac{\theta(v_{\alpha i}, v_{\beta j})}{\tau}) + \sum_{j\neq i}\exp(\frac{\theta(v_{\alpha i}, v_{\alpha j})}{\tau})} \tag{10}$$

To sum up, after n epochs of training according to the loss in Eq. 4, NaP selects the top r most similar CDPs based on the current epoch's embedding similarity. These CDPs are then treated as positive samples, and the training continues using the loss described in Eq. 10.

4 EXPERIMENTS

In this section, we empirically evaluate the quality of produced node embedding on node classification using two public benchmark datasets: GOOD benchmark and Facebook100.

4.1 Datasets

We use 3 datasets from GOOD benchmark[6] and 15 datasets from Facebook100[22] for experiments. Datasets from Facebook100 are social networks of 100 universities in the US. Each university is viewed as a domain and each node stands for a student or faculty.

4.2 Experimental Setup

4.2.1 Data settings. We divide the dataset according to GOOD[6]. Specifically, for the Facebook100, we randomly use 9 domains as the source domains for training, 1 domain (Emory) for validation, and 15 others for testing.

Table 1: Experiments results of all baselines and NaP. The bold font represents the top-1 performance and the underline represents the second performance across the self-supervised methods.

Dataset	Facebook100					GOOD benchmark		
	Santa	Wake	Bucknell	Colgate	Wesleyan	Twitch	CBAS	Cora
Domain	university					language	color	degree
DGI	87.08%	83.02%	89.24%	89.55%	88.52%	53.34%	<u>52.86%</u>	46.61%
GRACE	87.88%	82.70%	90.12%	82.09%	<u>90.80%</u>	58.00%	48.10%	50.85%
GCA	89.10%	82.71%	<u>93.01%</u>	91.18%	90.11%	60.14%	50.00%	<u>50.97%</u>
COSTA	89.93%	75.29%	91.46%	<u>91.52%</u>	88.36%	49.40%	45.24%	48.09%
BGRL	88.80%	<u>83.61%</u>	91.59%	85.18%	82.41%	**63.25%**	49.05%	40.63%
MVGRL	<u>90.12%</u>	78.58%	91.45%	88.38%	90.13%	53.98%	50.95%	47.15%
Ours	**91.06%**	**86.55%**	**93.26%**	**93.18%**	**91.51%**	<u>61.08%</u>	**53.33%**	**51.31%**
improve	+0.94%	+2.94%	+0.25%	+1.66%	+0.71%	-2.17%	+0.47%	+0.34%
GCN	92.10%	87.14%	94.47%	93.24%	92.10%	51.65%	65.24%	59.39%

4.2.2 Model and Metric settings. We use 6 contrastive methods: DGI, GRACE, GCA, COSTA, BGRL, MVGRL[8, 20, 24, 34, 36, 37] for self-supervised methods, and use GCN[12] as supervised baselines. The checkpoint for OOD testing is decided based on the result obtained from OOD validation domains. The reported results represent the average accuracy from three independent runs.

4.2.3 Results and Analysis.

(1) NaP surpasses baselines. As shown in the Table.1, NaP outperforms almost all GCL baselines. It is worth noting that NaP surpasses all four baselines - DGI, GRACE, GCA and COSTA[24, 34, 36, 37] - that use InfoNCE loss, with an improvement of up to 11.68%. Furthermore, NaP outperforms BGRL, which uses BYOL[20] as the loss function, and MVGRL, which uses JSD[8] as the loss function, on the majority of datasets. Last but not least, compared to GCN[12], NaP has a relatively good performance considering we use significantly fewer labels.

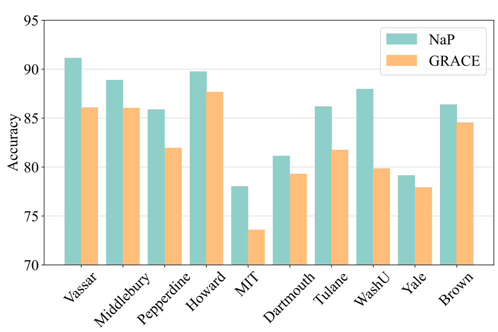

Figure 4: Experiments result of NaP and GRACE on 10 OOD target domains from Facebook100.

(2) NaP's strategy is highly effective. As shown in Fig.4, NaP achieves higher accuracy on 10 additional domains. Since this experiment utilized GRACE as a warm-up stage, NaP's superior OOD generalization ability demonstrates the effectiveness of the proposed strategy in this paper.

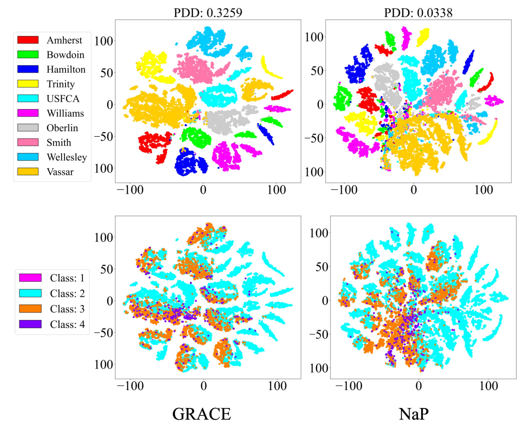

Figure 5: t-SNE visualization and PDD of node embedding.

(3) NaP narrows the distance between domains. As shown in Fig. 5, compared to GRACE, the embedding obtained by NaP exhibits a smaller PDD. More importantly, as the PDD decreases, the node distributions between different domains with the same label become closer.

Table 2: The similarity comparison of different CDPs.

	Input Feature	Embedding
All CDPs	0.0015	0.0199
Transformed CDPs	0.0282	0.8523
Other CDPs	-0.0010	-0.0891

(4) The CDPs transformed by NaP exhibit semantic similarity in the input space. As shown in Table.2 the cosine similarity of all transformed CDPs is significantly higher than that of all CDPs and the remaining CDPs. This demonstrates that NaP indeed transforms the most semantically similar CDPs into positive samples.

5 CONCLUSION

In this work, we investigate the OOD generalization capability of traditional graph contrastive learning methods. We argue that cross-domain pairs (CDPs) make the domains distribution shift larger and hinder the model's OOD generalization capability. Based on this, we propose to transfer the most semantically similar CDPs as positive samples. Comprehensive experiments show that our method NaP significantly benefits the OOD generalization capability of graph contrastive learning methods.

6 ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (Grant No.U21B2046, No.62202448), the Strategic Priority Research Program of the CAS under Grants No. XDB0680302.

REFERENCES

[1] Wendong Bi, Xueqi Cheng, Bingbing Xu, Xiaoqian Sun, Li Xu, and Huawei Shen. 2023. Bridged-gnn: Knowledge bridge learning for effective knowledge transfer. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 99–109.

[2] Wendong Bi, Bingbing Xu, Xiaoqian Sun, Li Xu, Huawei Shen, and Xueqi Cheng. 2023. Predicting the silent majority on graphs: Knowledge transferable graph neural network. In *Proceedings of the ACM Web Conference 2023*. 274–285.

[3] Guoxin Chen, Yongqing Wang, Fangda Guo, Qinglang Guo, Jiangli Shao, Huawei Shen, and Xueqi Cheng. 2023. Causality and independence enhancement for biased node classification. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 203–212.

[4] Mucong Ding, Kezhi Kong, Jiuhai Chen, John Kirchenbauer, Micah Goldblum, David Wipf, Furong Huang, and Tom Goldstein. 2021. A closer look at distribution shifts and out-of-distribution generalization on graphs. (2021).

[5] Wenzheng Feng, Jie Zhang, Yuxiao Dong, Yu Han, Huanbo Luan, Qian Xu, Qiang Yang, Evgeny Kharlamov, and Jie Tang. 2020. Graph random neural networks for semi-supervised learning on graphs. *Advances in neural information processing systems* 33 (2020), 22092–22103.

[6] Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. 2022. Good: A graph out-of-distribution benchmark. *Advances in Neural Information Processing Systems* 35 (2022), 2059–2073.

[7] H Hafidi, M Ghogho, P Ciblat, and A Swami. [n. d.]. Graphcl: Contrastive self-supervised learning of graph representations. arXiv 2020. *arXiv preprint arXiv:2007.08025* ([n. d.]).

[8] Kaveh Hassani and Amir Hosein Khasahmadi. 2020. Contrastive multi-view representation learning on graphs. In *International conference on machine learning*. PMLR, 4116–4126.

[9] Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. 2022. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 594–604.

[10] Shoubo Hu, Kun Zhang, Zhitang Chen, and Laiwan Chan. 2020. Domain generalization via multidomain discriminant analysis. In *Uncertainty in Artificial Intelligence*. PMLR, 292–302.

[11] Ming Jin, Yizhen Zheng, Yuan-Fang Li, Chen Gong, Chuan Zhou, and Shirui Pan. 2021. Multi-scale contrastive siamese networks for self-supervised graph representation learning. *arXiv preprint arXiv:2105.05682* (2021).

[12] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[13] Haoyang Li, Ziwei Zhang, Xin Wang, and Wenwu Zhu. 2022. Learning invariant graph representations for out-of-distribution generalization. *Advances in Neural Information Processing Systems* 35 (2022), 11828–11841.

[14] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. 2018. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[15] Costas Mavromatis and George Karypis. 2021. Graph infoclust: Maximizing coarse-grain mutual information in graphs. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 541–553.

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

[17] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. 2013. Domain generalization via invariant feature representation. In *International conference on machine learning*. PMLR, 10–18.

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

[19] Jiezhong Qiu, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongxia Yang, Ming Ding, Kuansan Wang, and Jie Tang. 2020. Gcc: Graph contrastive coding for graph neural network pre-training. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 1150–1160.

[20] Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Mehdi Azabou, Eva L Dyer, Remi Munos, Petar Veličković, and Michal Valko. 2021. Large-scale representation learning on graphs via bootstrapping. *arXiv preprint arXiv:2102.06514* (2021).

[21] Peifeng Tong, Wu Su, He Li, Jialin Ding, Zhan Haoxiang, and Song Xi Chen. 2023. Distribution free domain generalization. In *International Conference on Machine Learning*. PMLR, 34369–34378.

[22] Amanda L Traud, Peter J Mucha, and Mason A Porter. 2012. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications* 391, 16 (2012), 4165–4180.

[23] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[24] Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. 2018. Deep graph infomax. *arXiv preprint arXiv:1809.10341* (2018).

[25] Vikas Verma, Thang Luong, Kenji Kawaguchi, Hieu Pham, and Quoc Le. 2021. Towards domain-agnostic contrastive learning. In *International Conference on Machine Learning*. PMLR, 10530–10541.

[26] Xiao Wang, Nian Liu, Hui Han, and Chuan Shi. 2021. Self-supervised heterogeneous graph neural network with co-contrastive learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 1726–1736.

[27] Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. 2019. Simplifying graph convolutional networks. In *International conference on machine learning*. PMLR, 6861–6871.

[28] Qitian Wu, Hengrui Zhang, Junchi Yan, and David Wipf. 2022. Handling distribution shifts on graphs: An invariance perspective. *arXiv preprint arXiv:2202.02466* (2022).

[29] Ying-Xin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. 2022. Discovering invariant rationales for graph neural networks. *arXiv preprint arXiv:2201.12872* (2022).

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

[31] Yuning You, Tianlong Chen, Yang Shen, and Zhangyang Wang. 2021. Graph contrastive learning automated. In *International Conference on Machine Learning*. PMLR, 12121–12132.

[32] Jiaqi Zeng and Pengtao Xie. 2021. Contrastive self-supervised learning for graph classification. In *Proceedings of the AAAI conference on Artificial Intelligence*, Vol. 35. 10824–10832.

[33] Shengyu Zhang, Kun Kuang, Jiezhong Qiu, Jin Yu, Zhou Zhao, Hongxia Yang, Zhongfei Zhang, and Fei Wu. 2021. Stable prediction on graphs with agnostic distribution shift. *arXiv preprint arXiv:2110.03865* (2021).

[34] Yifei Zhang, Hao Zhu, Zixing Song, Piotr Koniusz, and Irwin King. 2022. COSTA: covariance-preserving feature augmentation for graph contrastive learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2524–2534.

[35] Zeyang Zhang, Xin Wang, Ziwei Zhang, Haoyang Li, Zhou Qin, and Wenwu Zhu. 2022. Dynamic graph neural networks under spatio-temporal distribution shift. *Advances in Neural Information Processing Systems* 35 (2022), 6074–6089.

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

[37] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. 2021. Graph contrastive learning with adaptive augmentation. In *Proceedings of the Web Conference 2021*. 2069–2080.