Improving Graph Contrastive Learning via Adaptive Positive Sampling

Jiaming Zhuo[1], Feiyang Qin[1], Can Cui[1], Kun Fu[1], Bingxin Niu[1], Mengzhu Wang[1],
Yuanfang Guo[2], Chuan Wang[3*], Zhen Wang[4,5], Xiaochun Cao[6], Liang Yang[1*]
[1]School of Artificial Intelligence, Hebei University of Technology, Tianjin, China
[2]School of Computer Science and Engineering, Beihang University, Beijing, China
[3]Institute of Information Engineering Chinese Academy of Sciences, Beijing, China
[4]School of Cybersecurity, Northwestern Polytechnical University, Xi'an, China
[5]OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an, China
[6]School of Cyber Science and Technology, Sun Yat-sen University, Shenzhen, China
jiaming.zhuo@outlook.com, feiyangqinzzu@163.com, 594021820@qq.com, fukun@hebut.edu.cn,
niubingxin666@163.com, dreamkily@gmail.com, andyguo@buaa.edu.cn, wangchuan@iie.ac.cn,
w-zhen@nwpu.edu.cn, caoxiaochun@mail.sysu.edu.cn, yangliang@vip.qq.com

Abstract

Graph Contrastive Learning (GCL), a Self-Supervised Learning (SSL) architecture tailored for graphs, has shown notable potential for mitigating label scarcity. Its core idea is to amplify feature similarities between the positive sample pairs and reduce them between the negative sample pairs. Unfortunately, most existing GCLs consistently present suboptimal performances on both homophilic and heterophilic graphs. This is primarily attributed to two limitations of positive sampling, that is, incomplete local sampling and blind sampling. To address these limitations, this paper introduces a novel GCL framework with an adaptive positive sampling module, named grapH contrastivE Adaptive posiTive Samples (HEATS). Motivated by the observation that the affinity matrix corresponding to optimal positive sample sets has a block-diagonal structure with equal weights within each block, a self-expressive learning objective incorporating the block and idempotent constraint is presented. This learning objective and the contrastive learning objective are iteratively optimized to improve the adaptability and robustness of HEATS. Extensive experiments on graphs and images validate the effectiveness and generality of HEATS.

1. Introduction

Self-supervised learning (SSL), has been a highly regarded methodology within the unsupervised learning field. It generates supervision signals for model training by extracting latent data patterns from the unannotated data [8]. As a notable branch of SSL, Contrastive Learning (CL) emphasizes producing the supervision signals using the established rule, i.e., bringing similar samples (positive samples) closer and pushing dissimilar samples (negative samples) apart. Based on this rule, Graph Contrastive Learning (GCL) extends the applied range to non-Euclidean graphs by employing techniques like graph augmentations and graph encoders. It has achieved remarkable performance on various downstream tasks, such as node classification [15, 32].

To generate the discriminative node representations from GCLs, the positive sample set of each node should contain all nodes from the same class (i.e., TRUE positive samples), while the negative ones should consist of nodes from different classes (i.e., TRUE negative samples). Subsequently, to extend the TRUE positive sample sets, certain research endeavors in GCLs focus on identifying semantically relevant samples within node neighborhoods [12, 20, 26, 30, 34], leveraging the homophily assumption [1, 16]. Despite their consistent performance improvements on the homophilic graphs, where adjacent nodes typically belong to the same class, the improvements are limited. More critically, their robustness to the heterophilic graphs, where adjacent nodes tend to be from different classes, is severely constrained.

These drawbacks are attributed to two limitations of positive sampling. (1) Incomplete local sampling. Most graphs are sparse, and nodes of the same class are typically situated beyond each other's neighborhoods. Therefore, the positive sample sets for most GCLs tend to be incomplete, which could lead to the absence of shared information between the same class. (2) Blind sampling. In dealing with blindness caused by a lack of supervised guidance, the endeavors of

*Chuan Wang and Liang Yang are corresponding authors.

GCLs to establish the criterion for positive sampling beyond the neighborhoods are insufficient. Concretely, most GCLs always center on exploring semantic relevance in neighborhoods utilizing various homophily measures [6, 12, 34]. In fact, these criteria do not apply beyond neighborhoods because of their local property. Due to the incompleteness and blindness of positive sampling, GCLs can not obtain all TRUE positive sample pairs. This hampers the representation ability of GCLs, resulting in their poor performances on both homophilic and heterophilic graphs.

To address these deficiencies, this paper proposes a novel GCL framework, named grapH contrastivE Adaptive posiTive Sampling (HEATS). The idea is to devise a criterion to guide global positive sampling. Toward this end, the characteristic of an affinity matrix (termed positive sample matrix) associated with the optimal positive sample set is first investigated. The conclusion drawn is that the optimal positive sample matrix obeys the block diagonal property (BDP) and is idempotent, as depicted in Figure 1(a). To be specific, this matrix can be decomposed into a batch of diagonal submatrices (blocks), where each block characterizes semantic relevance among nodes from the same class. Moreover, each block is fully connected and the weights within it are equal. Motivated by this, a novel positive sampling module is presented, focusing on constructing such block-diagonal and idempotent affinity matrices. To capture long-range dependencies, positive sample matrices are generated by optimizing a self-expressive objective incorporating block and idempotent constraints, leveraging the features of all nodes within batches. In light of the intrinsic denoising capability of GCLs [14, 35], an alternating update mechanism of matrix construction and contrastive optimization is introduced to obtain reciprocal benefits, thereby improving the robustness and adaptability of HEATS. In theory, HEATS has a stricter lower bound on the mutual information (MI) between node attributes and node embeddings compared to the baselines that select positive samples from neighborhoods, which guarantees its effectiveness and robustness.

The contributions of this paper are summarized below.

- We investigate the characteristics of an affinity matrix associated with the optimal positive sample set.
- We present a novel graph contrastive learning framework HEATS with an adaptive positive sampling module.
- We theoretically analyze the effectiveness and robustness of the proposed HEATS from a Mutual Information Maximization perspective.
- We conduct extensive evaluations to demonstrate the superior performance and generality of HEATS.

2. Preliminaries

This section starts by explaining the notations used throughout the paper. It then introduces the basic concepts in Graph Contrastive Learning (GCL).

2.1. Notations

Capital italic letters stand for sets (e.g., set $\mathcal{V}$), capital bold letters represent matrices (e.g., matrix $\mathbf{Q}$), bold lowercase letters term vectors (e.g., vector $\mathbf{q}_v$), and lowercase letters denote scalars (e.g., scalar $q_{v,u}$).

For generality, this paper focuses on attribute undirected graphs. It considers a graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathbf{X})$, where the node set $\mathcal{V}$ contains n node instances $\{(\mathbf{x}_v, \mathbf{y}_v)\}_{v\in\mathcal{V}}$. $\mathbf{x}_v \in \mathbb{R}^f$ and $\mathbf{y}_v \in \mathbb{R}^c$ denote the attribute and label vector of the node v, respectively. The node embeddings $\mathbf{H} \in \mathbb{R}^{n\times d}$ are learned on entire graphs in an unsupervised manner, then utilized in downstream tasks, such as node classification, where the labels $\mathbf{Y}$ are used for fine-tuning linear classifiers.

2.2. Graph Contrastive Learning

Being typical of a graph-based self-supervised learning SSL architecture, GCL aims to generate supervised signals based on a predefined rule. It involves bringing the similar samples (i.e., positive samples) closer while pushing the dissimilar samples (i.e., negative samples) apart simultaneously. Take GRACE [32], which is an oft-discussed baseline, as an example, its architecture is described below.

In line with SimCLR [2], a contrastive learning baseline in computer vision (CV), GRACE adopts a two-channel architecture. For graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathbf{X})$, both channels are responsible for generating node embeddings (represented as $\mathbf{H}$ and $\tilde{\mathbf{H}}$) for its two augmented graphs through graph augmentations [32] and encoders [11]. The description of these processes can be found in the appendix.

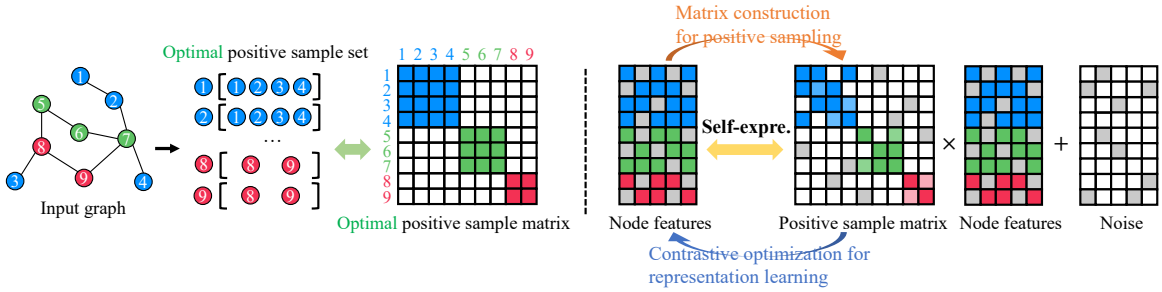
Once the node features are obtained, InfoNCE [22] loss, is calculated as a guide for its update. For node v, the positive sample of the anchor node ($\mathbf{h}_v$) is the same node ($\tilde{\mathbf{h}}_v$) in another view, while the negative samples are other nodes in both views. This contrastive loss (denoted as ℓ_{gc}) can be formulated as

$$\begin{aligned} &\ell_{gc}(\mathbf{h}_v, \tilde{\mathbf{h}}_v) = -log\frac{\mathrm{po}_v}{\mathrm{po}_v + \mathrm{ne}_v}, \\ &\mathrm{po}_v = \mathrm{e}^{\frac{\theta(\mathbf{h}_v, \tilde{\mathbf{h}}_v)}{\tau}},\ \mathrm{ne}_v = \sum_{t\in V\setminus v} \mathrm{e}^{\frac{\theta(\mathbf{h}_v, \mathbf{h}_t)}{\tau}} + \mathrm{e}^{\frac{\theta(\mathbf{h}_v, \tilde{\mathbf{h}}_t)}{\tau}}, \end{aligned} \tag{1}$$

where $\theta(\mathbf{h}_v, \tilde{\mathbf{h}}_v) = s(f(\mathbf{h}_v), f(\tilde{\mathbf{h}}_v))$, $s(,)$ terms the cosine similarity, and $f(\cdot)$ denotes the projection head [2], which is a two-layer multilayer perceptron (MLP). τ represents a scalar, which is positively correlated with the uniformity of the feature distribution. The overall contrastive loss is set as the average over all nodes based on the mutually symmetric form of Equation 1. This can be formulated as

$$\mathcal{L}_{grace} = \frac{1}{2|V|}\sum_{v\in V}\left(\ell_{gc}(\mathbf{h}_v, \tilde{\mathbf{h}}_v) + \ell_{gc}(\tilde{\mathbf{h}}_v, \mathbf{h}_v)\right). \tag{2}$$

In addition to the above pairwise positive sampling, several local positive sampling techniques have been presented



(a) Optimal positive sample set for GCLs and the corresponding matrix. (b) The proposed HEATS Framework.

Figure 1. Overview of the proposed GCL framework, named HEATS, and its design motivation. (a) An intuitive example of the optimal positive sample matrix for GCLs, where the colors of nodes stand for classes. (b) The proposed HEATS framework with adaptive positive sampling. The optimal positive sample set should consist of all nodes from the same class. Accordingly, the optimal positive sample matrix should be block-diagonal, where each block is fully connected and weights within it are equal.

[12, 34]. They propose to add the specified neighbor nodes into the positive sample sets, leveraging the homophily assumption [1, 16].

3. Methodology

In this section, following the introduction of design motivation, a graph contrastive learning (GCL) framework, named grapH contrastivE Adaptive posiTive Sampling (HEATS), is devised. Subsequently, theoretical analysis of its efficacy are presented.

3.1. Motivation

As previously discussed in the introduction, existing positive sampling strategies utilized in GCLs tend to be incomplete and blind, leading to a loss of discriminative power of node representations.

Definition 1. *Block Diagonal Property (BDP) [9]. The given matrix* $\mathbf{Z} \in \mathbb{R}^{n\times n}$ *satisfies block diagonal property if it can be decomposed into submatrices as follows:*

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{Z}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{Z}_k \end{bmatrix},$$

where $\mathbf{Z}_i \in \mathbb{R}^{n_i\times n_i}$ *stands for the* i*-th square submatrix (block), and* $\sum_{i=0}^{k} n_i = n$.

An intuitive and optimal solution is provide target nodes with positive sample sets that consists of all nodes from the same class (i.e., optimal positive sample set), as depicted in Figure 1(a). It is evident that the relationships among these positive samples can be succinctly expressed as a block diagonal matrix, where each block describes the relationships among a class of nodes. Additionally, this matrix should exhibit the following properties: each block is full connected and weights within it are equal. Therefore, the optimal positive sample matrix should be formulated as:

$$\mathbf{Z} = \begin{bmatrix} \frac{1}{n_1}\mathbf{1}_{n_1}\mathbf{1}_{n_1}^\top & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2}\mathbf{1}_{n_2}\mathbf{1}_{n_2}^\top & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{n_k}\mathbf{1}_{n_k}\mathbf{1}_{n_k}^\top \end{bmatrix}, \quad (3)$$

where $\mathbf{1}_{n_i} \in \mathbb{R}^{n_i}$ denotes an all-one column vector.

3.2. HEATS framework

Based on the above analysis, this paper devises a novel GCL framework HEATS, which directs the positive sampling by employing constructed block diagonal affinity matrices. As overviewed in Figure 1(b), HEATS consists of two key components: (1) *Matrix Construction* and (2) *Contrastive Optimization*. These two components engage in an alternating refinement process, gradually converging to the optimum.

3.2.1 Matrix Construction

This component aims to construct affinity matrices, which possess inter-class sparsity and intra-class connectivity (i.e., obey BDP), to guide the selection of positive samples. Particularly, it adopts a self-expressive learning objective, commonly utilized in subspace clustering [13], to acquire such affinity matrices. This objective posits that each sample can be represented by linearly combining all other samples in the same subspace (i.e., block). This objective can be formulated as the following equation:

$$\underset{\mathbf{Z},\mathbf{E}}{\operatorname{argmin}}\, \mathcal{O}(\mathbf{Z}) + \lambda\mathcal{T}(\mathbf{E}), \quad \text{s.t.}\, \mathbf{X} = \mathbf{Z}\mathbf{X} + \mathbf{E}, \quad (4)$$

where $\mathbf{Z}$ denotes the affinity matrix, which depicts the combination coefficients, and $\mathbf{E}$ stands for the noise. Moreover, $\mathcal{O}$ and $\mathcal{T}$ term the constraint on affinity matrices and noise matrices, respectively. λ is a hyperparameter that balances the two terms of the objective function.

To obtain a high-quality affinity matrix, as expressed in Equation 3, several constraints should be incorporated into the learning objective: **idempotent** and k**-block**. Firstly, the affinity matrix $\mathbf{Z}$ in Equation 3 is idempotent [27] because of $\mathbf{Z}_i = \mathbf{Z}_i \times \mathbf{Z}_i = \mathbf{Z}_i^2$. Secondly, the number of blocks in the affinity matrices should be controlled as k. Furthermore, the affinity matrices also should be normalized, symmetric, and non-negative. To achieve these properties and enhance effectiveness, the self-expressive learning objective can be reformulated as:

$$\begin{aligned} &\underset{\mathbf{Z},\mathbf{S}}{\operatorname{argmin}} \|\mathbf{Z}-\mathbf{S}\|_F^2 + \gamma\|\mathbf{S}\|_{id} + \lambda\|\mathbf{E}\|_{2,1}, \\ &\text{s.t. } \mathbf{H} = \mathbf{Z}\mathbf{H} + \mathbf{E}, \\ &\mathbf{S}\mathbf{1}_n = \mathbf{1}_n, \mathbf{S} = \mathbf{S}^\top, \mathbf{S} \geq \mathbf{0}, \operatorname{Tr}(\mathbf{S}) = k, \end{aligned} \tag{5}$$

where $\mathbf{S}$ denotes an intermediate-term to mitigate the loss of representation capability. And, $\|\mathbf{S}\|_{id} = \|\mathbf{S} - \mathbf{S}^2\|_F^2$ terms the idempotent constraint and γ is a scalar that controls the impacts of this term. The terms $\mathbf{S}\mathbf{1}_n = \mathbf{1}_n$, $\mathbf{S} = \mathbf{S}^\top$, and $\mathbf{S} \geq \mathbf{0}$ stand for the column-normalized, symmetric, and nonnegative constraints, respectively. Based on the above constraints, the term $\operatorname{Tr}(\mathbf{S}) = k$ is introduced to make the matrix have k blocks. $\|\mathbf{E}\|_{2,1} = \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{n}([\mathbf{E}]_{ij})^2}$ is the $\ell_{2,1}$-norm of noise, which characterizes sample-specific outliers. Finally, the affinity matrix, a guide of positive sampling, can be obtained by optimizing Equation 5. Please refer to the appendix for optimization details. Once the affinity matrix $\mathbf{M}$ (i.e., $\mathbf{Z}$ or $\mathbf{S}$) is obtained, it is necessary to sparsify it to enhance training efficiency. The sparsification process can be formulated as:

$$m_{v,u} = \begin{cases} 0, & m_{v,u} < \beta, \\ m_{v,u}, & \text{otherwise.} \end{cases}$$

where β denotes the threshold parameter, which is utilized to indicate filter out the values below it.

Theorem 1. *The affinity matrix* $\mathbf{M}$*, which is generated by optimizing Equation (5), satisfies the BDP in Definition 1.*

Theorem 1 theoretically demonstrates that the proposed matrix construction strategy aligns with the motivation. For the proof, please refer to the appendix.

3.2.2 Contrastive Optimization

This subsection delineates the methodology for formulating the contrastive loss leveraging the constructed affinity matrix $\mathbf{M}$. Firstly, it is utilized to direct the sampling process. By identifying non-zero values in the vector $\mathbf{m}_v$, nodes that are semantically associated with node v can be marked out. These nodes collectively constitute the positive sample set of node v, denoted as $\mathcal{P}_v^M = \{u \,|\, m_{v,u}{>}0\}$. Following the baselines [2, 32], the negative sample set is composed of the remaining nodes, that is $\mathcal{N}_v^M = V \setminus \{\mathcal{P}_v^M \cup v\}$. Secondly, the matrix $\mathbf{M}$ is also interpreted as a measure of the relationships, and the values within it are treated as weights.

In implementation, HEATS follows the baseline GRACE [32], which is mentioned in Section 2.2. Therefore, the contrastive loss of HEATS with respect to anchor node $\mathbf{h}_v$ can be reformulated from Equation 1 as

$$\begin{aligned} \ell_{ht}(\mathbf{h}_v, \tilde{\mathbf{h}}_v) &= -log\frac{\mathrm{po}_v}{\mathrm{po}_v + \mathrm{ne}_v}, \\ \mathrm{po}_v &= \mathrm{e}^{\frac{\theta(\mathbf{h}_v, \tilde{\mathbf{h}}_v)}{\tau}} + \sum_{u \in \mathcal{P}_v^M} m_{v,u} \cdot \mathrm{e}^{\frac{\theta(\mathbf{h}_v, \mathbf{h}_u)}{\tau}}, \\ \mathrm{ne}_v &= \sum_{t \in \mathcal{N}_v^M} \mathrm{e}^{\frac{\theta(\mathbf{h}_v, \tilde{\mathbf{h}}_t)}{\tau}} + \sum_{t \in \mathcal{N}_v^M} \mathrm{e}^{\frac{\theta(\mathbf{h}_v, \mathbf{h}_t)}{\tau}}, \end{aligned} \tag{6}$$

The overall contrastive loss of HEATS, denoted as $\mathcal{L}_{heats}$,

$$\mathcal{L}_{heats} = \frac{1}{2|V|} \sum_{v \in V} \left(\ell_{ht}(\mathbf{h}_v, \tilde{\mathbf{h}}_v) + \ell_{ht}(\tilde{\mathbf{h}}_v, \mathbf{h}_v) \right). \tag{7}$$

In comparison to the existing GCLs [12, 20, 34], the proposed HEATS exhibits several characteristics. (1) Generality. Since the positive sample matrices are constructed globally, that is, independent of the underlying graph topology, HEATS can be reasonably extended to non-graph domains such as computer vision (CV). (2) Robustness. Due to the fact that the positive sample matrix is obtained by optimizing a robust self-expressive learning objective, as formulated in Equation (5), the robustness is strengthened through positive sampling on these matrices.

Theorem 2. *The contrastive loss of HEATS* ($\mathcal{L}_{heats}$) *is a more stringent estimate of mutual information (MI) between node attributes and embeddings than that of the local baseline HomoGCL [12], that is,*

$$\mathcal{L}_{homogcl} \leq \mathcal{L}_{heats} \leq I(\mathbf{X}; \mathbf{H}, \tilde{\mathbf{H}}), \tag{8}$$

where $\mathbf{X}$ *denotes the node attributes, and* $\mathbf{H}$ *and* $\tilde{\mathbf{H}}$ *represent the node embeddings in two augmented views.*

Theorem 2 demonstrates that the proposed HEATS provides a refined approximation to the TRUE mutual information, enabling it to promote model robustness and stability.

3.2.3 Alternating Update

To enhance the adaptivity of HEATS, the two fundamental components, *matrix construction* and *contrastive optimization*, engage in iterative updates in an alternating fashion, as visually depicted in Figure 1(b). The former is tasked with solving Equation 5 to construct positive sample matrices in the feature space. The latter focuses on representation (feature) learning by optimizing the contrastive loss using these

Table 1. Statistics of twelve graph datasets. Homophily is the edge homophily in [17].

Dataset	Nodes	Edges	Features	Classes	Homophily
Cora	2,708	5,278	1,433	7	0.81
CiteSeer	3,327	4,552	3,703	6	0.74
PubMed	19,717	44,324	500	3	0.80
Wiki-CS	11,701	216,123	300	10	0.65
Computers	13,752	245,861	767	10	0.78
Photo	7,650	238,163	745	8	0.83
Cornell	183	295	1,703	5	0.13
Texas	183	309	1,703	5	0.11
Wisconsin	251	499	1,703	5	0.20
Chameleon	2,277	36,101	2,325	5	0.23
Squirrel	5,201	217,073	2,089	5	0.22
Actor	7,600	33,544	931	5	0.22

positive sample matrices, which is formulated as Equation 6. These two components are updated alternately.

The alternating update procedure offers several benefits to GCLs. (1) Adaptive positive samples. Note that the positive sample matrices are constructed in the feature space. Hence, this procedure can continually provide adaptive positive samples for the subsequent contrastive optimization, adapting to feature changes. (2) Discriminative node representations. GCLs have the denoising ability, to some extent, which is attributed to the fact that it obtains the invariant information [28]. As a result, the discriminative feature space can be provided for the construction of positive sample matrices, eventually resulting in more discriminative features through contrastive learning. Furthermore, benefits also involve the risk reduction of each component falling into the local optimum and accelerating the convergence speed.

4. Experiments

In this section, the proposed framework HEATS is validated by empirically evaluating its performances on node and image classification tasks in the unsupervised setting. Next, an in-depth understanding of the efficacy of this framework is provided through several experiment analyses.

4.1. Experiment Setup

Datasets. To illustrate the effectiveness and generality of the proposed HEATS, twelve non-Euclidean graph datasets, and three 2-D spatial image datasets are adopted in the experiments. According to graph homophily [17], these graph datasets can be divided into two categories: six **homophilic graphs**, including Cora, CiteSeer, PubMed, Wiki-CS, Computers, and Photo, and six **heterophilic graphs**, containing Cornell, Texas, Wisconsin, Chameleon, Squirrel, and Actor. Moreover, the image datasets are CIFAR-10, STL-10, and CIFAR-100. The statistics of the graph datasets are presented in Table 1. Due to space limitations, the introduction of all datasets is presented in the appendix.

Splitting. To ensure experimental fairness, the dataset splitting follows commonly employed schemes. Specifically, for three homophilic graphs (Cora, CiteSeer, and PubMed) and all six heterophilic graphs, the training, validation, and testing sets constitute 48%, 32%, and 20% of the data, respectively. For the remaining homophilic graphs (Wiki-CS, Computers, and Photo), the proportions amount to 10%, 10% and 80%. Besides, for the CIFAR-10 dataset, 5000 and 1000 images per class are selected for training and testing, respectively. For the STL-10 dataset, 10500 and 800 images per class are assigned to the training and testing sets, respectively. For the CIFAR-100 dataset, 500 and 100 images per class are served in training and testing, respectively.

Baselines. These experiments involve two types of downstream tasks: node classification for graphs and image classification. For the node classification, the baselines include two types: **semi-supervised** graph neural networks (GCN [11], GAT [24] and JKNet [29]), and **unsupervised** graph learning models (Deepwalk [18], Node2vec [4], GAE [10], VGAE [10], DGI [25], MVGRL [5], GRACE [32], GCA [33], BGRL [21], SELENE [31], and HomoGCL [12]). For the image classification, the baseline is CL model SimCLR [2], which is implemented with two backbones (ResNet-18 and ResNet-50) [7]. Please refer to the appendix for an introduction to related works.

Configurations. In the node classification, the proposed HEATS framework is implemented with the configurations following GRACE [32]. To be specific, each branch corresponds to an augmented graph, which is initially obtained through the edge deletion and attribute masking [3], with the random ratio among $\{0.2, 0.4, 0.6, 0.8\}$. Furthermore, the node features are obtained through a two-layer GCN [11] encoding and a two-layer MLP [19] projecting, where both dimensions are 64. The node features are trained by minimizing the contrastive loss, where the temperature coefficient is taken from $\{0.2, 0.4, 0.6, 0.8, 1\}$. In this process, an Adam optimizer with the learning rate of 0.01 and the weight decay rate among $\{0, 5 \times 10^{-5}\}$ is to be employed for optimization and regularization, respectively. Minibatch training is adopted for larger node numbers, such as PubMed. For the unique parameters of HEATS, namely γ and λ, both are selected from $\{1 \times 10^{-3}, 1 \times 10^{-2}, 1 \times 10^{-1}, 1, 10\}$. Furthermore, the number of blocks k is selected from a set with no more than the number of classes, and its impact is analyzed in Section 4.3.2. All experiments are conducted in PyTorch on a single RTX4090 24GB GPU.

Evaluation Protocol. For both node and image classification tasks, the experimental evaluations adopt the standard linear evaluation protocol [25]. Concretely, all models undergo initial unsupervised pre-training, followed by utilizating the obtained embeddings to train linear classifiers and subsequently presenting the test accuracy results. Throughout the evaluation phases, the Adam optimizer with a learning rate of 0.01 is utilized. The experiment results are re-

Table 2. Node classification accuracy (mean ± std) is reported for six homophilic datasets. The top-performing unsupervised model is denoted in **bold**, and the second-best unsupervised model in underline. The second column specifies the training information.

Model	Training Data	Cora	CiteSeer	PubMed	Wiki-CS	Computers	Photo
GCN	A, X, Y	85.77 ± 0.25	73.68 ± 0.31	88.13 ± 0.28	76.89 ± 0.37	86.34 ± 0.48	92.35 ± 0.25
GAT	A, X, Y	86.37 ± 0.30	74.32 ± 0.27	87.62 ± 0.26	77.42 ± 0.19	87.06 ± 0.35	92.64 ± 0.42
JKNet	A, X, Y	85.93 ± 1.35	74.37 ± 1.53	87.68 ± 0.30	79.52 ± 0.21	85.28 ± 0.72	92.68 ± 0.13
DeepWalk	A	73.96 ± 0.12	61.91 ± 0.42	74.79 ± 0.98	74.35 ± 0.06	85.68 ± 0.06	89.44 ± 0.11
Node2Vec	A	75.87 ± 0.22	62.54 ± 0.13	76.49 ± 0.32	71.79 ± 0.05	84.39 ± 0.08	89.67 ± 0.12
GAE	A, X	76.83 ± 1.22	65.43 ± 1.13	76.52 ± 0.33	70.15 ± 0.01	85.27 ± 0.19	91.62 ± 0.13
VGAE	A, X	79.36 ± 0.83	69.18 ± 0.27	79.17 ± 0.44	76.63 ± 0.19	86.37 ± 0.21	92.20 ± 0.11
DGI	A, X	85.90 ± 0.57	72.57 ± 0.23	83.52 ± 1.24	75.73 ± 0.13	84.09 ± 0.39	91.49 ± 0.25
MVGRL	A, X	86.77 ± 0.33	73.71 ± 0.48	84.63 ± 0.73	77.97 ± 0.18	87.09 ± 0.27	92.01 ± 0.13
GRACE	A, X	84.79 ± 0.64	72.94 ± 0.72	84.51 ± 0.68	79.16 ± 0.36	87.21 ± 0.44	92.65 ± 0.32
GCA	A, X	85.16 ± 0.51	72.73 ± 0.45	85.22 ± 0.73	79.35 ± 0.12	87.84 ± 0.27	92.78 ± 0.17
BGRL	A, X	85.37 ± 0.74	73.45 ± 0.83	84.61 ± 0.32	78.74 ± 0.22	88.92 ± 0.33	93.24 ± 0.29
SELENE	A, X	85.28 ± 0.83	73.48 ± 0.65	84.70 ± 0.52	78.31 ± 0.63	88.13 ± 0.51	92.93 ± 0.34
HomoGCL	A, X	85.02 ± 0.68	73.67 ± 0.78	82.33 ± 0.49	77.47 ± 0.45	87.84 ± 0.28	93.59 ± 0.27
HEATS	A, X	$\mathbf{87.10 \pm 1.40}$	$\mathbf{75.26 \pm 1.25}$	$\mathbf{85.41 \pm 0.82}$	$\mathbf{79.99 \pm 1.59}$	$\mathbf{89.29 \pm 1.12}$	$\mathbf{94.65 \pm 1.71}$

ported as an average of over ten random runs.

4.2. Results on benchmark datasets

Results on Homophilic Graphs. Table 2 presents the experimental results of node classification on six homophilic graphs. It is evident that compared to all unsupervised baseline models, the proposed HEATS achieves the best performance across all datasets. It even surpasses all supervised comparison models on datasets other than PubMed. To be specific, on the CiteSeer dataset, HEATS outperforms the second-best unsupervised model (i.e., MVGRL) by 1.55% and the baseline model (i.e., GRACE) by 2.32% in classification accuracy. This highlights the effectiveness of HEATS in exploring positive samples by modeling high-order relationships.

Results on Homophilic Graphs. A similar phenomenon can be observed from the experiment results on heterophilic graphs, as illustrated in Table 3. To be specific, compared to the baseline model (i.e., GRACE), HEATS achieves performance improvements across all datasets. Particularly noteworthy are improvements in accuracy by 12.11%, 7.29%, and 15.88% on the Cornell, Texas, and Wisconsin datasets, respectively. This emphasizes the contribution of the proposed positive sampling based on matrix construction in enriching the universality of the baselines. Moreover, HEATS outperforms other unsupervised baselines on three datasets, except the Chameleon, Squirrel, and Actor datasets, showcasing its superior performance.

Visualization. This experiment aims to intuitively demonstrate the representation ability of the proposed HEATS. For this purpose, the t-SNE [23] method is exploited to perform feature reduction and visualization of the trained representations. Figure 2 exhibits the experiment results (scatter plots) on three benchmark datasets (i.e., Cora, CiteSeer, and Photo), where colors stand for the classes of nodes. It can be observed that compared to all the comparison models (i.e., GRACE and BGRL), the proposed HEATS framework is enabled to produce the most informative representations. As exemplified by the Cora and CiteSeer dataset, compared to the baseline model (i.e., GRACE), HEATS achieves more compact clusters in node embeddings. Specifically, the embeddings of the same class are closer, while the embeddings of different classes exhibit more significant differences. The experiment results emphasize the effectiveness of HEATS in enhancing representation ability.

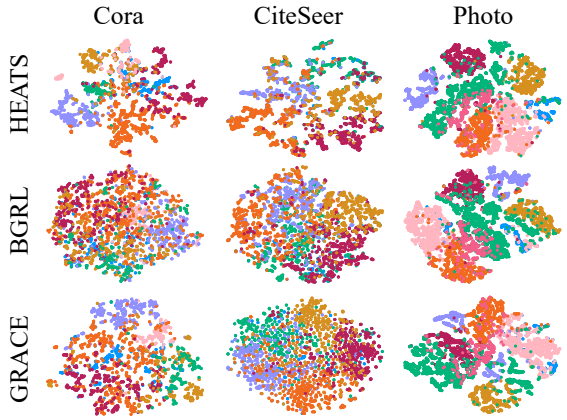


Figure 2. t-SNE visualization of node embeddings from HEATS, BGRL, and GRACE on Cora, CiteSeer, and Photo datasets.

Results on Image Datasets. As previously mentioned, the proposed HEATS designs a positive sampling strategy from a global perspective, which effectively mitigates graph structural constraints. In essence, this improves its generality to various domains. To substantiate this claim, the experiment compares a contrastive learning (CL) model tailored for images (i.e., SimCLR) with the proposed variants.

Table 3. Node classification accuracy (mean ± std) is reported for six heterophilic datasets. The top-performing unsupervised model is denoted in **bold**, and the second-best unsupervised model in underline. The second column specifies the training information.

Model	Training Data	Cornell	Texas	Wisconsin	Chameleon	Squirrel	Actor
GCN	A, X, Y	55.14 ± 7.57	55.68 ± 9.61	58.42 ± 5.10	59.82 ± 2.58	36.89 ± 1.34	30.64 ± 1.49
GAT	A, X, Y	58.92 ± 3.32	58.38 ± 4.45	55.29 ± 8.71	60.26 ± 2.50	40.72 ± 1.55	27.44 ± 0.89
JKNet	A, X, Y	56.49 ± 3.22	65.35 ± 4.86	51.37 ± 3.21	60.31 ± 2.76	44.24 ± 2.11	36.47 ± 0.51
DeepWalk	A	39.18 ± 5.57	46.49 ± 6.49	33.53 ± 4.92	47.74 ± 2.05	32.93 ± 1.58	22.78 ± 0.64
Node2Vec	A	42.94 ± 7.46	41.92 ± 7.76	37.45 ± 7.09	41.93 ± 3.29	22.84 ± 0.72	28.28 ± 1.27
GAE	A, X	58.85 ± 3.21	58.64 ± 4.53	52.55 ± 3.80	33.84 ± 2.77	28.03 ± 1.61	28.03 ± 1.18
VGAE	A, X	59.19 ± 4.09	59.20 ± 4.26	56.67 ± 5.51	35.22 ± 2.71	29.48 ± 1.48	26.99 ± 1.56
DGI	A, X	63.35 ± 4.61	60.59 ± 7.56	55.41 ± 5.96	39.95 ± 1.75	31.80 ± 0.77	29.82 ± 0.69
MVGRL	A, X	64.30 ± 5.43	62.38 ± 5.61	62.37 ± 4.32	$\mathbf{51.07 \pm 2.68}$	35.47 ± 1.29	$\mathbf{30.02 \pm 0.70}$
GRACE	A, X	54.86 ± 6.95	57.57 ± 5.68	50.00 ± 5.83	48.05 ± 1.81	31.33 ± 1.22	29.01 ± 0.78
GCA	A, X	55.41 ± 4.56	59.46 ± 6.16	50.78 ± 4.06	49.80 ± 1.81	35.50 ± 0.91	29.65 ± 1.47
BGRL	A, X	57.30 ± 5.51	59.19 ± 5.85	52.35 ± 4.12	47.46 ± 2.74	32.64 ± 0.78	29.86 ± 0.75
SELENE	A, X	59.94 ± 5.12	61.87 ± 4.25	61.87 ± 4.79	42.13 ± 2.15	33.28 ± 0.82	30.12 ± 0.76
HomoGCL	A, X	48.64 ± 2.59	54.05 ± 2.32	39.21 ± 5.75	48.68 ± 1.16	$\mathbf{38.71 \pm 0.85}$	28.81 ± 0.78
HEATS	A, X	$\mathbf{66.97 \pm 6.74}$	$\mathbf{64.86 \pm 4.68}$	$\mathbf{65.88 \pm 5.56}$	49.96 ± 1.86	36.24 ± 1.11	29.91 ± 1.16

Table 4. Image classification accuracy (mean) of SimCLR and its HEATS-based variants. ↑ dentoes the performance improvement.

Backbone	**Method**	CIFAR-10	STL-10	CIFAR-100
ResNet-18	SimCLR	78.59	78.44	49.60
	+HEATS	79.19	78.83	52.62
	↑	**+0.60**	**+0.39**	**+3.02**
ResNet-50	SimCLR	75.03	76.73	49.31
	+HEATS	77.58	77.80	53.81
	↑	**+2.55**	**+1.07**	**+4.50**

The configurations of these variants align with that of the baselines, including standard parameters such as learning rate and training epochs. Please refer to the appendix for experiment setups. Table 4 exhibits the image classification results after 30 epochs using ResNet-18 and ResNet-50 backbones. It can be seen that compared to the baseline (i.e., SimCLR), the proposed HEATS framework consistently improves performance across all three datasets. To be specific, HEATS showcases superior performance compared to the two baselines utilizing ResNet-18 and ResNet-50 backbones on the CIFAR-100 dataset by 3.02% and 4.5%, respectively, accentuating the generality to image-centric tasks. More importantly, it provides valuable insights for applying HEATS in a wide range of domains.

4.3. Further Analysis

4.3.1 Effectiveness Justify

The experiment intends to justify the effectiveness of the two proposed approaches to enhance positive sampling, i.e., matrix construction and alternating updates, to pursue the block-diagonal matrices. Therefore, three matrices, which include the adjacent matrix, the matrix computed using the initial features, and the matrix computed using the features after 200 iterations, are visualized in Figure 3 to provide an intuitive understanding.

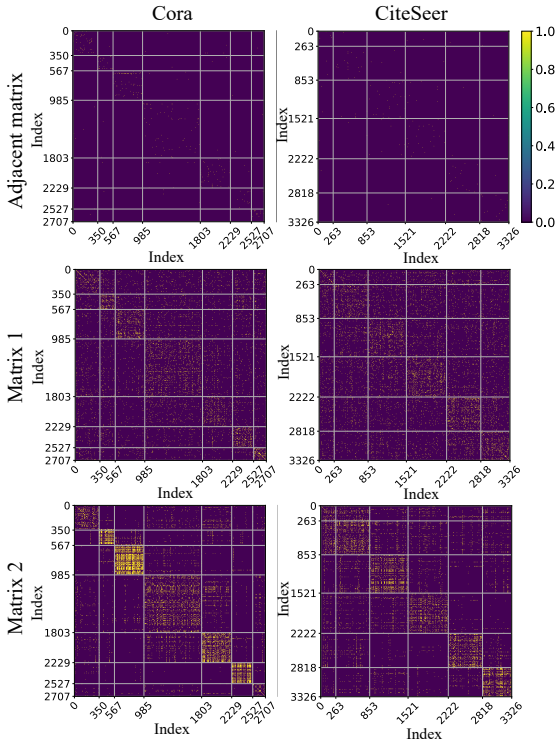
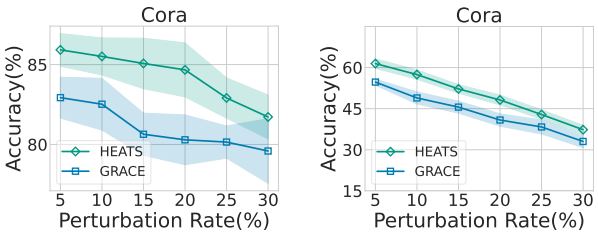


Figure 3. Visualization of the adjacent matrices, the matrices computed using the initial features (Matrix 1), and the matrices computed using the learned features after 200 iterations (Matrix 2) on the Cora and CiteSeer datasets. The nodes are sorted by class. The weights between nodes are not considered here.

Based on the observations in Figure 3, three significant conclusions can be drawn. (1) The constructed matrices are approximately block diagonal. In particular, on all datasets

Figure 4. Impact of the number of blocks k.



(a) Topology attacks (b) Attribute attacks

Figure 5. Results of GCLs under graph attacks

(i.e., Cora and CiteSeer), the edges between nodes from the same class outnumber those between nodes from different classes. This indicates that the proposed HEATS framework can serve as an effective strategy to obtain more TRUE positive samples. (2) Compared to the adjacent matrices, the constructed matrices exhibit denser diagonal blocks. This implies that the proposed matrix construction approach is feasible in contrastive learning by offering more TRUE positive samples beyond graph topology. (3) Compared to the matrices constructed in the initial feature space, those constructed in the updated feature space exhibit a more distinct block-diagonal structure, e.g., results of the Cora dataset. This emphasizes the effectiveness of the proposed alternating update approach. In summary, the experiment confirms the efficacy of the two proposed approaches in constructing block-diagonal matrices for positive sample collection.

4.3.2 Hyperparameter Study

The experiment aims to provide a guide on choosing the parameter k, which represents the number of blocks, by investigating its impact on the performance. This parameter is selected from 2 to the number of classes, as setting it to 1 seems ineffective. Figure 4 exhibits the experiment results on three datasets (i.e., Cora, CiteSeer, and Cornell). Incorporating the results in Table 2, it can be observed that HEATS consistently performs well within a specific parameter range. For example, on the Cora dataset, when the parameter selects from the set $\{4, 5, 6\}$, HEATS consistently outperforms the baseline model (i.e., GRACE) by 2.21%, 2.31%, and 2.25%, respectively. This indicates HEATS insensitivity to changes in this parameter. Additionally, with a small value (e.g., 2), HEATS performs better than the baseline GRACE, which offers guidance to select the parameter.

4.3.3 Robustness Analysis

This experiment is intended to evaluate the robustness of the proposed HEATS framework to noise. The topology attack (adding edges) and the attribute attack (flipping attributes) are applied to create noisy data on the Cora dataset. Figure 5 shows the performance variation of the GCL, HEATS, and the baseline GRACE under topology and attribute attacks.

It can be seen from Figure 5a that GRACE is somewhat resistant to robust against topology attacks, but the performance degrades significantly with higher attack levels. For example, the accuracy decreases by about 2% compared to the attack-free accuracy at an edge deletion rate of 0.1. This is attributed to their centers on capturing the invariant information from multiple augmented graphs, to some extent, resulting in them being insensitive to topology changes. However, the accuracy decreases 5% when the edge deletion rate increases to 0.3. In contrast, HEAT exhibits greater robustness to this attack. Specifically, it achieves about 4.5% increase in accuracy compared to the baseline GRACE when the edges deletion reaches an extreme 0.2%. This is mainly due to HEATS incorporating noise simulation and alternating update approach to extract high-order positive samples, thereby enhancing the robustness. Furthermore, as shown in Figure 5b, HEATS consistently outperforms the baseline GRACE, showcasing its denoising ability.

5. Conclusions

In this paper, a novel graph contrastive learning framework, named HEATS, is proposed to address two key limitations of the existing positive sampling techniques: incomplete local sampling and blind sampling. Drawing inspiration from the observation that the affinity matrix associated with the optimal positive sample set is block-diagonal and idempotent, the idea of constructing such matrices to guide positive sampling is proposed. To enhance the adaptability and robustness of HEATS, an approach that alternately updates the positive sampling and contrastive optimization is proposed. Extensive experiments reveal the efficacy and robustness of HEATS. Future research directions include improving the scalability of the framework to accommodate large datasets and improving its generality to multimodal data.

6. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62376088, 61972442, 62272020, 62102413, 62132006, U2001202, and U22B2036), in part by the National Science Fund for Distinguished Young Scholars (No. 62025602), in part by the National Social Science Fund of China under Grant 22VMG037, in part by the Natural Science Foundation of Hebei Province of China under Grant F2020202040, and in part by the Tencent Foundation and XPLORER PRIZE.

References

[1] Kristen M Altenburger and Johan Ugander. Monophily in social networks introduces similarity among friends-of-friends. *Nature human behaviour*, 2(4):284–290, 2018. 1, 3

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020. 2, 4, 5

[3] Kaize Ding, Zhe Xu, Hanghang Tong, and Huan Liu. Data augmentation for deep graph learning: A survey. *SIGKDD Explor.*, 24(2):61–77, 2022. 5

[4] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *SIGKDD*, pages 855–864, 2016. 5

[5] Kaveh Hassani and Amir Hosein Khas Ahmadi. Contrastive multi-view representation learning on graphs. In *ICML*, pages 4116–4126, 2020. 5

[6] Dongxiao He, Jitao Zhao, Rui Guo, Zhiyong Feng, Di Jin, Yuxiao Huang, Zhen Wang, and Weixiong Zhang. Contrastive learning meets homophily: Two birds with one stone. In *ICML*, pages 12775–12789, 2023. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[8] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In *NeurIPS*, pages 15637–15648, 2019. 1

[9] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012. 3

[10] Thomas N. Kipf and Max Welling. Variational graph auto-encoders. *CoRR*, abs/1611.07308, 2016. 5

[11] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 2, 5

[12] Wen-Zhi Li, Chang-Dong Wang, Hui Xiong, and Jian-Huang Lai. Homogcl: Rethinking homophily in graph contrastive learning. In *SIGKDD*, pages 1341–1352, 2023. 1, 2, 3, 4, 5

[13] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):171–184, 2013. 3

[14] Yao Ma, Xiaorui Liu, Tong Zhao, Yozen Liu, Jiliang Tang, and Neil Shah. A unified view on graph neural networks as graph signal denoising, 2020. 2

[15] Yixuan Ma, Xiaolin Zhang, Peng Zhang, and Kun Zhan. Entropy neural estimation for graph contrastive learning. In *MM*, pages 435–443, 2023. 1

[16] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001. 1, 3

[17] Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-gcn: Geometric graph convolutional networks. In *ICLR*, 2020. 5

[18] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: online learning of social representations. In *SIGKDD*, pages 701–710, 2014. 5

[19] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986. 5

[20] Xiao Shen, Dewang Sun, Shirui Pan, Xi Zhou, and Laurence T. Yang. Neighbor contrastive learning on learnable graph augmentation. In *AAAI*, pages 9782–9791, 2023. 1, 4

[21] Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Rémi Munos, Petar Veličković, and Michal Valko. Bootstrapped representation learning on graphs. 2021. 5

[22] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. 2

[23] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 6

[24] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018. 5

[25] Petar Velickovic, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. Deep graph infomax. In *ICLR*, 2019. 5

[26] Haonan Wang, Jieyu Zhang, Qi Zhu, and Wei Huang. Augmentation-free graph contrastive learning. *CoRR*, abs/2204.04874, 2022. 1

[27] Lai Wei, Shiteng Liu, Rigui Zhou, and Changming Zhu. Learning idempotent representation for subspace clustering. *CoRR*, abs/2207.14431, 2022. 4

[28] Tete Xiao, Xiaolong Wang, Alexei A. Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. In *ICLR*, 2021. 5

[29] Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. In *ICML*, pages 5449–5458, 2018. 5

[30] Hengrui Zhang, Qitian Wu, Yu Wang, Shaofeng Zhang, Junchi Yan, and Philip S. Yu. Localized contrastive learning on graphs. *CoRR*, abs/2212.04604, 2022. 1

[31] Zhiqiang Zhong, Guadalupe Gonzalez, Daniele Grattarola, and Jun Pang. Unsupervised network embedding beyond homophily. *arXiv preprint arXiv:2203.10866*, 2022. 5

[32] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep graph contrastive representation learning. *CoRR*, abs/2006.04131, 2020. 1, 2, 4, 5

[33] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Graph contrastive learning with adaptive augmentation. In *WWW*, 2021. 5

[34] Jiaming Zhuo, Can Cui, Kun Fu, Bingxin Niu, Dongxiao He, Chuan Wang, Yuanfang Guo, Zhen Wang, Xiaochun Cao, and Liang Yang. Graph contrastive learning reimagined: Exploring universality. In *WWW*, 2024. 1, 2, 3, 4

[35] Yunhao Zou and Ying Fu. Estimating fine-grained noise model via contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12682–12691, 2022. 2