Understanding Graph Contrastive Learning From A Statistical Perspective

Anonymous authors
Paper under double-blind review

Abstract

Although recent advances have prompted the prosperity in graph contrastive learning, the researches on universal principles for model design and desirable properties of latent representations are still inadequate. From a statistical perspective, this paper proposes two principles for guidance and constructs a general graph self-supervised framework. Reformulating data augmentation as a mixture process, the first one, termed consistency principle, lays stress on exploring and mapping cross-view common information to consistent and essence-revealing representations. For the purpose of instantiation, four statistical indicators are employed to estimate and maximize the correlation between representations from various views, whose accordant variation trend during training implies the extraction of common content. With awareness of the insufficiency of a solo consistency principle, suffering from degenerated and coupled solutions, a decorrelation principle is put forward to encourage diverse and informative representations. Accordingly, two specific strategies, performing in representation space and eigen spectral space, respectively, are propounded to decouple various representation channels. Under two principles, various combinations of concrete implementations derive a family of methods. Provably, after decomposition and analysis for the commonly used *InfoNCE* loss, we clarify that the approaches based on mutual information maximization implicitly fulfill the two principles and are covered within our framework. The comparison experiments with current state-of-the-arts demonstrate the effectiveness and sufficiency of two principles for high-quality graph representations. Furthermore, visual studies reveal how certain principles affect learned representations.

1 Introduction

Independent of costly manual annotations, Self-Supervised Learning (SSL) extracts meaningful information from unlabeled data sources and learns useful representations using a specific proxy task. Based on various proxy objectives, a series of SSL paradigms have been proposed. As a distinguished member of the SSL family, contrastive learning has exhibited great prospects and yielded brilliant results in various fields spanning computer vision (He et al., 2020; Chen et al., 2020; Chen & He, 2021; Grill et al., 2020) and graphs (You et al., 2020; Zhu et al., 2021; Hassani & Khasahmadi, 2020; Xu et al., 2021). On the basis of the reliance on negative samples, the general contrastive learning methods can be divided into two families: *negative-rely* approaches (Chen et al., 2020; He et al., 2020; You et al., 2020; Hassani & Khasahmadi, 2020) and *negative-free* approaches (Grill et al., 2020; Chen & He, 2021; Bielak et al., 2021). The former learns informative representations by pushing negative pairs away while the latter adopts special strategies (*e.g.*, asymmetric architectures (Grill et al., 2020)) to prevent collapsed solutions.

As a distinctive characteristic, contrastive learning takes the form of multi-view learning, where multiple views can be naturally acquired (*e.g.*, image and text (Radford et al., 2021)) or artificially generated (*e.g.*, distorting the raw data via augmentation (Chen et al., 2020; Zhu et al., 2021)). The multiple views of a specific instance can be regarded as a data pair sampled from the joint distribution of various data sources, describing the instance object from different aspects. Under contrastive learning scheme, the neural models are trained to pull together representations from various views of the same instance (*i.e.*, positive pairs) while pushing apart those from different instances. This paradigm is usually explained as estimating the mutual information between two views from an *information theoretic* perspective (Xu et al., 2021; Zhu et al., 2021). Some research efforts based on information theory have been put into understanding and guiding contrastive learning, such as view selection (Tian et al., 2020) and contrastive objective design (Tsai et al., 2021).

For multi-view data, the common information across multiple sources is usually essence-revealing and facilitates various downstream tasks (Tian et al., 2020; Lyu et al., 2022). One of the research topics in this paper is to train a neural encoder to mine and map the view-invariant common information in graph data to representation space. Intuitively, if an encoder excels at capturing common information in various sources, the representations from different views should exhibit high consistency and correlation. Inspired by this, from a statistical dependence perspective, we propose one principle of *cross-view maximum consistency* for contrastive graph representation learning. Under the consistency principle, neural encoders are encouraged to map common information in multiple sources to *consistent* representations, which from various views are highly correlated. This principle naturally motivates us to search for appropriate statistical indicators to measure cross-view latent correlation in representation space. Four classical metrics, including Distance Correlation (Székely et al., 2007), RV-coefficient (Robert & Escoufier, 1976), simple Matrix Correlation (Smilde et al., 2008), and Hilbert-Schmidt Independence Criterion (Gretton et al., 2005), are thoroughly investigated and employed to realize the consistency principle. Compared with the notoriously hard estimation of mutual information, explicitly dependent on the data probability distribution, these indicators can be readily obtained from the empirical data without the participation of parameterized estimators.

The consistency principle emphasizes the strong statistical correlation between presentations from various views of the same instance, which potentially ignores the internal state of individual representation. One consequence is dimensional collapse (Hua et al., 2021; Jing et al., 2022), which severely hampers the diversity and expressiveness of learned representations, and commonly exists in negative-free contrastive learning methods. A pioneer work (Jing et al., 2022) blames dimensional collapse on strong augmentation along feature dimension and implicit regularization. We analyze the cause of dimensional collapse from the perspective of objective function, and draw the conclusion that the collapsed representations are a shortcut solution under the specific self-supervised objective. Statistically, the dimensional collapse presents that various representation channels (*i.e.*, dimensions) are tightly coupled and highly correlated, making the representation vectors only span a lower-dimensional subspace. Accordingly, another principle of *between-channel minimum dependence* is proposed to learn diverse and informative representations by decorrelating various representation channels. Concretely, we adopt two strategies to realize this principle: the first utilizes statistical metrics to directly decouple different representation channels, and the second regularizes representations in eigen spectral space by minishing data distribution differences along various principal directions.

To sum up, we make the following contributions over the peer works:

- We investigate what good representations should be in graph contrastive learning. As a response, from a statistical perspective, two complementary principles, cross-view maximum consistency and between-channel minimum dependence, are proposed to mine view-invariant common information and learn diverse representations.
- To learn augmentation-invariant and consistent representations across views, four statistical indicators are employed to instantiate the consistency principle and their dynamic behaviors during training are thoroughly analyzed.
- Two strategies, performing in representation space and eigen spectral space, respectively, are proposed to achieve diverse representations by decorrelating various channels.
- We provide an explanation for the behavior of negative samples from a decorrelation perspective and incorporate the negative-rely self-supervised methods into our framework.
- Empirically, extensive experiments demonstrate the sufficiency of two principles for high-quality graph representations. Ablation studies and visual analysis further reveal the working mechanism of the two principles and typical phenomena during training.

2 Related Work

2.1 Graph Contrastive Learning

Inspired by the prosperity in computer vision field, some research efforts have been devoted to generalizing contrastive learning to graph data. For most current methods, despite the differences in view design, network architecture, and contrastive objectives, their core idea is to maximize the mutual information between learned representations from various views. Enlightened by the *InfoMax* principle (Linsker, 1988), Deep Graph Infomax (DGI) (Veličković et al., 2018) and InfoGraph (Sun et al., 2020) learn node-level and graph-level representations by maximizing mutual information between patch-level representations and a graph-level summary vector based on Jenson-Shannon estimator (Nowozin et al., 2016). Embedding the *InfoNCE* (Gutmann & Hyvärinen, 2010) loss into the contrastive framework, GraphCL (You et al., 2020) utilizes various priors to design four types of graph augmentations and systematically investigates the influences of various combinations of graph augmentations on downstream tasks. MVGRL (Hassani & Khasahmadi, 2020) employs graph diffusion (Klicpera et al., 2019) to generate new views and studies the effects of different mutual information estimators. GRACE (Zhu et al., 2020) and GCA (Zhu et al., 2021) utilize node attribute masking and edge perturbation to construct multiple augmented views and adopt the *InfoNCE* loss as the objective function. From the perspective of information theory, InfoGCL (Xu et al., 2021) explains how to construct contrastive learning models for particular tasks and datasets. Despite the diversity, most current works are carried out within the framework of information theory. This paper attempts to illustrate from a statistical point of view what makes for good graph contrastive learning.

2.2 Statistical Correlation

Statistical correlation analysis investigates the degree of dependence between random variables and presents it with proper indicators. RV-coefficient (Robert & Escoufier, 1976) can be regarded as a multivariate generalization of the squared Pearson correlation (Benesty et al., 2009) and measure the linear closeness of two high-dimensional variables, which is broadly applied in the bioinformatics field. Mutual information (Gutmann & Hyvärinen, 2010) can capture non-linear dependencies between two high-dimensional variables. Due to the explicit dependence on the probability distribution, it is intractable to directly obtain mutual information from empirical data. Hilbert-Schmidt Independence Criterion (Gretton et al., 2005) estimates the correlation between two variables in Reproducing Kernel Hilbert Space. Based on characteristic function, distance correlation (Székely et al., 2007) constructs a measurement to describe non-linear dependencies between two high-dimensional variables.

3 Two Principles for Graph Contrastive Learning

3.1 Notations an Preliminaries

Notations. A graph is denoted by $G(\mathbf{A},\mathbf{X}) \in \mathcal{G}$ with node set $\mathcal{V} = \{v_1, ..., v_N\}$, where $|\mathcal{V}| = N$ indicates the number of nodes. Each node $v_i \in \mathcal{V}$ has a D-dimensional feature vector $\mathbf{x}_i \in \mathbb{R}^D$. Feature matrix $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]^\top \in \mathbb{R}^{N\times D}$ contains feature information of all nodes within graph G and adjacency matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ describes the topology connection between nodes.

Graph View Generation. In this paper, multiple views are artificially generated through graph augmentation. A transformation $\tau \in \mathcal{T} : G(\mathbf{A},\mathbf{X}) \to G'(\mathbf{A}',\mathbf{X}')$ maps the original graph to an augmented version, where $\mathcal{T}$ denotes the whole augmentation function space. Specifically, the graph augmentation τ is jointly realized from two aspects of topology structure and feature. In topology-level, *edge removal* randomly removes edges of a certain ratio p_e from the original graph structure. In feature-level, for feature matrix $\mathbf{X} \in \mathbb{R}^{N\times D}$, *feature masking* randomly sets feature channels of a specific number $D \cdot p_f$ to zero, where p_f is the masking ratio (Zhu et al., 2020).

Basic Framework. This paper focuses on generating high-quality node-level representations, and the basic model architecture follows the common practice of previous works. First, two graph augmentation functions τ_A and τ_B are randomly sampled from $\mathcal{T}$, and then two various views $G'_A(\mathbf{A}'_A, \mathbf{X}'_A) = \tau_A(G)$ and $G'_B(\mathbf{A}'_B, \mathbf{X}'_B) = \tau_B(G)$ are generated from their respective transformations. The two augmented versions are fed into a shared neural encoder $f_\theta(\cdot)$ with learnable parameters θ to obtain representations $\widetilde{\mathbf{H}}_A = [\tilde{\mathbf{h}}_1^A, ..., \tilde{\mathbf{h}}_N^A]^\top \in \mathbb{R}^{N\times d}$ and $\widetilde{\mathbf{H}}_B = [\tilde{\mathbf{h}}_1^B, ..., \tilde{\mathbf{h}}_N^B]^\top \in \mathbb{R}^{N\times d}$. For the ease of subsequent discussion, $\widetilde{\mathbf{H}}_A$ and $\widetilde{\mathbf{H}}_B$ are further normalized into $\mathbf{H}_A = [\mathbf{h}_1^A, ..., \mathbf{h}_N^A]^\top$ and $\mathbf{H}_B = [\mathbf{h}_1^B, ..., \mathbf{h}_N^B]^\top$ along sample direction so that each representation channel in normalized representation matrix is subject to a distribution with 0-mean and 1-standard deviation.

3.2 Cross-view Maximum Consistency

3.2.1 Formulating Augmentation as A Mixing Process and Consistency Principle

For an augmentation transformation $\tau \in \mathcal{T}$, we assume that the augmented view G' is generated through a mixing process between the original graph G and a random unknown noise graph $\hat{G}$:

$$G' = \tau(G) = g(\{G, \hat{G}\}), \tag{1}$$

where $g : \mathcal{G} \times \mathcal{G} \to \mathcal{G}$ is an unknown mixing function.

Then, the multi-view generation process can be modeled as

$$G'_A = \tau_A(G) = g(\{G, \hat{G}_A\}), \quad G'_B = \tau_B(G) = g(\{G, \hat{G}_B\}). \tag{2}$$

$\hat{G}_A$ and $\hat{G}_B$ are considered independent, that is $p(\hat{G}_A, \hat{G}_B) = p(\hat{G}_A) \cdot p(\hat{G}_B)$, for τ_A and τ_B are acquired from $\mathcal{T}$ separately. We assume that the information within each view G'_A (or G'_B) can be partitioned into two separated parts:

1) a common (shared) component invariant across (G'_A, G'_B), which is related to G;
2) an individual (private) component variant across (G'_A, G'_B), which is relevant to $\hat{G}_A$ (or $\hat{G}_B$).

Thus, the interested problem is specified as extracting augmentation-invariant essential information while discarding view-specific private contents. We stress that our goal is not to recover the original graph G, but to obtain high-quality and essence-revealing representations, facilitating subsequent downstream tasks. To this end, we expect that a smooth mapping $f_\theta : \mathcal{G} \to \mathcal{H}$ can extract and project the common information from G' to the *representation space* $\mathcal{H} \subseteq \mathbb{R}^d$. Under such a mapping, the representations from various views are consistent in characterizing objects, that is, they should be correlated from each other. We use a d-dimensional random variable H_A with distribution $p(H_A)$ to describe the projection results (that is, node representations) from the view G'_A, where the representation matrix $\mathbf{H}_A \in \mathbb{R}^{N \times d}$ contains N empirical observations of H_A. Symmetrically, the same setting applies to variable H_B and representations $\mathbf{H}_B$. The sample pairs,

$$(\mathbf{h}_1^A, \mathbf{h}_1^B), \ldots, (\mathbf{h}_N^A, \mathbf{h}_N^B) \in \mathcal{H} \times \mathcal{H}, \tag{3}$$

are acquired from the joint distribution $p(H_A, H_B)$ of H_A and H_B, each of which characterizes the same node from two various views.

To make the representations learned by the encoder f_θ can truly reflect the common information across various views, we propose the following *consistency principle*:

Principle 1 [Cross-View Maximum Consistency (CVMC)]. *The representations from various views of the same instance should be consistent in describing objects and statistically correlated.*

The consistency principle addresses the problem of extracting common information as strengthening the consistency (or correlation) between variables H_A and H_B under any augmentation transformations and can be formulated as

$$\max_{f_\theta} Cor(H_A, H_B), \tag{4}$$

where $Cor(\cdot)$ is a statistical indicator measuring consistency or correlation between random variables.

3.2.2 Four Statistical Indicators

Here, we introduce four statistical indicators to instantiate $Cor(\cdot)$ in Eq. (4). To be formal, two general variables are first defined for the following statement. We denote P_{YZ} and $P_Y P_Z$ as the joint distribution and the product of marginal distributions of two variables $Y \in \mathcal{Y} \subseteq \mathbb{R}^{d_1}$ and $Z \in \mathcal{Z} \subseteq \mathbb{R}^{d_2}$. Several observed samples $\{(\mathbf{y}_i, \mathbf{z}_i) | i = 1, \ldots, n\}$ are acquired from the joint distribution. $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]^\top \in \mathbb{R}^{n \times d_1}$ and $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n]^\top \in \mathbb{R}^{n \times d_2}$ collect the samples with respect to individual variables.

Hilbert-Schmidt Independence Criterion. Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005) measures both linear and nonlinear correlation between two random variables in reproducing kernel Hilbert space (RKHS). Taking two transformations $\phi : \mathcal{Y} \to \mathcal{Y}'$ and $\psi : \mathcal{Z} \to \mathcal{Z}'$, where $\mathcal{Y}'$ and $\mathcal{Z}'$ denote RKHS, HSIC evaluates the cross-covariance operators between the RKHSs of Y and Z:

$$\mathit{HSIC}(Y, Z) = \|\mathbb{E}_{Y,Z \sim P_{YZ}}[\phi(Y)\psi(Z)^\top] - \mathbb{E}_{Y \sim P_Y}[\phi(Y)]\mathbb{E}_{Z \sim P_Z}[\psi(Z)]^\top\|_{HS}^2, \tag{5}$$

where $\| \cdot \|_{HS}^2$ denotes the Hilbert-Schmidt norm.

Given finite samples $\mathbf{Y}$ and $\mathbf{Z}$, under kernel functions $k(\cdot, \cdot)$ and $l(\cdot, \cdot)$, HSIC can be directly estimated:

$$HSIC(Y,Z) = \frac{1}{(n-1)^2} tr(\mathbf{KJLJ}), \tag{6}$$

where $tr(\cdot)$ denotes the matrix trace, $\mathbf{K}_{ij} = k(\mathbf{y}_i, \mathbf{y}_j)$ and $\mathbf{L}_{ij} = l(\mathbf{z}_i, \mathbf{z}_j)$ are the kernel Gram matrices, and $\mathbf{J} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top \in \mathbb{R}^{n\times n}$ is the centering matrix, in which $\mathbf{I}$ is the identity matrix and all elements in $\mathbf{1} \in \mathbb{R}^n$ are 1. $HSIC(Y,Z) = 0$ holds if and only if Y and Z are independent, and larger values mean more correlation. Without relying on the explicit probability distribution, HSIC is computationally convenient for empirical estimation.

Distance Correlation. From the perspective of characteristic functions, Distance Correlation (DC) (Székely et al., 2007) measures both linear and nonlinear association between two arbitrary dimensional variables. For $\mathbf{Y}$ and $\mathbf{Z}$, define

$$\begin{aligned} &a_{ij} = \|\mathbf{y}_i - \mathbf{y}_j\|_2, \quad \overline{a}_{i\cdot} = \tfrac{1}{n}\textstyle\sum_{j=1}^n a_{ij}, \\ &\overline{a}_{\cdot j} = \tfrac{1}{n}\textstyle\sum_{i=1}^n a_{ij}, \quad \overline{a}_{\cdot\cdot} = \tfrac{1}{n^2}\textstyle\sum_{i=1}^n\sum_{j=1}^n a_{ij}, \quad A_{ij} = a_{ij} - \overline{a}_{i\cdot} - \overline{a}_{\cdot j} + \overline{a}_{\cdot\cdot}\ . \end{aligned} \tag{7}$$

Analogously, define $b_{ij} = \|\mathbf{z}_i - \mathbf{z}_j\|_2$ and $B_{ij} = b_{ij} - \overline{b}_{i\cdot} - \overline{b}_{\cdot j} + \overline{b}_{\cdot\cdot}$ for $i,j = 1,2,\ldots,n$. The empirical *distance covariance* $V(Y,Z)$ is the nonnegative number defined by

$$V^2(Y,Z) = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n A_{ij}B_{ij}\ . \tag{8}$$

Then, the distance correlation $DC(Y,Z)$ is defined by

$$DC^2(Y,Z) = \begin{cases} \frac{V^2(Y,Z)}{\sqrt{V^2(Y,Y)\cdot V^2(Z,Z)}}, & V^2(Y,Y)\cdot V^2(Z,Z) > 0\ , \\ 0, & V^2(Y,Y)\cdot V^2(Z,Z) = 0\ . \end{cases} \tag{9}$$

Distance correlation satisfies $0 \le DC(Y,Z) \le 1$. $DC(Y,Z) = 0$ characterizes independence of Y and Z, and the strength of the correlation between two variables is positively correlated with the value of distance correlation.

RV-coefficient. RV-coefficient (RV) (Robert & Escoufier, 1976) can evaluate linear correlation between two arbitrary-dimensional random variables. Given $\mathbf{Y}$ and $\mathbf{Z}$, the RV-coefficient between variables Y and Z is defined as

$$RV(Y,Z) = \frac{tr(\mathbf{Y}\mathbf{Y}^\top\mathbf{Z}\mathbf{Z}^\top)}{\sqrt{tr[(\mathbf{Y}\mathbf{Y}^\top)^2]\cdot tr[(\mathbf{Z}\mathbf{Z}^\top)^2]}}. \tag{10}$$

RV-coefficient projects the linear correlation between two variables into the interval $[0,1]$, and larger values reflect stronger linear dependence. $RV(Y,Z) = 0$ stands if and only two variables are linearly independent from each other.

Matrix Correlation. The simple Matrix Correlation (MC) (Smilde et al., 2008) can convey the linear association between two equal-dimensional random variables. For $\mathbf{Y} \in \mathbb{R}^{n\times d_1}$ and $\mathbf{Z} \in \mathbb{R}^{n\times d_2}$, if $d_1 = d_2$, MC is defined as

$$MC(Y,Z) = \left|\frac{tr(\mathbf{Y}^\top\mathbf{Z})}{\sqrt{tr(\mathbf{Y}^\top\mathbf{Y})\cdot tr(\mathbf{Z}^\top\mathbf{Z})}}\right|. \tag{11}$$

where $|\cdot|$ denotes the absolute value. The values of MC drop in the interval $[0,1]$, and larger values mirror stronger linear correlation.

$Cor(\cdot)$ in Eq. (4) can be instantiated as one of the above statistical indicators. A summary and comparison between them are summarized in Table 4 of Appendix K.

3.3 Between-Channel Minimum Dependence

3.3.1 Dimensional Collapse and Decorrelation Principle

Focusing on learning consistent representations across views, the consistency principle imposes no constraints on the internal state of individual representation, which potentially leads to the issue of dimensional collapse. As shown in the left of Figure 1, dimensional collapse presents that various representation channels are coupled with each other and express similar information, which weakens expressive ability of the model and reduces diversity of representations.

Proposition 1. *Given a self-supervised objective $\mathcal{L}$, the neural models tend to find low-rank solutions satisfying the objective $\mathcal{L}$, unless there are relevant constraints in objective function or network architecture that can explicitly or implicitly restrain this tendency.*

Proposition 1 regards dimensional collapse as a shortcut optimal solution, and blames degenerated solutions on the combined effect of two factors, "lazy" behaviors of neural networks and lack of relevant constraints avoiding collapse. A detailed discussion is placed in Appendix A.

To prevent dimensional collapse and learn diverse and informative graph representations, we propose the following *decorrelation principle* to guide the design of objective functions:

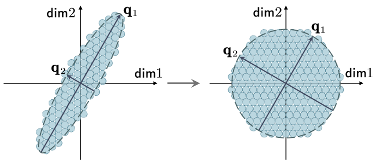


Figure 1: Dimensional collapse and effects of decorrelation in two-dimensional representation space. $\mathbf{q}_1$ and $\mathbf{q}_2$ represent two principal directions of data, relating to two eigenvectors of covariance matrix.

Principle 2 [Between-Channel Minimum Dependence (BCMD)]**.** *Various representation channels should be statistically independent from each other.*

3.3.2 Two Strategies

Here, two strategies are proposed to actualize the decorrelation principle.

Direct Channel Decorrelation. We reformulate variables H_A and H_B as

$$H_A = (H_A^1, H_A^2, \dots, H_A^d), \quad H_B = (H_B^1, H_B^2, \dots, H_B^d), \tag{12}$$

where H_A^i (H_B^i) is a one-dimensional random variable associated with the i-th channel of representation matrix $\mathbf{H}_A$ ($\mathbf{H}_B$). A natural approach to realize the decorrelation principle is to directly relieve the statistical dependence between various channels. Our strategy of direct channel decorrelation (DCD) performs from two aspects of intra-view and inter-view:

$$\mathcal{L}_{dcd} = \frac{1}{d(d-1)} \sum_{i=1}^{d} \sum_{j=1, j \neq i}^{d} (\underbrace{Cor(H_A^i, H_A^j) + Cor(H_B^i, H_B^j)}_{\text{intra-view}} + \underbrace{Cor(H_A^i, H_B^j)}_{\text{inter-view}}), \tag{13}$$

where $Cor(\cdot)$ denotes a statistical indicator. For simplicity, we instantiate it as the square of Pearson correlation coefficient in practice.

Spectral Regularization. One appearance of dimensional collapse is that data points show different forms of distributions along various principal directions, appearing loose distributions in some directions (*e.g.*, $\mathbf{q}_1$ in the left of Figure 1) with larger variance and presenting tight distributions in other directions (*e.g.*, $\mathbf{q}_2$ in the left of Figure 1) with smaller variance.

Property 1. *For covariance matrix $\mathbf{\Sigma_H} = \frac{1}{N}\mathbf{H}^\top\mathbf{H} \in \mathbb{R}^{d \times d}$, which has d eigenvectors $[\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_d]$ associated with d eigenvalues $[\lambda_1, \lambda_2, \dots, \lambda_d]$, the variance of data $\mathbf{H}$ along the k-th principal direction (i.e., the direction of $\mathbf{q}_k$) is equal to λ_k.*

Proof. Please refer to Appendix D. □

When data distributions along various principal directions take the same forms (*i.e.*, have equal variances), dimensional collapse disappears naturally. To this end, we put forward a spectral regularization (SR) strategy to reduce the distribution difference:

$$\mathcal{L}_{sr} = std(\lambda(\mathbf{\Sigma}_A)) + std(\lambda(\mathbf{\Sigma}_B)), \tag{14}$$

where $\mathbf{\Sigma}_A = \frac{1}{N}\mathbf{H}_A^\top\mathbf{H}_A$ and $\mathbf{\Sigma}_B = \frac{1}{N}\mathbf{H}_B^\top\mathbf{H}_B$ are covariance matrices, $\lambda(\cdot)$ denotes all eigenvalues of a matrix, and $std(\cdot)$ indicates the standard deviation of all eigenvalues. When $\mathbf{H}_A$ ($\mathbf{H}_B$) is completely decorrelated, the eigenvalues of its covariance matrix are equal. Our strategy of spectral regularization realizes the decorrelation principle in eigen spectral space, which relaxedly performs ZCA Whitening (Bell & Sejnowski, 1997) operation by minimizing the objective (14). The differences and connections between our strategies and ZCA Whitening are discussed in Appendix F.

Theorem 1. *For representation matrix $\mathbf{H} \in \mathbb{R}^{N \times d}$, corresponding to a d-dimensional variable H, the entropy of H under empirical data $\mathbf{H}$ is maximized when $\mathbf{H}$ is completely decorrelated.*

Proof. Please refer to Appendix E. □

As entropy is a measurement of the amount of information, Theorem 1 indicates that the decorrelation principle facilitates the extraction of more sufficient common information. From this perspective, the decorrelation principle can be seen as a complement and enhancement to the consistency principle. Besides, in conjunction with Theorem 1, Proposition 1 indicates that the training process of neural network presents a phenomenon of entropy reduction.

3.4 Overall Framework and Objective Function

Overall Framework. The overall contrastive learning framework under two principles is as follows

$$\begin{tikzcd}
 & & & & \mathit{CVMC} \arrow[dl, bend right] \arrow[dr, bend left] & & & & \\
G \arrow[r, "\tau_A"] & G'_A \arrow[r, "f_\theta(\cdot)"] & \widetilde{\mathbf{H}}_A \arrow[r, "Norm"] & \mathbf{H}_A & & \mathbf{H}_B & \widetilde{\mathbf{H}}_B \arrow[l, "Norm"'] & G'_B \arrow[l, "f_\theta(\cdot)"'] & G \arrow[l, "\tau_B"'] \\
 & & & & \mathit{BCMD} \arrow[ul, bend left] \arrow[ur, bend right] & & & &
\end{tikzcd},$$

where $Norm$ denotes the normalization operation along sample direction.

Objective Function. The learning objective related to two principles is

$$\mathcal{L} = \alpha \mathcal{L}_{cvmc} + \beta \mathcal{L}_{bcmd}, \tag{15}$$

where α and β are two weighted coefficients, $\mathcal{L}_{bcmd}$ can be instantiated as $\mathcal{L}_{dcd}$ or $\mathcal{L}_{sr}$, and $\mathcal{L}_{cvmc}$ is related to Eq. (4). We call $\mathcal{L}_{cvmc}$ *consistency term* and $\mathcal{L}_{bcmd}$ *decorrelation term.*

Connection to Mutual Information Maximization. Appendix B explains the connection between two principles and negative-rely methods based on mutual information maximization.

Relationship with Graph Smoothness. Appendix G demonstrates the relationship between decorrelation principle and graph smoothness.

Comparison with Two Peer Works. The comparisons with two peer works, Spectral Contrastive Loss (HaoChen et al., 2021) and VICReg (Bardes et al., 2022), are put in Appendix C.

4 Experiments

4.1 Datasets and Experimental Setup

To assess our approach, seven widely used benchmark datasets are adopted for experimental study, including three citation networks **Cora**, **Citeseer** and **Pubmed** (Sen et al., 2008), two co-purchase networks **Amazon-Computers** and **Amazon-Photo** (Shchur et al., 2019), and two co-authorship network **Coauthor-CS** and **Coauthor-Physics** (Shchur et al., 2019). The details of the datasets are placed in Appendix H. The model is implemented by Graph Convolutional Network (GCN) (Kipf & Welling, 2016a). The model parameters are initialized via Xavier initialization (Glorot & Bengio, 2010) and trained by Adam optimizer (Kingma & Ba, 2017). The networks are first trained to in a fully unsupervised manner, and the learned representations are evaluated by a simple linear classifier.

4.2 Comparison Experiments and Ablation Studies

Comparison with State-of-the-Art. We compare our framework under two principles with other state-of-the-art methods on node classification task under the simple linear classifier. The unsupervised baselines cover DeepWalk (Perozzi et al., 2014), GAE (Kipf & Welling, 2016b), DGI (Veličković et al., 2018), GMI (Peng et al., 2020), GRACE (Zhu et al., 2020), GCA (Zhu et al., 2021), G-BT (Bielak et al., 2021) InfoGCL (Xu et al., 2021), CCA-SSG (Zhang et al., 2021) and MVGRL (Hassani & Khasahmadi, 2020). Besides, some supervised models including multi-layer perceptron (MLP), C&S (Huang et al., 2021), GCN (Kipf & Welling, 2016a) and GAT (Veličković et al., 2017) are also adopted as baselines. For our method, we investigate various combinations of four statistical indicators for consistency principle and two strategies for decorrelation principle. The experimental results are summarized in Table 1. Regarding ours, the left of "-" denotes the employed statistical indicator while its right represents the adopted strategy to decouple various channels. The linear kernel function is applied to realizing HSIC. Overall, three main findings can be observed:

1) Our approach acquires the best performance on six of seven datasets, demonstrating that the two principles can instruct the model to learn high-quality representations.

Table 1: Node classification accuracy with standard deviation in percentage on seven datasets. OOM indicates Out-Of-Memory on a 32GB GPU. The **bold** font highlights the best results.

	Algorithm	Cora	Citeseer	Pubmed	Computers	Photo	CS	Physics
	MLP	57.8 ± 0.2	54.2 ± 0.1	72.8 ± 0.2	79.81 ± 0.06	86.36 ± 0.08	91.32 ± 0.11	94.21 ± 0.04
	GCN	81.5	70.3	79.0	86.51 ± 0.54	92.42 ± 0.22	93.03 ± 0.31	95.65 ± 0.16
	GAT	83.0 ± 0.7	72.5 ± 0.7	79.0 ± 0.3	86.93 ± 0.29	92.56 ± 0.35	92.31 ± 0.24	95.47 ± 0.15
	Plain Linear + C&S	81.1 ± 0.3	72.1 ± 0.4	78.4 ± 0.2	87.23 ± 0.15	92.95 ± 0.12	93.11 ± 0.15	95.32 ± 0.06
Unsupervised	DeepWalk	68.5 ± 0.5	49.8 ± 0.2	66.2 ± 0.7	85.68 ± 0.06	89.44 ± 0.11	84.61 ± 0.22	91.77 ± 0.15
	GAE	72.1 ± 0.5	66.5 ± 0.4	71.8 ± 0.6	85.27 ± 0.19	91.62 ± 0.13	90.01 ± 0.71	94.92 ± 0.07
	GMI	83.0 ± 0.3	72.4 ± 0.1	79.9 ± 0.2	82.21 ± 0.31	90.68 ± 0.17	OOM	OOM
	GRACE	80.5 ± 0.3	69.2 ± 0.2	80.1 ± 0.2	86.53 ± 0.28	92.24 ± 0.17	92.98 ± 0.05	95.32 ± 0.03
	CCA-SSG	84.2 ± 0.4	73.1 ± 0.3	81.6 ± 0.4	88.74 ± 0.28	93.14 ± 0.14	93.31 ± 0.22	95.38 ± 0.06
	GCA	80.7 ± 0.2	69.8 ± 0.4	79.5 ± 0.5	87.85 ± 0.31	92.49 ± 0.09	93.10 ± 0.01	**95.68 ± 0.05**
	G-BT	84.0 ± 0.4	73.0 ± 0.3	80.7 ± 0.4	88.14 ± 0.33	92.63 ± 0.44	92.95 ± 0.17	95.07 ± 0.17
	InfoGCL	83.5 ± 0.3	73.5 ± 0.4	79.1 ± 0.2	-	-	-	-
	DGI	82.3 ± 0.6	71.8 ± 0.7	76.8 ± 0.6	83.95 ± 0.47	91.61 ± 0.22	92.15 ± 0.63	94.51 ± 0.52
	MVGRL	83.7 ± 0.6	**73.6 ± 0.3**	79.9 ± 0.2	87.52 ± 0.11	91.74 ± 0.07	92.11 ± 0.12	95.33 ± 0.03
	DC-DCD (Ours)	83.2 ± 0.5	72.6 ± 0.3	79.1 ± 0.6	88.41 ± 0.32	93.02 ± 0.15	93.58 ± 0.13	95.34 ± 0.09
	DC-SR (Ours)	83.6 ± 0.4	72.5 ± 0.4	79.2 ± 0.5	88.49 ± 0.33	**93.31 ± 0.13**	93.41 ± 0.24	95.50 ± 0.04
	RV-DCD (Ours)	83.6 ± 0.4	72.8 ± 0.4	79.4 ± 0.6	88.57 ± 0.28	92.93 ± 0.22	93.53 ± 0.18	95.47 ± 0.02
	RV-SR (Ours)	83.2 ± 0.5	72.4 ± 0.5	78.8 ± 0.7	88.49 ± 0.27	92.83 ± 0.19	93.48 ± 0.16	95.46 ± 0.09
	HSIC-DCD (Ours)	82.9 ± 0.5	72.5 ± 0.4	79.5 ± 0.4	88.39 ± 0.28	93.04 ± 0.18	93.51 ± 0.22	95.45 ± 0.04
	HSIC-SR (Ours)	82.9 ± 0.6	72.2 ± 0.6	79.1 ± 0.5	88.16 ± 0.37	92.87 ± 0.16	93.32 ± 0.31	95.45 ± 0.07
	MC-DCD (Ours)	**84.5 ± 0.3**	**73.6 ± 0.3**	**81.7 ± 0.3**	88.70 ± 0.31	93.14 ± 0.15	**93.60 ± 0.08**	95.42 ± 0.12
	MC-SR (Ours)	84.4 ± 0.4	73.5 ± 0.4	81.5 ± 0.4	**88.78 ± 0.25**	93.09 ± 0.14	93.56 ± 0.11	95.38 ± 0.08

2) Two decorrelation strategies achieve almost the same effect, implying their similar inherence in driving representations toward diversity.

3) In general, employing matrix correlation (MC) to maximize cross-view consistency makes the best benefit. One reason is that MC, measuring the linear correlation between two equal-dimensional variables, places higher requirements on consistency and thus demands networks to extract more common information.

Abalation Studies on Two Principles. We conduct controlled experiments to assess the impact of individual principle and the experimental results are presented in Table 2. Only optimizing the consistency term without decoupling various channels achieves view-invariant yet non-informative representations, thus leading to suboptimal results. Within expectation, only considering the decorrelation between various dimensions, making the model learn decoupled yet meaningless representations, results in poor performance.

Table 2: Ablation study on two principles. "Δ" indicates the absence of this term.

Dataset	Cora	Pubmed	CS	Physics
DC-Δ	78.8	70.4	91.04	95.22
RV-Δ	75.2	69.8	91.12	95.23
HSIC-Δ	74.9	68.5	91.18	95.17
MC-Δ	79.5	73.8	92.01	95.14
Δ-DCD	53.6	48.2	26.63	54.48
Δ-SR	52.7	47.7	25.93	57.53

4.3 Exploratory Experiments on the Consistency Principle

Under the principle of cross-view maximum consistency, we attempt to extract and map the common information from various augmented views to consistent representations by maximizing their statistical dependencies. The common information among various views is inherent and does not vary with the statistical indicators. Intuitively, good representations learned under a specific measurement should accordingly satisfy other indicators. We visualize the training dynamic on Cora and Pubmed in Figure 2. Concretely, except for the indicator supervising model training, we also show synergistic changes of other indicators. For mutual information, here, the *InfoNCE* (Gutmann & Hyvärinen, 2010) is used as mutual information estimator. It can be observed that the curves of various indicators show a similar trend in each figure. This phenomenon suggests that the consistency principle indeed guides the model to extract common information and learn consistent representations.

4.4 Exploratory Experiments on the Decorrelation Principle

This subsection aims at providing empirical supports for Proposition 1 and evaluates the ability of two proposed objectives to prevent dimensional collapse and learn diverse representations. Figure 3 shows relevant experimental results on Cora and Coauthor-CS, where MC is employed to maximize cross-view consistency, and more experiments are placed in Appendix I. The left of each subfigure in

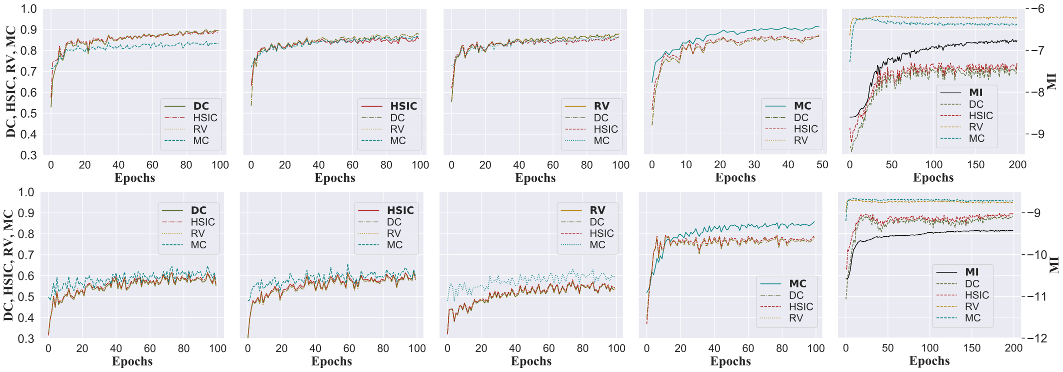
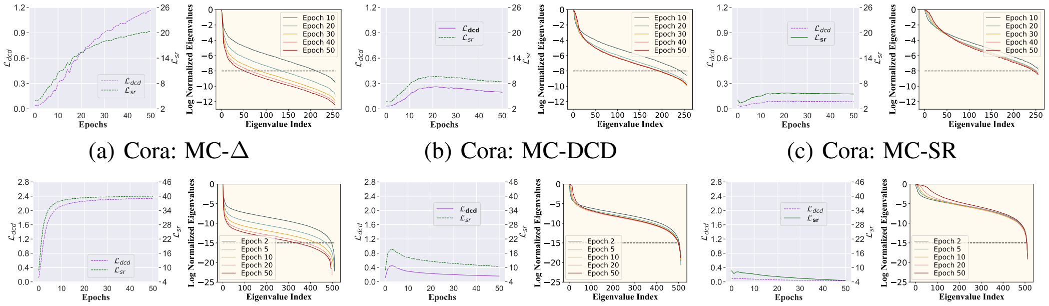


Figure 2: The variation trend of various statistical indicators during the training on Cora and Pubmed datasets. The top row: Cora; the bottom row: Pubmed. The kernel function of HSIC is Gaussian kernel. In each figure, the solid line denotes the optimized indicator in the training stage, which is also **bolded** in the legend, while the dashed lines describe the coordinated variations of other indicators.



(a) Cora: MC-Δ

(b) Cora: MC-DCD

(c) Cora: MC-SR

(d) CS: MC-Δ

(e) CS: MC-DCD

(f) CS: MC-SR

Figure 3: The joint changes of two objectives and the evolution of eigenvalues of covariance matrix. The solid line in the left of each subfigure presents the change of the optimized objective, while the dashed line describes the corresponding response of another objective.

Figure 3 presents the joint variations of $\mathcal{L}_{dcd}$ and $\mathcal{L}_{sr}$ under various settings, while the right shows the evolution of eigenvalues of covariance matrix of representations. In the left of Figure 3(a, d), at the beginning of training, $\mathcal{L}_{dcd}$ and $\mathcal{L}_{sr}$ are small, suggesting that the representations do not fall into dimensional collapse. Without decorrelation term, both $\mathcal{L}_{dcd}$ and $\mathcal{L}_{sr}$ continuously increase with the training process, demonstrating the representations gradually tend to dimensional collapsed solution. Accordingly, the right subfigure shows that the distributions of eigenvalues of covariance matrix under various epochs present significant differences. The more the iterations, the more uneven the eigenvalues. These phenomena support Proposition 1. As shown in Figure 3(b, c, e, f), under $\mathcal{L}_{dcd}$ or $\mathcal{L}_{sr}$, the tendency for dimensional collapse is suppressed, and eigenvalue distributions under different epochs become more consistent. In every subfigure, two losses present almost the same trend.

5 Conclusion

In this paper, we have concentrated on what makes for good graph representations in self-supervised learning and proposed two principles to construct a general framework. Statistically, the first principle requires the representations of various views of the same instance to correlate to each other while the second one demands the independence between various representation channels. The theoretical and empirical results demonstrate the rationality and effectiveness of the two principles. Ablation experiments and visual studies further uncover the working mechanisms of individual principle. Besides, we analyze the relationship between mutual information maximization and the two principles, and treat it as an instance under our framework. While our framework allows for a variety of statistical indicators for instantiation, it is an interesting and promising topic how to choose the most appropriate metrics for specific tasks and datasets. We believe that our work opens avenues for designing more effective graph self-supervised learning architectures and objective functions. Our framework is not just for graph data, and it is left for future work to generalize it to other fields.

ETHICS STATEMENT

We have carefully read ICLR Code of Ethics and promise to adhere to it. Our paper does not involve ethics issues and potential negative societal impacts mentioned in ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

All experiments run on a server with Intel(R) Xeon(R) 6230R CPU @ 2.1GHz and TITAN RTX GPU in Ubuntu 18.04. For the experiments in Table 1, we adopt the public splits on Cora, Citeser and Pubmed, and a random 1:1:8 split for training/validation/testing on the other datasets without standard split. For all unsupervised learning methods, we first train networks to learn node representations in a fully unsupervised manner and then evaluate quality of the learned representations by training and testing a simple linear classifier. To make a fair comparison, for the methods without adopting the same settings as ours, we conduct experiments to get related results based on their official source code. Our source code is available in an anonymous repository `https://github.com/ICLR2023-ID3781/ICLR2023ID3781`.

REFERENCES

N.A. Ahmed and D.V. Gokhale. Entropy expressions and their estimators for multivariate distributions. *IEEE Transactions on Information Theory*, pp. 688–692, 1989.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 531–540. PMLR, 2018.

Anthony J. Bell and Terrence J. Sejnowski. The "independent components" of natural scenes are edge filters. *Vision Research*, 37:3327–3338, 1997.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pp. 1–4. Springer, 2009.

Piotr Bielak, Tomasz Kajdanowicz, and Nitesh V. Chawla. Graph barlow twins: A self-supervised representation learning framework for graphs, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 1597–1607. PMLR, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15750–15758, 2021.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256. PMLR, 2010.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic Learning Theory*, pp. 63–77. Springer Berlin Heidelberg, 2005.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, pp. 21271–21284. Curran Associates, Inc., 2020.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 297–304, 2010.

Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Advances in Neural Information Processing Systems*, pp. 5000–5011. Curran Associates, Inc., 2021.

Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 4116–4126. PMLR, 2020.

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 (CVPR)*, 2020.

Michael D Hirschhorn. The am-gm inequality. *Mathematical Intelligencer*, pp. 7–7, 2007.

Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. On feature decorrelation in self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9598–9608, 2021.

Qian Huang, Horace He, Abhay Singh, Ser-Nam Lim, and Austin Benson. Combining label propagation and simple models out-performs graph neural networks. In *International Conference on Learning Representations*, 2021.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*. OpenReview.net, 2016a.

Thomas N. Kipf and Max Welling. Variational graph auto-encoders, 2016b.

Johannes Klicpera, Stefan Weiß enberger, and Stephan Günnemann. Diffusion improves graph learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Qimai Li, Zhichao Han, and Xiao-ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the 32th AAAI Conference on Artificial Intelligence*, pp. 3538–3545. AAAI Press, 2018.

R. Linsker. Self-organization in a perceptual network. *Computer*, 21:105–117, 1988.

Qi Lyu, Xiao Fu, Weiran Wang, and Songtao Lu. Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective. In *International Conference on Learning Representations*, 2022.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 271–279, 2016.

Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aaron van den Oord, Sergey Levine, and Pierre Sermanet. Wasserstein dependency measure for representation learning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019.

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. Association for Computing Machinery, 2020.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 701–710. Association for Computing Machinery, 2014.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 8748–8763. PMLR, 2021.

Paul Robert and Yves Escoufier. A unifying tool for linear multivariate statistical methods: the rv-coefficient. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 25:257–265, 1976.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29:93–93, 2008.

Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation, 2019.

A. K. Smilde, H. A. L. Kiers, S. Bijlsma, C. M. Rubingh, and M. J. van Erk. Matrix correlations for high-dimensional data: the modified RV-coefficient. *Bioinformatics*, 25:401–405, 2008.

Fan-Yun Sun, Jordan Hoffman, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *International Conference on Learning Representations*, 2020.

Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35:2769 – 2794, 2007.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems*, pp. 6827–6839. Curran Associates, Inc., 2020.

Yao-Hung Hubert Tsai, Martin Q. Ma, Muqiao Yang, Han Zhao, Louis-Philippe Morency, and Ruslan Salakhutdinov. Self-supervised representation learning with relative predictive coding. In *International Conference on Learning Representations*, 2021.

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

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*. OpenReview.net, 2017.

Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. In *International Conference on Learning Representations*. OpenReview.net, 2018.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Dongkuan Xu, Wei Cheng, Dongsheng Luo, Haifeng Chen, and Xiang Zhang. Infogcl: Information-aware graph contrastive learning. In *Advances in Neural Information Processing Systems*, pp. 30414–30425. Curran Associates, Inc., 2021.

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

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

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

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Graph contrastive learning with adaptive augmentation. In *Proceedings of the Web Conference 2021*, WWW '21, pp. 2069–2080. Association for Computing Machinery, 2021.

A Interpretation and Discussion of Proposition 1

For convenience, we restate Proposition 1.

Proposition 1. *Given a self-supervised objective $\mathcal{L}$, the neural models tend to find low-rank solutions satisfying the objective $\mathcal{L}$, unless there are relevant constraints in objective function or network architecture that can explicitly or implicitly restrain this tendency.*

Proposition 1 actually describes the behaviors of neural networks during training under self-supervised mode. The *low-rank solutions* mean that the learned representations are redundant and non-informative, where various channels (*i.e.*, dimensions) are coupled and correlated to each other. This phenomenon is so-called dimensional collapse, restricting representations to a low-dimensional subspace, which is shown in Figure 4(a). Proposition 1 blames the dimensional collapse on the "lazy" behaviors of neural networks, that is, the networks take shortcuts to realize the objective function. When dimensional collapsed representations can satisfy the objective, the neural model does not bother to learn decoupled and diverse representations.

According to (Smilde et al., 2008; Székely et al., 2007), as far as distance correlation, RV-coefficient, and matrix correlation, the measurement of the correlation between two high-dimensional variables H_A and H_B can be maximized when they satisfy that

$$H_A = a \cdot H_B, \tag{16}$$

where a is a non-zero real number. In this circumstance, even if dimensional collapse occurs (*e.g.*, all channels in a representation matrix are equal), the principle of cross-view maximum consistency can be still realized as long as two representation matrices satisfy that $\mathbf{H}_A = a \cdot \mathbf{H}_B$. The consistency principle makes no requirements for the relationship between various channels, and the employed statistical indicators do not contain explicit or implicit constraints to prevent low-ran solutions and dimensional collapse. In this circumstance, Proposition 1 thinks that the model tends to learn simple shortcut solutions to satisfy the objective, which is accompanied by dimensional collapse. The empirical studies in subsection 4.4 confirm the rationality of Proposition 1.

Due to the insufficiency of the consistency principle, it is necessary to add additional constraints to achieve decorrelation between various dimensions. As shown in Figure 4(b), the decorrelation strategy makes for diverse and decoupled representations.

B Connection to Mutual Information Maximization

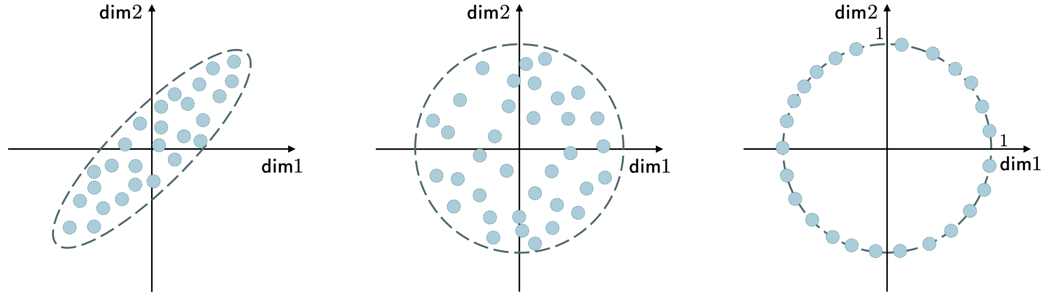
Most existing graph contrastive learning methods are based on mutual information maximization between two views, which explicitly rely on negative samples. The mutual information is defined as the KL-divergence between the joint probability distribution of two variables and the product of their marginal probability distributions. Nevertheless, explicit dependence on probability distributions makes it intractable to directly derive mutual information from empirical samples. Given the limitations of mutual information estimation, a feasible scheme is to derive lower bounds on mutual information (Gutmann & Hyvärinen, 2010; Belghazi et al., 2018; Ozair et al., 2019). As the most common one, the *InfoNCE* is defined as

$$\mathcal{L}_{\texttt{NCE}} \triangleq \mathop{\mathbb{E}}_{\substack{(\mathbf{z},\mathbf{y})\sim p_{\text{pos}} \\ \{\mathbf{z}_i^-\}_{i=1}^M \sim p_{\texttt{data}}}} \left[-\log \frac{e^{f(\mathbf{z})^\top f(\mathbf{y})/\tau}}{e^{f(\mathbf{z})^\top f(\mathbf{y})/\tau} + \sum_i e^{f(\mathbf{z}_i^-)^\top f(\mathbf{y})/\tau}} \right], \tag{17}$$

where $\tau > 0$ is a temperature hyperparameter, $M \in \mathbb{Z}_+$ is a fixed number of negative samples, $p_{\texttt{data}}$ is the data distribution over $\mathbb{R}^D$, and p_{pos} denotes the joint distribution of positive pairs over $\mathbb{R}^D \times \mathbb{R}^D$. $f : \mathbb{R}^D \to \mathcal{S}^{d-1}$ is an encoder mapping data to l_2 normalized representation vectors of dimension d, where $\mathcal{S}^{d-1}$ denotes the unit hypersphere. A previous research (Wang & Isola, 2020) analyzes the *InfoNCE* loss by decomposing it into two terms: 1) alignment term and 2) uniformity term.

The alignment term is defined as the expected distance between positive pairs in presentation space:

$$\mathcal{L}_{\texttt{align}} \triangleq \mathop{\mathbb{E}}_{(\mathbf{z},\mathbf{y})\sim p_{\text{pos}}} \|f(\mathbf{z}) - f(\mathbf{y})\|_2^\alpha, \tag{18}$$



(a) dimensional collapse (b) consistency + decorrelation (c) alignment + uniformity

Figure 4: A diagram of dimensional collapse and coping strategies in 2-dimensional space. In Figure 4(a), the dimensional collapse occurs, and the two dimensions of samples present a clear correlation. In Figure 4(b), the decorrelation strategy decouples two dimensions and makes for diverse representations. In Figure 4(c), the uniformity term brings representations to distribute on the unit hypersphere $\mathcal{S}^1$ uniformly.

where $\alpha > 0$. The alignment term plays a similar role to our consistency term, both of which attempt to improve closeness between samples from the joint distribution. Eq. (18) directly pulls two data points closer, which usually requires that $\mathbf{z}$ and $\mathbf{y}$ belong to the same domain. Differently, our scheme utilizes statistical indicators to maximize the consistency between representations from various views of the same instance (*i.e.*, representations from the joint distribution), which allows for greater flexibility because the statistical indicators can measure the correlation between two variables with different properties.

The uniformity term is defined as the logarithm of the mean Gaussian potential:

$$\mathcal{L}_{\texttt{uniform}} \triangleq \mathop{\mathbb{E}}_{\mathbf{z},\mathbf{y}\sim p_{\texttt{data}}} e^{-t\|f(\mathbf{z})-f(\mathbf{y})\|_2^2}, \tag{19}$$

where $t > 0$. As shown in Figure 4(c), the uniformity term attempts to make data points distribute on the unit hypersphere uniformly by pushing negative samples away from each other. Under the conditions that $D > d - 1$ and $p_{\texttt{data}}$ has bound density, (Wang & Isola, 2020) clarifies that it is possible to achieve perfect uniformity, where the distribution of $f(\mathbf{z})$ for $\mathbf{z} \sim p_{\texttt{data}}$ is a uniform distribution on $\mathcal{S}^{d-1}$.

Theorem 2. *When the learned representations scatter over the unit hypersphere $\mathcal{S}^{d-1}$ uniformly, that is, the distribution $p_{\texttt{rep}}$ of representation $\mathbf{h} \in \mathbb{R}^d$ is a uniform distribution on $\mathcal{S}^{d-1}$, the representations are completely decorrelated and no dimensional collapse occurs.*

Proof. We rewrite $\mathbf{h}$ as $\mathbf{h} = (h_1, h_2, \ldots, h_d)$. A feasible proof strategy is to prove that any two dimensions are linearly independent from each other. Without loss of generality, here, we consider proving the independence between the first two dimensions. The Pearson correlation coefficient is adopted as the measurement of correlation. Under the distribution $p_{\texttt{rep}}$, the Person correlation coefficient between the first two dimensions can be formulated as

$$\rho = \frac{\mathbb{E}_{\mathbf{h}\sim p_{\texttt{rep}}}\left[(h_1-\bar{h}_1)(h_2-\bar{h}_2)\right]}{\sqrt{\mathbb{E}_{\mathbf{h}\sim p_{\texttt{rep}}}\left[(h_1-\bar{h}_1)^2\right]}\sqrt{\mathbb{E}_{\mathbf{h}\sim p_{\texttt{rep}}}\left[(h_2-\bar{h}_2)^2\right]}}, \tag{20}$$

where $\bar{h}_1 = \mathbb{E}_{\mathbf{h}\sim p_{\texttt{rep}}} h_1$ and $\bar{h}_2 = \mathbb{E}_{\mathbf{h}\sim p_{\texttt{rep}}} h_2$. For the representations distribute on the zero-centered unit hypersphere uniformly, it can be known that $\bar{h}_1 = \bar{h}_2 = 0$. To prove $\rho = 0$, we just need to confirm that $\mathbb{E}_{\mathbf{h}\sim p_{\texttt{rep}}}[h_1 \cdot h_2] = 0$. As shown in Figure 5, the point $\mathbf{h}' = (h_1, h_2)$ is actually a projection of $\mathbf{h}$ onto the plane spanned by the first two dimensions. We assume that $\mathbf{h}' = (h_1, h_2)$ is subject to a distribution p' on the two-dimensional projection plane. p' can be formulated as

$$p' = \int_0^1 p'(r)\mathrm{d}r, \quad \text{where } p'(r) = \left\{ \begin{array}{ll} p', & \|\mathbf{h}'\|_2 = r \\ 0, & \|\mathbf{h}'\|_2 \neq r \end{array} \right. . \tag{21}$$

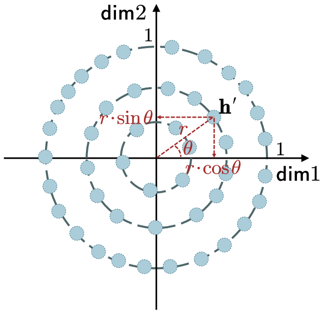


Figure 5: An illustration of projections in the two-dimensional projection plane. The points on the circle of a specific radius are subject to a uniform distribution.

Hence, we further have the following formula:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{h}\sim p_{\text{rep}}}[h_1\cdot h_2]\\
=&\mathbb{E}_{\mathbf{h}'\sim p'}[h_1\cdot h_2]\\
=&\mathbb{E}_{\mathbf{h}'\sim\int_0^1 p'(r)\mathrm{d}r}[h_1\cdot h_2]\\
=&\int_0^1\left[\int_{\mathbf{h}'\sim p'(r)}[h_1\cdot h_2]\cdot p'(r)\mathrm{d}\mathbf{h}'\right]\mathrm{d}r
\end{aligned}\quad. \tag{22}$$

For a point $\mathbf{h}'=(h_1,h_2)$ with $\|\mathbf{h}'\|_2=r$, whose angle with the axis of the first dimension is θ, it can be known that $h_1=r\cdot\cos\theta$ and $h_2=r\cdot\sin\theta$. For the representations distribute on the hypersphere $\mathcal{S}^{d-1}$ uniformly, on the two-dimensional projection plane, the distribution of the points on the circle of radius r is also a uniform distribution. Thus, it can be known that $\int_{\mathbf{h}'\sim p'(r)}[h_1\cdot h_2]\cdot p'(r)\mathrm{d}\mathbf{h}'\propto\int_0^{2\pi}r^2\sin\theta\cos\theta\mathrm{d}\theta$. It is obvious that $\int_0^{2\pi}r^2\sin\theta\cos\theta\mathrm{d}\theta$ is equal to 0. Naturally, we can know that $\mathbb{E}_{\mathbf{h}\sim p_{\text{rep}}}[h_1\cdot h_2]=0$ in Eq. 22, meaning $\rho=0$ in Eq. (20). The above proof process can be applied to any two dimension. Thus, the representations are completely decorrelated without dimensional collapse. We complete the proof. □

Theorem 2 demonstrates that the uniformity term potentially achieves the effect of decoupling various dimensions by making the representations scatter uniformly on the unit hypersphere. In this sense, the uniformity term can be regarded as a concrete implementation of our principle of between-channel minimum dependence.

The above analysis demonstrates that the two terms obtained by decomposing the *InfoNCE* loss can be regarded as the concrete implements of our two principles. The mutual information maximization can be explained based on the two principles, which can be seen as an instance of our framework.

C Comparison with Two Peer Works

C.1 Comparison with Spectral Contrastive Loss

Like our proposed spectral regularization, a peer work (HaoChen et al., 2021) also adopts spectral decomposition to design contrastive loss function. However, our strategy differs from the spectral contrastive loss in (HaoChen et al., 2021) in many aspects. **1) Operation object.** (HaoChen et al., 2021) first constructs an weighted adjacency matrix for all augmented data samples and performs spectral decomposition for the *normalized adjacency matrix*. The object of our spectral regularization is the *covariance matrix* of node representations. **2) Focus.** Our SR strategy focus on reducing the differences of various *eigenvalues* while (HaoChen et al., 2021) concentrates on generating *eigenvectors* as embeddings. **3) Purpose.** The purpose of our spectral regularization is to mitigate dimensional collapse and learn diverse representations while (HaoChen et al., 2021) expects to obtain sample embeddings through spectral decomposition. Besides, their method relies on negative samples while our approach is negative-free.

C.2 Comparison with VICReg

A self-supervised method called VICReg (Bardes et al., 2022) adopts a similar line as ours, that is, extracting invariant information from two different views and applying specific strategies to prevent collapsed solutions. However, the following aspects differentiate our work from VICReg. **1) Network architecture.** After representation encoder, VICReg maps the representations into a embedding space by an expander (i.e., a projection head). The loss is computed in the embedding space. Our network architecture does not include additional projection heads, and all the loss is computed based on the node representations. **2) Implementation of invariant information extraction.** To extract invariant information in various views, VICReg applies mean-squared error to the embeddings from two views, which is a direct method to reduce the Euclidean distances between the embeddings from various views. Deviating from their strategy, from a statistical perspective, our consistency principle utilizes statistical indicators to maximize the statistical correlation between the representations from various views by regarding them as the empirical samples of multi-dimensional variables. The consistency principle allows for a more flexible choice of statistical indicators. **3) Prevention of collapsed solutions.** To prevent collapsed solutions, VICReg makes the squares of the non-diagonal elements of the covariance matrix of embeddings trend to 0, which can enhance the diversity of the representations. Our work proposes to reduce coupling between different representation channels by minishing their statistical correlation. Accordingly, two specific strategies are proposed: a) directly reducing the statistical correlation between various channels with specific statistical indicators; b) regularizing representations in spectral space. The experiments show similar effects of two strategies.

D Proof of Property 1

Here, we provide the proof of Property 1. Figure 6 can help to understand Property 1. For convenience, we restate Property 1:

Property 1. *For covariance matrix $\mathbf{\Sigma_H} = \frac{1}{N}\mathbf{H}^\top\mathbf{H} \in \mathbb{R}^{d\times d}$, which has d eigenvectors $[\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_d]$ associated with d eigenvalues $[\lambda_1, \lambda_2, \dots, \lambda_d]$, the variance of data $\mathbf{H}$ along the k-th principal direction (i.e., the direction of $\mathbf{q}_k$) is equal to λ_k.*

Proof. For N d-dimensional data points $\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_N]^\top \in \mathbb{R}^{N\times d}$, having been normalized to 0-mean along sample direction (*i.e.*, $\frac{1}{N}\sum_{i=1}^{N}\mathbf{h}_i = \mathbf{0}$), its covariance matrix is $\mathbf{\Sigma_H} = \frac{1}{N}\mathbf{H}^\top\mathbf{H}$. After eigendecomposition for $\mathbf{\Sigma_H}$, we can obtain d unit orthogonal eigenvectors $[\mathbf{q}_1, \dots, \mathbf{q}_d]$ associated to eigenvalues $[\lambda_1, \dots, \lambda_d]$, respectively. According to $\frac{1}{N}\mathbf{H}^\top\mathbf{H}\mathbf{q}_k = \lambda_k\mathbf{q}_k$, it can be known that

$$\frac{1}{N}\mathbf{q}_k^\top\mathbf{H}^\top\mathbf{H}\mathbf{q}_k = \lambda_k\mathbf{q}_k^\top\mathbf{q}_k = \lambda_k. \tag{23}$$

Taking a principal direction $\mathbf{q}_k$ as explanation, the projection of a sample $\mathbf{h}_i$ onto this direction is $z_i = \mathbf{q}_k^\top\mathbf{h}_i$, and the mean of all projections is

$$\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i = \frac{1}{N}\sum_{i=1}^{N}\mathbf{q}_k^\top\mathbf{h}_i = 0. \tag{24}$$

Thus, along of the principal direction $\mathbf{q}_k$, the variance is

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})^2 \\
=&\frac{1}{N}\sum_{i=1}^{N}\mathbf{q}_k^\top\mathbf{h}_i\mathbf{h}_i^\top\mathbf{q}_k \\
=&\frac{1}{N}\mathbf{q}_k^\top(\sum_{i=1}^{N}\mathbf{h}_i\mathbf{h}_i^\top)\mathbf{q}_k \\
=&\frac{1}{N}\mathbf{q}_k^\top\mathbf{H}^\top\mathbf{H}\mathbf{q}_k \\
=&\lambda_k.
\end{aligned} \tag{25}$$

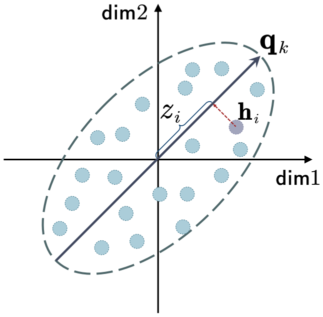


Figure 6: An illustration helping understanding Property 1.

The above equation demonstrates that the variance of data $\mathbf{H}$ along the direction $\mathbf{q}_k$ is equal to λ_k. We conclude the proof. □

E Proof of Theorem 1

The formal proof relies on the following property (Ahmed & Gokhale, 1989) of multivariate normal distribution:

Lemma 1. *Assuming a high-dimensional variable X obeys a d-dimensional Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ with mean μ and covariance matrix Σ, the entropy of variable X satisfies that*

$$Ent(X) = \frac{1}{2}\ln|\Sigma| + \frac{d}{2}(\ln 2\pi + 1), \tag{26}$$

where $Ent(X)$ is the entropy of variable X and $|\cdot|$ denotes the determinant of a matrix.

For convenience, we restate Theorem 1:

Theorem 1. *For representation matrix $\mathbf{H} \in \mathbb{R}^{N\times d}$, corresponding to a d-dimensional variable H, the entropy of H under empirical data $\mathbf{H}$ is maximized when $\mathbf{H}$ is completely decorrelated.*

Proof. Assuming the variable H obeys a d-dimensional normal distribution, according to Lemma 1, the entropy of variable H under empirical data $\mathbf{H}$ satisfies

$$Ent_{\mathbf{H}}(H) \propto \ln|\Sigma_{\mathbf{H}}|, \tag{27}$$

where $Ent_{\mathbf{H}}(H)$ denotes the information entropy of variable H under empirical data $\mathbf{H}$ and $\Sigma_{\mathbf{H}} = \frac{1}{N}\mathbf{H}^\top\mathbf{H}$ is empirical covariance matrix. For the representation matrix $\mathbf{H}$ has been normalized along sample direction, the diagonal elements of $\Sigma_{\mathbf{H}}$ are all equal to 1. Thus, we can know that $\sum_{i=1}^{d}\lambda_i = tr(\Sigma_{\mathbf{H}}) = d$, where $\lambda_1, \lambda_2, \dots, \lambda_d$ are d eigenvalues of $\Sigma_{\mathbf{H}}$ and $tr(\cdot)$ denotes matrix trace. According to the properties of determinant and AM-GM Inequality (Hirschhorn, 2007), we can know that

$$|\Sigma_{\mathbf{H}}| = \prod_{i=1}^{d}\lambda_i \le \left(\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_d}{d}\right)^d = 1. \tag{28}$$

It can be known that $|\Sigma_{\mathbf{H}}|$ achieves the upper bound of 1 when all eigenvalues are equal to 1, meaning that representation matrix $\mathbf{H}$ is completely decorrelated. Meanwhile, according to Eq. (27), the entropy of variable H under empirical data $\mathbf{H}$ reaches a maximum value. We conclude the proof. □

F Comparison with ZCA Whitening and Toy Experiments

Definition 1 (ZCA Whitening). *For input data $\mathbf{H} \in \mathbb{R}^{N\times d}$ with N d-dimensional vectors, which has been normalized to zero-mean along sample direction, ZCA Whitening processes the data as follows:*

$$ZCA(\mathbf{H}) = \mathbf{H}\mathbf{Q}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{Q}^\top, \tag{29}$$

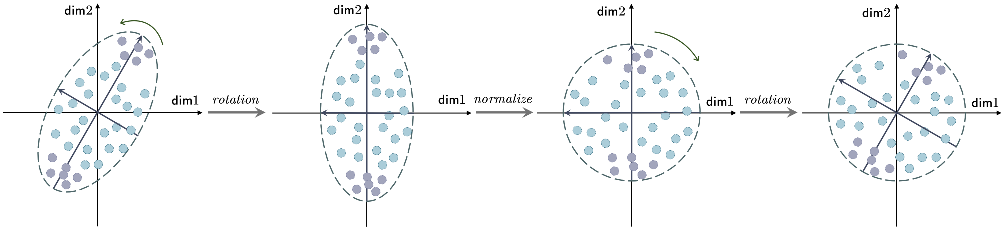
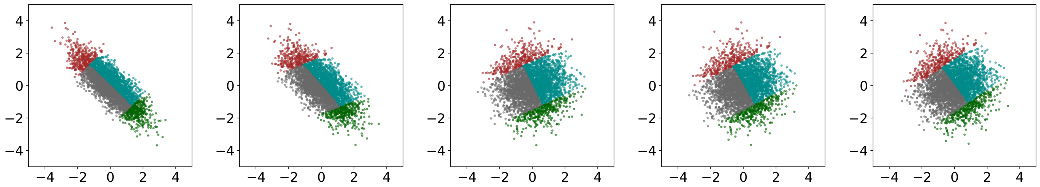


Figure 7: A diagram of ZCA whitening.



(a) Inputs (b) Random projection (c) Direct channel decorrelation (d) Spectral regularization (e) ZCA whitening

Figure 8: Visualizations of inputs and outputs of neural networks under various settings. For the convenience of visualization and comparison, the visualized points have been normalized to 0-mean and 1-variance. Color is used to reflect the relative positions of points. Best viewed in colors.

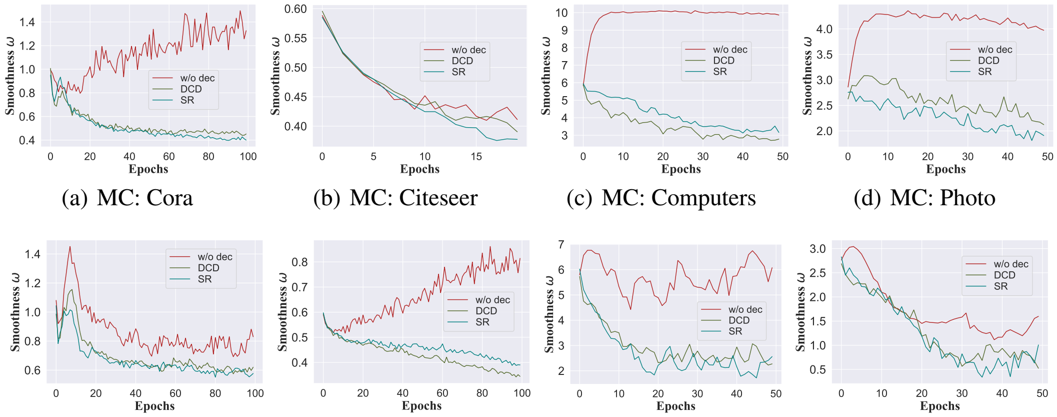
where $\mathbf{\Lambda} \in \mathbb{R}^{d\times d}$ *is a diagonal matrix filled with the eigenvalues of* $\mathbf{\Sigma} = \frac{1}{N}\mathbf{H}^\top\mathbf{H}$ *and* $\mathbf{Q} \in \mathbb{R}^{d\times d}$ *is the corresponding eigenvectors (i.e.,* $\mathbf{\Sigma} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$*).*

ZCA whitening (Bell & Sejnowski, 1997) realizes decorrelation between various channels (*i.e.*, dimensions) through matrix transformation. A vivid illustration of ZCA whitening is shown in Figure 7. Although both are aided by spectral properties, different from ZCA whitening, our strategy of spectral regularization decouples various channels by optimizing neural models under the specific objective function. Besides, our strategy can regulate the degree of decorrelation by controlling the objective objection, which can be regarded as a “soft” ZCA whitening.

We conduct some toy experiments to show the effects of ZCA whitening and our strategies. First, we build a simple neural network with three fully-connected layers. We sample 4,000 data points from a two-dimensional Gaussian distribution, then construct a data matrix $\mathbf{x} \in \mathbb{R}^{4,000\times 2}$, and further apply a rotation transformation to generate the inputs as shown in Figure 8(a). The input data represent a strong correlation between the two dimensions. A simple neural network with three fully-connected layers is built, and the outputs with randomly initialized parameters are shown in Figure 8(b), which still shows a strong dependency relationship between two dimensions. After applying ZAC whitening to the results of the random projection (*i.e.*, Figure 8(b)), the visualization in Figure 8(e), appearing as a circle, demonstrates its great ability of decorrelation. After being trained with our strategies of direct channel decorrelation and spectral regularization respectively, the network can output decoupled representations in Figure 8(c) and 8(d).

G Relationship between Decorrelation and Smoothness

Graph neural networks follow the paradigm of message aggregation, which has a smoothing effect on node representations and makes neighborhood nodes more similar. A graph usually is build under the homophily assumption, that is, nodes closely connected on a graph tend to have similar labels. Therefore, proper smoothing helps to learn good representations (Li et al., 2018). Nevertheless, over-smoothing makes all node representations collapse together and thus damages representations. Over-smoothing issue usually occurs in deep graph neural networks (Li et al., 2018). In this paper, we adopt one or two layers of GCNs as the backbone, so there is no over-smoothing issue.



(a) MC: Cora (b) MC: Citeseer (c) MC: Computers (d) MC: Photo

(e) DC: Cora (f) DC: Citeseer (g) DC: Computers (h) DC: Photo

Figure 9: Changes of graph smoothness during training under various decorrelation settings. The vertical axis represents $\omega(\mathbf{H}, \mathbf{A})$ in Definition 2. "w/o" denotes no decorrelation, "DCD" indicates direct channel decorrelation strategy, and "SR" indicates spectral regularization strategy. Here, matrix correlation is used to instantiate the consistency principle. The top row: matrix correlation; the bottom row: distance correlation. Best viewed in colors.

Here, we explore the relationship between decorrelation and graph smoothness. To this end, the graph smoothness is first defined as follows

Definition 2 (Graph Smoothness)**.** *For representation matrix* $\mathbf{H} \in \mathbb{R}^{N\times d}$ *on a graph with adjacency matrix* $\mathbf{A}$*, the graph smoothness can be measured by*

$$\omega(\mathbf{H}, \mathbf{A}) = \frac{\sum_{v_i \in \mathcal{V}} \sum_{v_j \in \mathcal{N}_{v_i}} A_{ij} \cdot \|\mathbf{h}_i - \mathbf{h}_j\|_2^2}{|\mathcal{E}| \cdot d},$$

where $|\mathcal{E}|$ *is the number of edges,* $\mathbf{h}_i$ *denotes the representation of node* v_i*, and* $\mathcal{N}_{v_i}$ *collects the neighbors of node* v_i*. The smaller the value of* $\omega(\mathbf{H}, \mathbf{A})$ *is, the higher the graph smoothness is.*

Empirically, we explore the relationship between decorrelation and graph smoothness. Concretely, under various decorrelation settings, we visualize the changes of graph smoothness on various datasets during training. As shown in Figure 9, with decorrelation operation, as the training process, the smoothness of the graph tends to improve. In other words, decorrelation principle potentially plays a smoothing effect on graphs, which makes neighborhood nodes more similar in representation space.

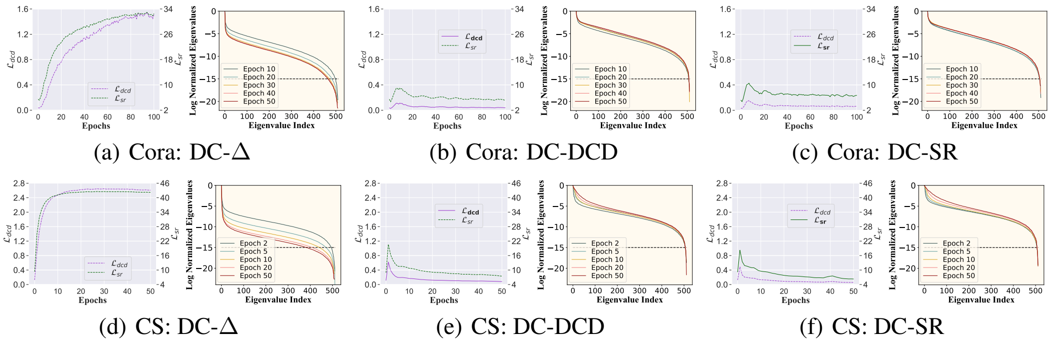
H Statistics of Seven Datasets

The statistics of seven experimental datasets are summarized in Table 3. The details of the datasets are as follows:

- **Cora**, **Citeseer** and **Pubmed** are three citation networks with nodes corresponding to documents and edges representing citation relationships. Each node (*i.e.*, document) has a class label indicating its category and is described by a bag-of-words feature vector.
- **Amazon-Computers** and **Amazon-Photo** are two graphs constructed from Amazon, representing co-purchase relationships between goods. The nodes indicate goods, and an edge is established between two nodes which are frequently bought together. A sparse bag-of-words feature about product reviews describes each node.
- **Coauthor-CS** and **Coauthor-Physics** are two academic network, where nodes represent authors and edges denote co-authorship relationships, respectively. Two nodes (*i.e.*, authors) are linked if they participate in a paper together.

Table 3: Statistics of the experimental datasets.

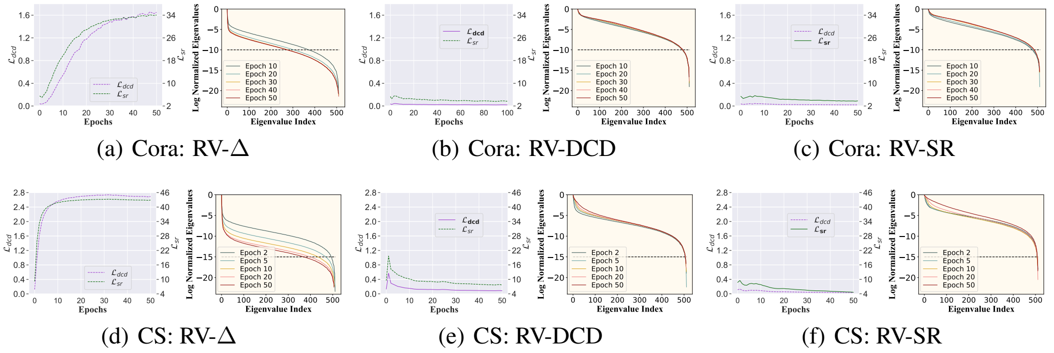
Dataset	Nodes	Edges	Features	Classes
Cora	2,708	5,429	1,433	7
Citeseer	3,327	4,732	3,703	6
Pubmed	19,717	44,338	500	3
Amazon-Computers	13,752	245,861	767	10
Amazon-Photo	7,650	119,081	745	8
Coauthor-CS	18,333	81,894	6,805	15
Coauthor-Physics	34,493	991,848	8,451	5



(a) Cora: DC-Δ (b) Cora: DC-DCD (c) Cora: DC-SR

(d) CS: DC-Δ (e) CS: DC-DCD (f) CS: DC-SR

Figure 10: The joint changes of two objectives and the evolution of eigenvalues of covariance matrix with Distance Correlation as the indicator measuring cross-view consistency.



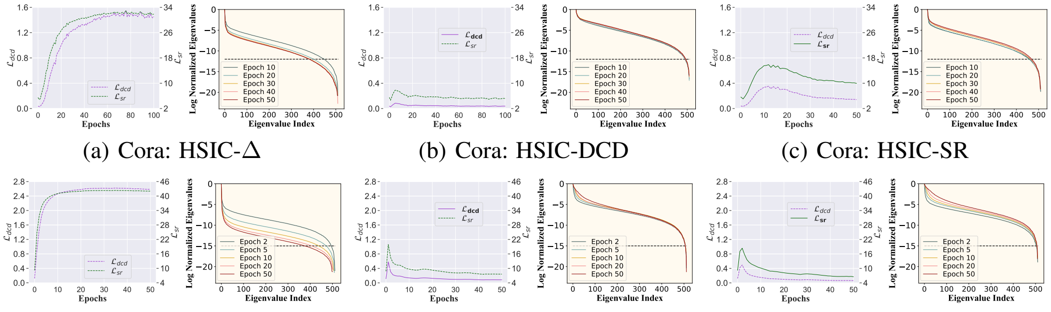
(a) Cora: RV-Δ (b) Cora: RV-DCD (c) Cora: RV-SR

(d) CS: RV-Δ (e) CS: RV-DCD (f) CS: RV-SR

Figure 11: The joint changes of two objectives and the evolution of eigenvalues of covariance matrix with RV-coefficient as the indicator measuring cross-view consistency.

I Supplement to the Experiments in Subsection 4.4

Here, we provide additional experiments as a complement to Subsection 4.4. Concretely, the same experiments are conducted on Cora and Coauthor-CS datasets with three other indicators (*i.e.*, DC, RV, and HSIC) as the measurements of cross-view consistency. The experiments are shown in Figure 10, 11, and 12. The observations and conclusions in Subsection 4.4 can still explain the supplementary experiments. Overall, under the objectives without relevant constraints for the relations between various representation channels, the model tends to learn low-rank solutions with redundancy, which is demonstrated by (a, d) in Figure 3, 10, 11, and 12. Besides, two proposed objectives can effectively prevent the model from this shortcut solution and have a consistent trend of change in the training.



(a) Cora: HSIC-Δ

(b) Cora: HSIC-DCD

(c) Cora: HSIC-SR

(d) CS: HSIC-Δ

(e) CS: HSIC-DCD

(f) CS: HSIC-SR

Figure 12: The joint changes of two objectives and the evolution of eigenvalues of covariance matrix with HSIC as the indicator measuring cross-view consistency.

J Visual Studies

J.1 t-SNE Visual Analysis

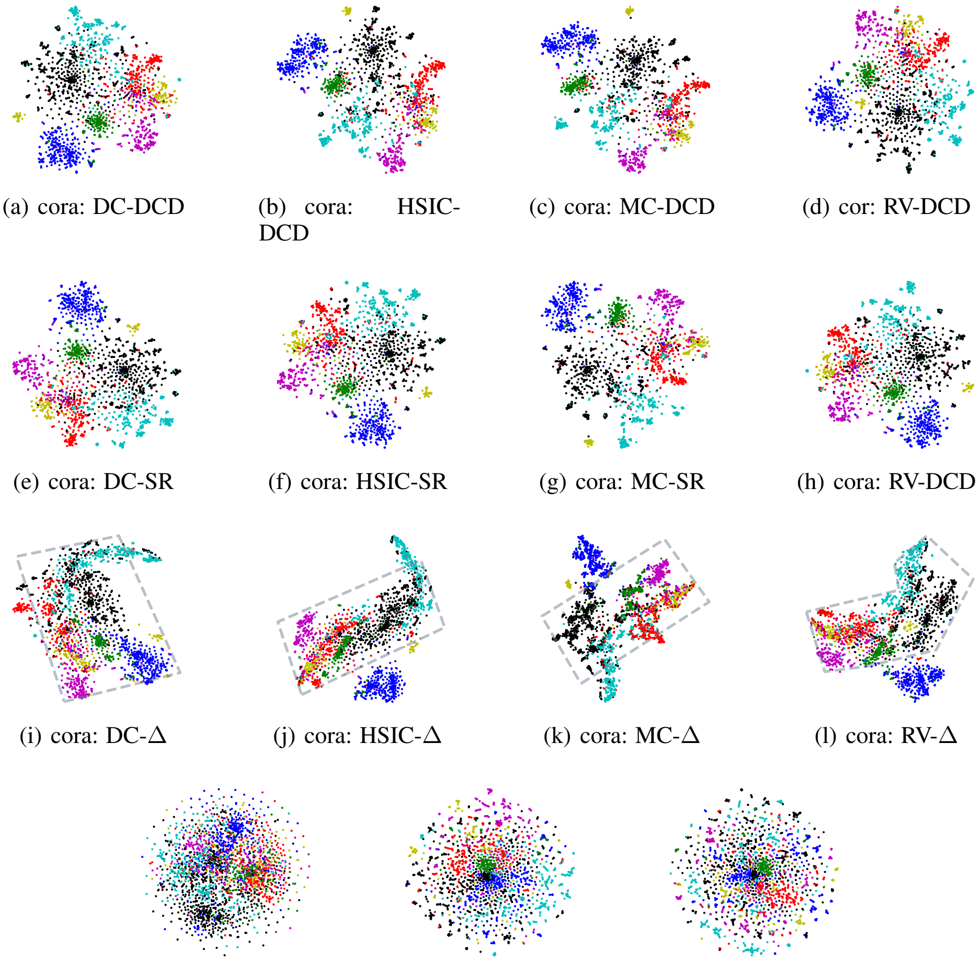
To better understand two principles, a series of t-SNE (Van der Maaten & Hinton, 2008) plots of the learned representations under different settings are drawn in Figure 13. As shown in Figure 13(m), the raw features are highly overlapping and do not present discriminable clusters. Besides, its shape in the 2-dimensional space takes on a circle, demonstrating that the two dimensions are not correlated and reflecting that no dimensional collapse occurs in the original feature space. The visualizations on the top two rows show that, under the guidance of two principles, the model can learn meaningful and interpretable representations, which are better gathered according to their real categories. As shown on the third row, without the decorrelation term, the model can still learn meaningful representations, whose projections in 2-dimensional space present discernible clusters. Nevertheless, the two dimensions exhibit a degree of linear correlation, where the gray boxes in the four figures assist in observing this phenomenon. The appearance in the two dimensional space can reflect dimensional collapse in the representation space, and the redundancy and correlation between various dimensions impair the quality of representations. In Figure 13(n, o), without consistency term, the 2-dimensional t-SNE embeddings take on mussy circular shapes, implying decoupled yet meaningless representations.

J.2 Correlation Matrix Visualization

In Figure 14, we visualize the absolute correlation matrices of the learned representations under various settings on Cora dataset. Concretely, for a representation matrix $\mathbf{H} \in \mathbb{R}^{N\times d}$, which has been normalized to 0-mean and 1-standard deviation, the absolute correlation matrix is $\mathbf{C} = |\frac{1}{N}\mathbf{H}^\top\mathbf{H}| \in [0,1]^{d\times d}$, where each element is the absolute value of Pearson correlation coefficient between two one-dimensional variables (*i.e.*, two channels). As shown on the bottom row of Figure 14, without decorrelation term, the off-diagonal elements are large, indicating that various channels of representation matrix are tightly correlated to each other and fail to capture diverse information. This phenomenon echoes the fifth row of Figure 13 and suggests the occurrence of dimensional collapse. The visualizations on the top four rows demonstrate that two proposed decorrelation strategies can effectively prevent dimensional collapse issue and facilitate learning highly disentangled and diverse representations.

K Summary of the Statistical Indicators

A summary of the five statistical indicators including mutual information are summarized in Table 4.



(a) cora: DC-DCD

(b) cora: HSIC-DCD

(c) cora: MC-DCD

(d) cor: RV-DCD

(e) cora: DC-SR

(f) cora: HSIC-SR

(g) cora: MC-SR

(h) cora: RV-DCD

(i) cora: DC-Δ

(j) cora: HSIC-Δ

(k) cora: MC-Δ

(l) cora: RV-Δ

(m) cora: raw features

(n) cora: Δ-DCD

(o) cora: Δ-SR

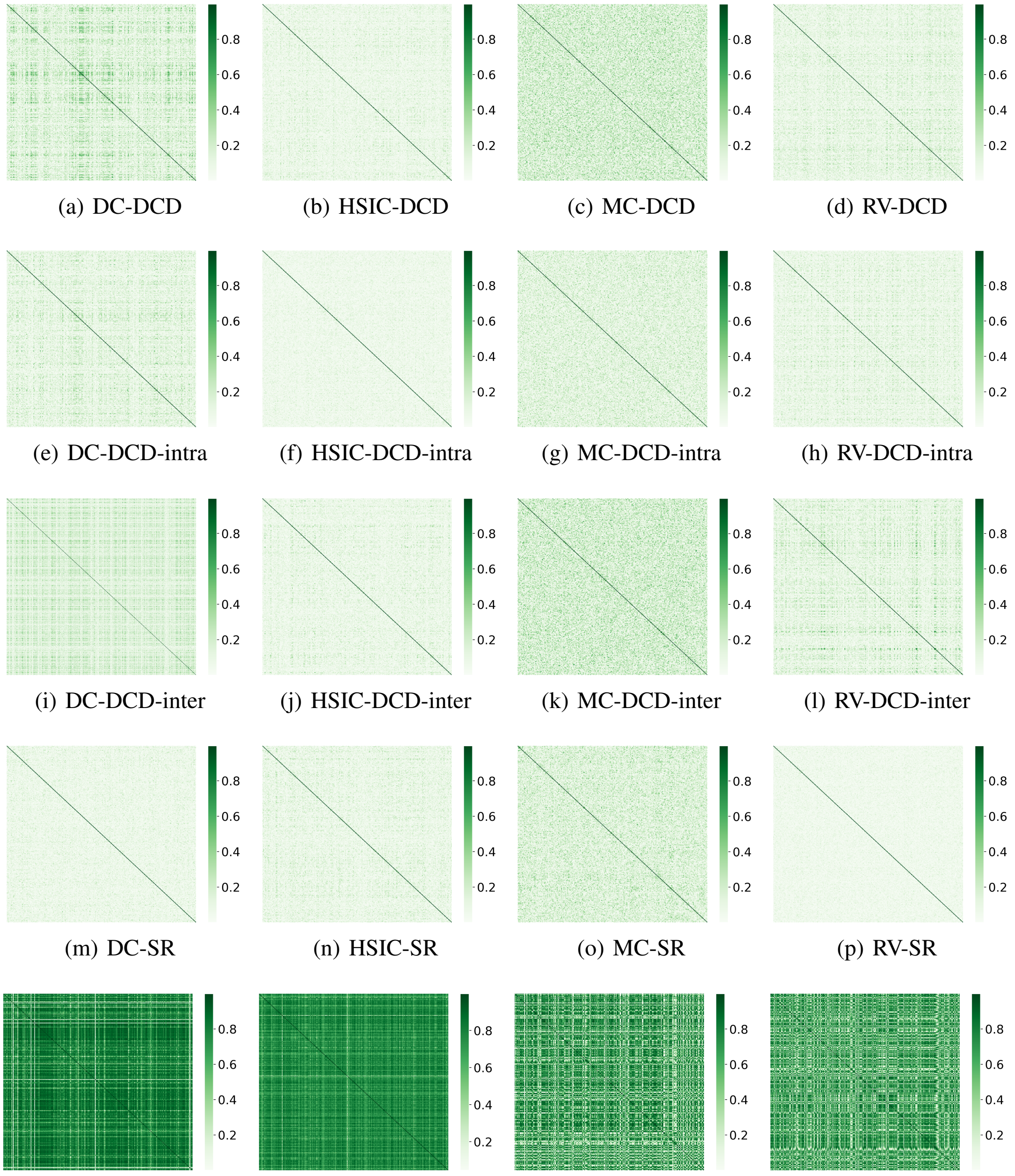
Figure 13: t-SNE visualizations of raw features and learned representations under various settings on Cora dataset. The colors indicate the categories of the sample points. Best viewed in colors.

Table 4: A summary of five statistical metrics. *Tractable*: whether statistical indicators can be obtained directly from empirical data. *Equal*: whether the dimensions of two variables need to be the same. *Linear/Nonlinear*: statistical relationships that can be captured. *Range*: theoretical value range. *Direction*: whether larger values indicate stronger correlation.

Indicator	Tractable	Equal	Linear/Nonlinear	Range	Direction
MI	✗	✗	Both	$[0, +\infty)$	✓
HSIC	✓	✗	Both	$[0, +\infty)$	✓
DC	✓	✗	Both	$[0, 1]$	✓
RV	✓	✗	Linear	$[0, 1]$	✓
MC	✓	✓	Linear	$[0, 1]$	✓

L PyTorch-Style Pseudocode

Here, we provide algorithm flow in the form of PyTorch-style pseudocode for four statistical indicators in subsection 3.2.2, two decorrelation strategies in subsection 3.3.2, and overall workflow.



(a) DC-DCD (b) HSIC-DCD (c) MC-DCD (d) RV-DCD

(e) DC-DCD-intra (f) HSIC-DCD-intra (g) MC-DCD-intra (h) RV-DCD-intra

(i) DC-DCD-inter (j) HSIC-DCD-inter (k) MC-DCD-inter (l) RV-DCD-inter

(m) DC-SR (n) HSIC-SR (o) MC-SR (p) RV-SR

(q) DC-Δ (r) HSIC-Δ (s) MC-Δ (t) RV-Δ

Figure 14: Visualizations of absolute correlation matrices of representations under various settings on Cora dataset. The second row shows the effects of intra-view decorrelation of DCD, while the third row describes inter-view decorrelation of DCD.

L.1 Pseudocode for Four Statistical Indicators

For two representation matrices $\mathbf{H}_A \in \mathbb{R}^{N\times d}$ and $\mathbf{H}_B \in \mathbb{R}^{N\times d}$ from views A and B, corresponding to two variables H_A and H_B, the calculation process of four statistical indicators are as follows.

Distance Correlation. The empirical evaluation of Distance Correlation (DC) between two variables is presented in Algorithm 1.

RV-coefficient. The empirical evaluation of RV-coefficient is provided in Algorithm 2.

Algorithm 1 PyTorch-style Code for Distance Correlation.

```
# H_A: representations from view A, shape=[N,d]
# H_B: representations from view B, shape=[N,d]

def DC(H_A, H_B):
    dis_matA = torch.sum(H_A * H_A, dim=1, keepdim=True) + torch.sum(H_A * H_A, dim=1,  \
               keepdim=True).T - 2 * torch.matmul(H_A, H_A.T)
    Dis_matA = dis_matA - torch.mean(dis_matA, dim=0, keepdim=True)  \
               - torch.mean(dis_matA, dim=1, keepdim=True) + torch.mean(dis_matA)
    dis_matB = torch.sum(H_B * H_B, dim=1, keepdim=True) + torch.sum(H_B * H_B, dim=1,  \
               keepdim=True).T - 2 * torch.matmul(H_B, H_B.T)
    Dis_matB = dis_matB - torch.mean(dis_matB, dim=0, keepdim=True)  \
               - torch.mean(dis_matB, dim=1, keepdim=True) + torch.mean(dis_matB)

    numer = torch.mean(Dis_matA * Dis_matB)
    denom = torch.sqrt(torch.mean(Dis_matA * Dis_matA) * torch.mean(Dis_matB * Dis_matB))
    DC_coe = numer / denom
    return DC_coe
```

Algorithm 2 PyTorch-style Code for RV-coefficient.

```
# H_A: representations from view A, shape=[N,d]
# H_B: representations from view B, shape=[N,d]

def RV(H_A, H_B):
    K_A = H_A @ H_A.T
    K_B = H_B @ H_B.T
    numerator = torch.trace(K_A @ K_B)
    denominator = torch.sqrt(torch.trace(K_A @ K_A) * torch.trace(K_B @ K_B))
    RV_coe = numerator / denominator
    return RV_coe
```

Matrix Correlation. The empirical evaluation of Matrix Correlation (MC) is presented in Algorithm 3. We expect that two variables have positive correlation, so we do not take the absolute value for MC in practice.

HSIC. The empirical evaluation of Hilbert-Schmidt Independence Criterion (HSIC) is provided in Algorithm 4. The function $\texttt{hsic1}(\texttt{H\_A}, \texttt{H\_B})$ strictly implements Eq. (6). After expanding the key step in $\texttt{hsic}(\texttt{H\_A}, \texttt{H\_B})$, we can obtain $\texttt{hsic2}(\texttt{H\_A}, \texttt{H\_B})$. The results of the two implementations are completely equivalent, but the latter has higher computational efficiency. In practice, like other indicators, we normalize HSIC in the form of function $\texttt{HSIC}(\texttt{H\_A}, \texttt{H\_B})$

In the above implementations, all N nodes are used to calculate the empirical estimation of the statistical indicators. It is permissible to only adopt a subset of all samples for estimation. Besides, for clarity and ease of understanding, there is a problem of redundant computations in the above implementations , which can be avoid in practice.

L.2 Pseudocode for Two Proposed Strategies for Decorrelation

For two normalized representation matrices $\mathbf{H}_A \in \mathbb{R}^{N \times d}$ and $\mathbf{H}_B \in \mathbb{R}^{N \times d}$, the algorithm flows for two strategies in subsection 3.3.2 are as follows.

Direct Channel Decorrelation. The specific calculation for Direct Channel Decorrelation is provided in Algorithm 5.

Spectral Regularization. The specific calculation for Spectral Regularization is provided in Algorithm 6.

Algorithm 3 PyTorch-style Code for Matrix Correlation.

```
# H_A: representations from view A, shape=[N,d]
# H_B: representations from view B, shape=[N,d]

def MC(H_A, H_B):
    numerator = torch.trace(H_A.T @ H_B)
    denominator = torch.sqrt(torch.trace(H_A.T @ H_A) * torch.trace(H_B.T @ H_B))
    MC_coe = numerator / denominator
    return MC_coe
```

Algorithm 4 PyTorch-style code for Hilbert-Schmidt Independence Criterion.

```
# H_A: representations from view A, shape=[N,d]
# H_B: representations from view B, shape=[N,d]

# linear kernel function
def linear_kernel(H_A):
    return torch.mm(H_A, H_A.T)

# gaussian kernel function
def gaussian_kernel(H_A, sigma):
    # sigma: width parameter
    dis_mat = torch.sum(H_A * H_A, dim=1, keepdim=True) + torch.sum(H_A * H_A, dim=1,  \
              keepdim=True).T - 2 * torch.matmul(H_A, H_A.T)
    return torch.exp(- dis_mat / sigma)

def hsic1(H_A, H_B):
    # utilize linear kernel
    K_A, K_B = linear_kernel(H_A), linear_kernel(H_B)
    # utilize linear kernel
    # K_A, K_B = gaussian_kernel(H_A, sigma), gaussian_kernel(H_B, sigma)
    N = H_A.shape[0]
    J = torch.eye(N) - torch.ones(N, N) / N
    return torch.trace(K_A @ J @ K_B @ J) / (N  - 1) ** 2

def hsic2(H_A, H_B):
    # utilize linear kernel
    K_A, K_B = linear_kernel(H_A), linear_kernel(H_B)
    # utilize linear kernel
    # K_A, K_B = gaussian_kernel(H_A, sigma), gaussian_kernel(H_B, sigma)
    N = H_A.shape[0]
    K_AB = torch.mm(K_A, K_B)
    hsic_ceo = torch.trace(K_AB) + torch.mean(K_A) * torch.mean(K_B) * N ** 2  \
           - 2 * torch.mean(K_AB) * N
    return hsic_ceo / (N - 1) ** 2

def HSIC(H_A, H_B):
    # normalize HSIC
    numerator = hsic2(H_A, H_B)
    denominator = torch.sqrt(hsic2(H_A, H_A) * hsic2(H_B, H_B))
    return numerator / denominator
```

Algorithm 5 PyTorch-style code for Direct Channel Decorrelation.

```
# H_A: representations from view A, shape=[N,d]
# H_B: representations from view B, shape=[N,d]

def loss_DCD(H_A, H_B):
    d = H_A.shape[1]  # dimension
    N = H_A.shape[0]  # number of nodes
    M = torch.ones(d, d) - torch.eye(d)  # mask matrix

    # intra-view correlation matrix
    c_A = torch.mm(H_A.T, H_A) / N
    c_B = torch.mm(H_B.T, H_B) / N
    # inter-view correlation matrix
    c_AB = torch.mm(H_A.T, H_B) / N

    loss_dcd_intra = (c_A * M).pow(2).sum() + (c_B * M).pow(2).sum()
    loss_dcd_inter = (c_AB * M).pow(2).sum()
    loss_dcd = loss_dcd_intra + loss_dcd_inter
    return loss_dcd / d / (d - 1)
```

L.3 Pseudocode for Overall Workflow

The overall workflow under two principles is summarized in Algorithm 7.

M Hyperparameter Sensitivity Analysis on Weighted Coefficients α and β

In this section, we conduct experiments to explore the effects of weighted coefficients α and β of the consistency term and the decorrelation term in Eq. (15) of the main text. Concretely, we evaluate the impact of various combinations of α and β on node classification accuracy on Cora dataset. The experimental results under various settings are presented in Figure 15. It can be found that the good results benefit from the appropriate values of the two hyperparameters. Actually, what matters is their ratio $\frac{\alpha}{\beta}$. For instance, in Figure 15(a), the performance is always satisfactory when $\frac{\alpha}{\beta}$ is about 5.

Algorithm 6 PyTorch-style code for Spectral Regularization.

```
# H_A: representations from view A, shape=[N,d]
# H_B: representations from view B, shape=[N,d]

def loss_SR(H_A, H_B):
    N = H_A.shape[0]  # number of nodes

    # calculate covariance matrix
    c_A = torch.mm(H_A.T, H_A) / N
    c_B = torch.mm(H_B.T, H_B) / N

    eigvals_A = torch.linalg.eigvals(c_A).float()
    std_A = torch.std(eigvals_A)
    eigvals_B = torch.linalg.eigvals(c_B).float()
    std_B = torch.std(eigvals_B)

    loss_sr = std_A + std_B
    return loss_sr
```

Algorithm 7 Overall Workflow under Two Principles

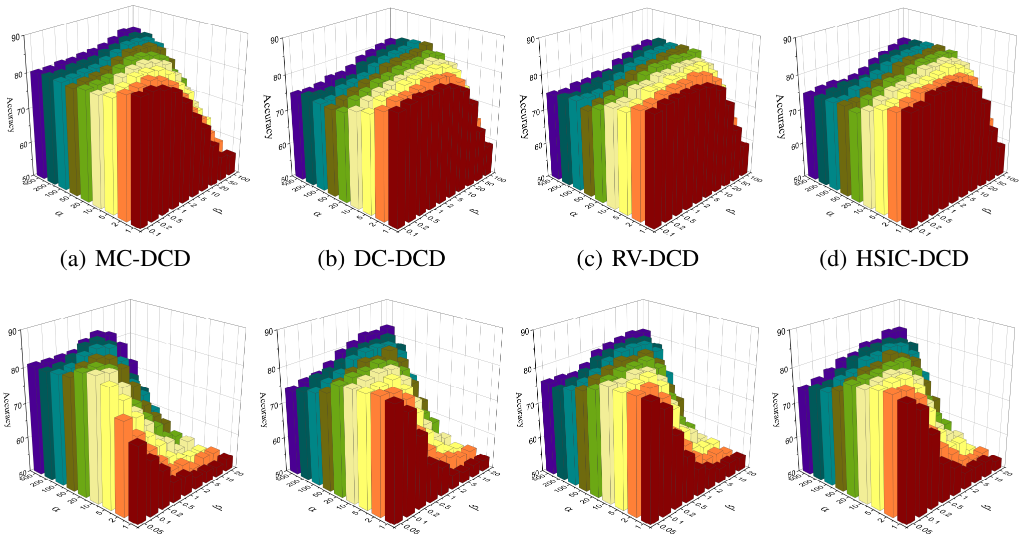
Input: A graph $G(\mathbf{A}, \mathbf{X})$ with N nodes, neural encoder f_θ, weighted coefficients α and β, augmentation function space $\mathcal{T}$, correlation indicator $Cor(\cdot, \cdot)$, training epochs T.

```
Initialize $f_\theta$;
repeat
    Randomly sample two augmentation functions $\tau_A$ and $\tau_B$ from $\mathcal{T}$;
    Generate two augmented views $G'_A(\mathbf{A}'_A, \mathbf{X}'_A) = \tau_A(G)$ and $G'_B(\mathbf{A}'_B, \mathbf{X}'_B) = \tau_B(G)$;
    Obtain node representations $\widetilde{\mathbf{H}}_A = f_\theta(\mathbf{A}'_A, \mathbf{X}'_A)$ and $\widetilde{\mathbf{H}}_B = f_\theta(\mathbf{A}'_B, \mathbf{X}'_B)$;
    Get normalized representations $\mathbf{H}_A$ and $\mathbf{H}_B$;
    Calculate consistency loss $\mathcal{L}_{cvmc}$ based on the given $Cor(\cdot, \cdot)$;
    Calculate decorrelation loss $\mathcal{L}_{bcmd}$ according to Eq. (13) or Eq. (14);
    Obtain the overall objective $\mathcal{L} = \alpha\mathcal{L}_{cvmc} + \beta\mathcal{L}_{bcmd}$;
    Update parameters $\theta$ through back propagation;
until reaching maximum training steps $T$
Get $\mathbf{H} = f_\theta(\mathbf{A}, \mathbf{X})$ for downstream tasks.
```

Similarly, in Figure 15(e), it can work well when $\frac{\alpha}{\beta}$ is approximately equal to 200. When applying our method to a new dataset, we can fix one hyperparameter (*e.g.*, α) and adjust the other one.

N Hyperparameter Sensitivity Analysis on Augmentation Intensity

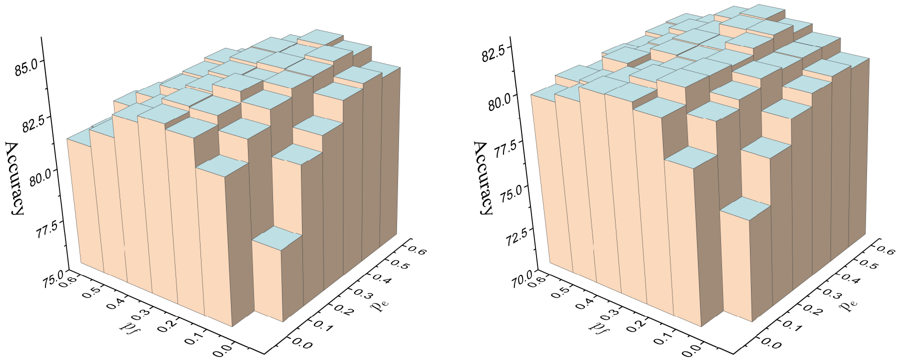
We study the influences of augmentation intensity on node classification accuracy. Various combinations of feature masking ratio p_f and edge removal ratio p_e are attempted on Cora and Pubmed datasets. As shown in Figure 16, the optimal performance is achieved under the best combination of p_e and p_f. Besides, when p_e and p_f are in a proper range, the experimental results are always competitive, which reflects the robustness of our framework. Besides, we can find that appropriately strong augmentations (*i.e.*, larger p_e and p_f) contribute to better performance for it is helpful to mine augmentation-invariant information.



(a) MC-DCD (b) DC-DCD (c) RV-DCD (d) HSIC-DCD

(e) MC-SR (f) DC-SR (g) RV-SR (h) HSIC-SR

Figure 15: Node classification accuracy under various combinations of weighted coefficients α and β on Cora dataset. In the caption of each subfigure, the left of "-" denotes the employed statistical indicator while its right represents the adopted decorrelation strategy. DC: Distance Correlation; RV: RV-coefficient; MC: Matrix Correlation; HSIC: Hilbert-Schmidt Independence Criterion; DCD: Direct Channel Decorrelation; SR: Spectral Regularization.



(a) Cora (b) Pubmed

Figure 16: Node classification accuracy under various combinations of feature masking ratio p_f and edge removal ratio p_e on Cora and Pubmed. When $p_e = 0$ and $p_f = 0$, the experimental results are 52.5 and 47.7 on Cora and Pubmed, which are not displayed for better visualization.